PERTANIKA JOURNAL OF SOCIAL SCIENCES & HUMANITIES

About the Journal

Overview

Pertanika Journal of Social Sciences & Humanities is the official journal of Universiti Putra Malaysia. It is an open-access online scientific journal. It publishes original scientific outputs. It neither accepts nor commissions third party content.

Recognised internationally as the leading peer-reviewed interdisciplinary journal devoted to the publication of original papers, it serves as a forum for practical approaches to improve quality in issues pertaining to social and behavioural sciences as well as the humanities.

Pertanika Journal of Social Science & Humanities is a **quarterly** (*March*, *June*, *September*, *and December*) periodical that considers for publication original articles as per its scope. The journal publishes in **English** as well as in **Bahasa Malaysia** and it is open for submission by authors from all over the world.

The journal is available world-wide.

Aims and scope

Pertanika Journal of Social Sciences & Humanities aims to develop as a pioneer journal for the social sciences with a focus on emerging issues pertaining to the social and behavioural sciences as well as the humanities. Areas relevant to the scope of the journal include Social Sciences—accounting, anthropology, archaeology and history, architecture and habitat, consumer and family economics, economics, education, finance, geography, law, management studies, media and communication studies, political sciences and public policy, population studies, psychology, sociology, technology management, and tourism; Humanities—arts and culture, dance, historical and civilisation studies, language and linguistics, literature, music, philosophy, religious studies, and sports.

History

Pertanika Journal of Social Sciences & Humanities was founded in 1993 and focuses on research in social and behavioural sciences as well as the humanities and its related fields.

Vision

To publish journal of international repute.

Mission

Our goal is to bring the highest quality research to the widest possible audience.

Quality

We aim for excellence, sustained by a responsible and professional approach to journal publishing. Submissions can expect to receive a decision within 120 days. The elapsed time from submission to publication for the articles averages 180 days. We are working towards decreasing the processing time with the help of our editors and the reviewers.

Abstracting and indexing of Pertanika

Pertanika Journal of Social Sciences & Humanities is now over 27 years old; this accumulated knowledge and experience has resulted the journal being abstracted and indexed in SCOPUS (Elsevier), Clarivate Web of Science (ESCI), EBSCO, DOAJ, Agricola, ASEAN CITATION INDEX, ISC, Microsoft Academic, Google Scholar, and MyCite.

Citing journal articles

The abbreviation for Pertanika Journal of Social Sciences & Humanities is *Pertanika J. Soc. Sci. & Hum.*

Publication policy

Pertanika policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. It prohibits as well publication of any manuscript that has already been published either in whole or substantial part elsewhere. It also does not permit publication of manuscript that has been published in full in proceedings.

Code of Ethics

The *Pertanika* journals and Universiti Putra Malaysia take seriously the responsibility of all of its journal publications to reflect the highest publication ethics. Thus, all journals and journal editors are expected to abide by the journal's codes of ethics. Refer to *Pertanika*'s **Code of Ethics** for full details, or visit the journal's web link at: http://www.pertanika.upm.edu.my/code_of_ethics.php

Originality

The author must ensure that when a manuscript is submitted to *Pertanika*, the manuscript must be an original work. The author should check the manuscript for any possible plagiarism using any program such as Turn-It-In or any other software before submitting the manuscripts to the *Pertanika* Editorial Office, Journal Division.

All submitted manuscripts must be in the journal's acceptable **similarity index range**:
≤ 20% – PASS; > 20% – REJECT.

International Standard Serial Number (ISSN)

An ISSN is an 8-digit code used to identify periodicals such as journals of all kinds and on all media–print and electronic.

Pertanika Journal of Social Sciences and Humanities: e-ISSN 2231-8534 (Online).

Lag time

A decision on acceptance or rejection of a manuscript is reached in 120 days (average). The elapsed time from submission to publication for the articles averages 180 days.

Authorship

Authors are not permitted to add or remove any names from the authorship provided at the time of initial submission without the consent of the journal's Chief Executive Editor.

Manuscript preparation

Most scientific papers are prepared according to a format called IMRAD. The term represents the first letters of the words *Introduction*, *Materials and Methods*, *Results*, *And Discussion*. IMRAD is simply a more 'defined' version of the "IBC" (*Introduction*, *Body*, *Conclusion*) format used for all academic writing. IMRAD indicates a pattern or format rather than a complete list of headings or components of research papers; the missing parts of a paper are: *Title*, *Authors*, *Keywords*, *Abstract*, *Conclusions*, *References*, and *Acknowledgement*. Additionally, some papers include *Appendices*. For manuscripts in **Bahasa Malaysia**, the title, abstract and keywords should be written in both **English** and **Bahasa Malaysia**.

The *Introduction* explains the scope and objective of the study in the light of current knowledge on the subject; the *Materials and Methods* describes how the study was conducted; the *Results* section reports what was found in the study; and the *Discussion* section explains meaning and significance of the results and provides suggestions for future directions of research. The manuscript must be prepared according to the journal's **Instruction to Authors** (http://www.pertanika.upm.edu.my/Resources/regular_issues/Regular_Issues_Instructions_to_Authors.pdf).

Editorial process

Authors who complete any submission are notified with an acknowledgement containing a manuscript ID on receipt of a manuscript, and upon the editorial decision regarding publication.

Pertanika follows a **double-blind peer -review process**. Manuscripts deemed suitable for publication are sent to reviewers. Authors are encouraged to suggest names of at least 3 potential reviewers at the time of submission of their manuscripts to *Pertanika*, but the editors will make the final selection and are not, however, bound by these suggestions.

Notification of the editorial decision is usually provided within 120 days from the receipt of manuscript. Publication of solicited manuscripts is not guaranteed. In most cases, manuscripts are accepted conditionally, pending an author's revision of the material.

As articles are double-blind reviewed, material that may identify authorship of the paper should be placed only on page 2 as described in the first-4-page format in Pertanika's Instruction to Authors (http://www.pertanika.upm.edu.my/Resources/regular_issues/Regular_Issues_Instructions_to_Authors.pdf).

The journal's peer review

In the peer review process, 2 or 3 referees independently evaluate the scientific quality of the submitted manuscripts. At least 2 referee reports are required to help make a decision.

Peer reviewers are experts chosen by journal editors to provide written assessment of the **strengths** and **weaknesses** of written research, with the aim of improving the reporting of research and identifying the most appropriate and highest quality material for the journal.

Operating and review process

What happens to a manuscript once it is submitted to *Pertanika*? Typically, there are 7 steps to the editorial review process:

1. The journal's Chief Executive Editor and the Editor-in-Chief examine the paper to determine whether it is relevance to journal needs in terms of novelty, impact, design, procedure, language as well as presentation and allow it to proceed to the reviewing process. If not appropriate, the manuscript is rejected outright and the author is informed.

2. The Chief Executive Editor sends the article-identifying information having been removed, to 2 to 3 reviewers. They are specialists in the subject matter of the article. The Chief Executive Editor requests that they complete the review within 3 weeks.

    Comments to authors are about the appropriateness and adequacy of the theoretical or conceptual framework, literature review, method, results and discussion, and conclusions. Reviewers often include suggestions for strengthening of the manuscript. Comments to the editor are in the nature of the significance of the work and its potential contribution to the research field.

3. The Editor-in-Chief examines the review reports and decides whether to accept or reject the manuscript, invite the authors to revise and resubmit the manuscript, or seek additional review reports. In rare instances, the manuscript is accepted with almost no revision. Almost

without exception, reviewers' comments (to the authors) are forwarded to the authors. If a revision is indicated, the editor provides guidelines to the authors for attending to the reviewers' suggestions and perhaps additional advice about revising the manuscript.

4. The authors decide whether and how to address the reviewers' comments and criticisms and the editor's concerns. The authors return a revised version of the paper to the Chief Executive Editor along with specific information describing how they have addressed the concerns of the reviewers and the editor, usually in a tabular form. The authors may also submit a rebuttal if there is a need especially when the authors disagree with certain comments provided by reviewers.

5. The Chief Executive Editor sends the revised manuscript out for re-review. Typically, at least 1 of the original reviewers will be asked to examine the article.

6. When the reviewers have completed their work, the Editor-in-Chief examines their comments and decides whether the manuscript is ready to be published, needs another round of revisions, or should be rejected. If the decision is to accept, the Chief Executive Editor is notified.

7. The Chief Executive Editor reserves the final right to accept or reject any material for publication, if the processing of a particular manuscript is deemed not to be in compliance with the S.O.P. of *Pertanika*. An acceptance notification is sent to all the authors.

9. The editorial office ensures that the manuscript adheres to the correct style (in-text citations, the reference list, and tables are typical areas of concern, clarity, and grammar). The authors are asked to respond to any minor queries by the editorial office. Following these corrections, page proofs are mailed to the corresponding authors for their final approval. At this point, **only essential changes are accepted**. Finally, the manuscript appears in the pages of the journal and is posted on-line.

ABSTRACTING AND INDEXING OF PERTANIKA JOURNALS

The journal is indexed in SCOPUS (Elsevier), Clarivate-Emerging Sources Citation Index (ESCI), BIOSIS, National Agricultural Science (NAL), Google Scholar, MyCite, ISC. In addition, Pertanika JSSH is recipient of "CREAM" Award conferred by Ministry of Higher Education (MoHE), Malaysia.

Pertanika Journal of Social Sciences & Humanities
Vol. 29 (S3) 2021

Contents

Language Education: Benchmarking and Standardization

Preface

Language learning aims to achieve a competency level that allows for effective communication with speakers of the language and accurate comprehension of materials written or presented in the language. Internationally recognized language examinations such as TOEFL and IELTS have long been used to determine a learner's proficiency in English. However, the Common European Framework of Reference for languages, commonly referred to by its acronym CEFR, is now becoming increasingly accepted as a major reference describing language learners' abilities. The CEFR is relevant to all languages as its descriptors and "can-do statements" reflect the major goals of language use regardless of language. The six levels of the CEFR, ranging from A1 to C2, have become popular proficiency descriptors. It is now common to hear one being described as having a B1 level of proficiency instead of having intermediate proficiency. The impact of the CEFR on educational systems has been felt in Europe, where it originated, and in other parts of the world. In Malaysia, for example, the CEFR has been adopted as a guide and is referenced in the National Education Blueprint with English language learners in the country expected to achieve a B1/B2 level at the end of their secondary education. Many institutions have also started to use the CEFR level descriptors as part of their language admission requirements. Subsequently, language assessment, curricula, courses, and learning materials have adjusted to this change.

This theme-based special issue of *Pertanika* JSSH on Language Education: Benchmarking and Standardization mainly looks at the responses by various institutions towards the increasing popularity of the CEFR and, in some instances, its formal implementation in the education system. Many authors of the twenty-three articles in this issue directly describe how their institutions have responded to the CEFR, including standardization and benchmarking in language assessment, goals, and activities. While most of the articles focus on teaching and learning the English language, a few articles discuss the teaching of other languages. As we acknowledge the importance of constructing our language curriculum and assessment to meet international standards and benchmarks, we hope that this special issue will provide insights into how best to attain standardization and benchmark language proficiency levels in a global context.

Finally, we would like to thank all the contributors for sharing their experiences and insights, the manuscript reviewers in ensuring the appropriate quality level, and the *Pertanika* staff for their patience and cooperation throughout the review process until the publication.

Guest Editors

Arshad Abd. Samad *(Prof. Dr.)*
Tan Wee Chun *(Dr.)*
Zailin Shah Hj. Yusoff *(Assoc. Prof. Dr.)*

Assessment of the Analytic Scale of Argumentative Writing (ASAW)

Vahid Nimehchisalem[1]*, Jayakaran Mukundan[2], Shameem Rafik-Galea[3] and Arshad Abd Samad[4]

[1]*Department of English, Faculty of Modern Languages and Communication, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[2]*Department of Language and Humanities Education, Faculty of Educational Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[3]*Faculty of Education, Languages and Psychology, SEGi University, Kota Damansara PJU 5, 47810 Petaling Jaya, Selangor, Malaysia*
[4]*School of Education, Taylor's University, No. 1, Jalan Taylor's, 47500 Subang Jaya, Selangor, Malaysia*

ABSTRACT

The Analytic Scale of Argumentative Writing (ASAW) was developed because of the need for a genre-specific scale to assess English as a Second Language (ESL) university student writers' argumentative essays. The present study reports the findings of field-testing ASAW. For this purpose, argumentative samples (n = 110) were collected and remote-scored by experienced raters (n = 5) who used ASAW. Overall, moderate to high inter-rater reliability (r = 0.7-0.9), as well as high (r = 0.84-0.92) and moderate to high (r = 0.70-0.77) intra-rater reliability coefficients after short (6-week) and long (9-week) rating intervals were obtained, respectively. Some established instruments were used to score the same essays rated using ASAW to test the concurrent validity of the scale. The scores assigned by the raters using the scale demonstrated moderate (r = 0.51) to high (r = 0.77) correlations with the scores awarded using several other standard instruments. The raters who used ASAW were given a questionnaire to evaluate the scale itself, and on average, the results indicated that the raters were highly satisfied with it. It took an average of 5.5 minutes for the raters to evaluate an essay, indicating it was economical. The study has useful implications for refinement of ASAW and development and validation of similar scales and benchmarks in the future.

Keywords: Analytic scale, argumentative writing, assessing writing, English as a second language, instrument evaluation

INTRODUCTION

English as a Second Language (ESL) learners' writing may be assessed through impressionistic or scale-based methods. Due to the problems of impressionistic measurement (Brennan et al., 2001), writing instructors are advised to use rating scales as guidelines that help them judge the learners' writing more objectively. Scales may be holistic or analytic. Holistic scales [e.g., Performance Descriptors for the TOEFL iBT® Test (Educational Testing Service, 2011)] help the rater assign a single score for students' overall writing ability. Thus, they are appropriate for large-scale language proficiency tests. Analytic scales [e.g., ESL Composition Profile (ESL-CP) (Jacobs et al., 1981)] allow raters to assign individual scores for each sub-trait (e.g., content or organization) and are suitable for diagnosing students' specific writing problems. Scales may also be generic or genre-specific. In contrast to generic scales that are all-purpose, genre-specific scales are sensitive to the unique features of the genre they assess. This specificity contributes to their construct validity (Cooper, 1999). Despite their costly development and administration procedures, genre-specific scales that are also analytic are instrumental instruction, assessment, and research tools.

Many writing scales are available in the literature. Most are generic and holistic (e.g., Performance Descriptors for the TOEFL iBT® Test), while some are generic and analytic (e.g., ESL-CP). A few analytic genre-specific scales are also available. For example, Connor and Lauer (1988)

developed the Argumentative Quality Scale (AQS) that focuses only on students' argumentative writing ability, leaving out traits like grammar or vocabulary. An analytic three-point scale includes three sub-scales of 'claim,' 'data,' and 'warrant,' following Toulmin's (1958) model of argument. Persuasive Appeals Scale (PAS) is another similar instrument, developed based on the Theory of Classical Rhetoric (Kinneavy, 1971), for evaluating persuasive appeals. It is a four-point scale with three sub-scales of 'rational,' 'credibility,' and 'affective' appeals (Connor & Lauer, 1988).

Yeh (1998) developed and compared two analytic scales for assessing argumentative essays for American school students. The first had the sub-scales of 'claim clarity,' 'reason strength,' and 'rebuttals to counterarguments' while the second focused on 'development, organization, focus, and clarity,' 'voice, and conventions.' Better test results were obtained for the second scale. The sub-scales of the first instrument explained only a third of the variance in holistic scores, while those of the second scale accounted for two-thirds of the variance in holistic scores (Yeh, 1998) obviously because it covered a wider scope of argumentative writing construct.

In New Zealand, Glasswell et al. (2001) developed six analytic genre-specific scales for assessing school students' ability to 'explain,' 'argue,' 'instruct,' 'classify,' 'inform' and 'recount'. Every scale had four sub-scales, 'audience awareness and purpose,' 'content inclusion,' 'coherence,' and 'language resources.' The scales were

tested for consequential validity, ease of use, relevance to the test context (Glasswell et al., 2001). Tests of reliability showed adjacent agreement consensus of (70-90%) and measurement correlations of r = 0.70-0.80 (Brown et al., 2004).

To the researchers' knowledge, only one university-level validated genre-specific scale is available in Malaysia, developed at the Universiti Kebangsaan Malaysia (Wong, 1989). Therefore, a data-based method was followed, in which 20 narratives purposively collected from the target students from different writing performance levels were analyzed. The scale was tested for its reliability and concurrent validity before being used for placement purposes (Wong, 1989).

Scale Validation

It is considered valid if an instrument measures what it claims to measure (Cronbach, 1971). Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). In other words, a valid instrument should have both evidential and consequential bases. According to Messick (1989), an instrument is considered evidentially valid if it is based on well-established and relevant theories; that is if it has construct validity. Additionally, Messick (1989) regards the instrument as consequentially valid if it has construct validity and if its users find it practical, satisfactory, and useful.

Writing scales are validated through qualitative and/or quantitative methods. A panel of experts familiar with the learning-testing situation for which the scale is being developed may be involved in the validation process. In addition, scales may be tested for their reliability and concurrent validity through statistical methods. How stringently a scale should be tested depends on the sensitivity of the decision based on the awarded scores about that scale. In the case of international high-stakes language tests, it is necessary to test the scale rigorously and continuously. However, such high standards are rarely expected from scales used in local tests.

Validity should be considered while developing (*a priori*) and after administrating (*a posteriori*) a scale (Weir, 2005). *A priori* validity is theory-based and has a judgmental and subjective nature; therefore, to be valid, an instrument should also go through *a posteriori* validation process, which provides empirical evidence on its relevance. *A posteriori* validity is determined by scoring, criterion-related and consequential validation (Weir, 2005). Scoring validity indicates the reliability or score consistency reached after repeated scale administrations to rate similar samples. The extent to which test scores correlate with a suitable external performance criterion is known as criterion-related validity. Finally, an instrument is consequentially valid if its stakeholders are satisfied with it. Factors like practicality are related to consequential validity; if an instrument is cost-effective, it will indicate higher consequential validity,

or according to Bachman and Palmer (1996), higher micro-/macro-level impact on its stakeholders. This study seeks to determine *a posteriori* validity of an analytic genre-specific scale, called the Analytic Scale of Argumentative Writing (after this referred to as 'ASAW' or 'the scale').

To address the gap in the literature, we developed an analytic genre-specific scale to help raters assess argumentative essays. What follows is a background on the results of our developmental study, which have previously been published in separate articles. As discussed in the next section, while the construct validity of ASAW was tested in our previous studies, the present paper is concerned more with its consequential validity.

Development of ASAW

ASAW was developed based on the Pyramid of Argumentation (Nimehchisalem, 2018). In an attempt to show the inter-relationship between the elements of communicative language competence and argumentation, this composite framework combines:

1. Theory of Communicative Language Ability (Bachman, 1990), composed of 'knowledge of language,' 'strategic competence' and 'psychophysiological mechanisms,' all interacting with the 'context of situation' and 'world knowledge;'
2. Taxonomy of Components of Language Competence (Bachman, 1990), including 'organizational competence' (the way texts are organized) and 'pragmatic competence' (the way texts are related to users' communicative goals and the features of language use context) (Bachman & Palmer, 1996);
3. Theory of Classical Rhetoric (Kinneavy, 1971) including 'ethical appeal,' 'rhetorical situation,' 'rhetorical style,' and 'arrangement' (with 'emotional appeals' excluded in the Pyramid of Argumentation to differentiate argumentative from persuasive writing, and with 'logical appeals' replaced by Toulmin's Model of Argument); and
4. Model of Argument (Toulmin, 1958) consisting of claim, data (supporting the claim), warrant (bridging the claim and data), backing (supporting the warrant), rebuttal (accounting for counterarguments), and qualifiers (indicating the certainty of the argument).

An evaluative criteria checklist was developed based on this theoretical framework, the previous scales, and the related literature. It went through three complementary studies to be operationalized:

1. A survey elicited experienced (≥2 years) Malaysian ESL writing lecturers' (n = 88) views on the importance, comprehensiveness, and clarity of the scale items. Principal Component Analysis was used to explore the experts' views on the essential dimensions

of argumentative writing. The survey results suggested grouping the criteria under three domains of 'content,' 'organization,' and 'language,' which cumulatively explained 57.4% of the variance (Nimehchisalem & Mukundan, 2011).

2. A focus group study involved female Malaysian senior lecturers (n = 4) with a minimum of 5 years of teaching and rating experience. They identified 'task fulfillment,' 'content' and 'organization' (*highly important*); 'vocabulary' and 'style' (*important*); and finally 'grammar' and 'mechanics' (*fairly important*) as the essential dimensions in evaluating argumentative essays (Nimehchisalem et al., 2012).

3. A data-based analysis of argumentative samples (n = 20) that had been collected from the target students resulted in the descriptors of 'content' and 'organization' sub-scales of ASAW (Nimehchisalem & Mukundan, 2013).

A scale emerged with five sub-scales of 'content,' 'organization,' 'language conventions,' 'vocabulary,' and 'overall effectiveness' with equal weights assigned to each sub-scale. A score converter was added to ASAW to help raters convert the scores to their corresponding grade in the university grading system (Appendix 1). As this brief background illustrates, ASAW has gone through several stages to strengthen its theoretical foundation and validity. The present study was done further to test its validity, reliability, and economy.

Objective and Research Questions

The objective of this study was to test the reliability, concurrent validity, economy of ASAW, and micro-level consequential validity. The following research questions were addressed:

1. How consistently are the scores assigned for the same written samples by different experienced raters using ASAW?

2. Is there a significant correlation between the:
   - learners' 'total' scores assigned to their essays using ASAW and their general English proficiency band scores?
   - ASAW 'content' scores and the 'total' scores assigned to similar essays using AQS?
   - ASAW 'content' scores and the 'total' scores assigned to similar essays using PAS?
   - ASAW 'content,' 'organization,' 'language conventions,' 'vocabulary' as well as 'total' scores, and the scores were given to the same samples based on ESL-CP?
   - ASAW 'overall effectiveness' and 'total' scores compated to similar essays using Tests of Written English Scoring Guide (TWE-SG)?

3. To what extent are the raters who used ASAW satisfied with it?
4. Is ASAW an economic scale?

METHOD

The quantitative method was used to test the reliability and concurrent validity of the scores awarded using ASAW and its economy. In addition, both quantitative and qualitative methods were used to examine the raters' satisfaction.

Tasks

Inter-rater reliability may decrease if raters are given written samples with different topics (Weir, 1993), evaluating writing scales on several different topics (Reid, 1990). Therefore, eight similar tasks with different argumentative topics, prompting 300-word argumentative essays in 60 minutes, were developed following the guidelines offered by Bachman and Palmer (1996), Breland et al. (1999), Hamp-Lyons, (1991), Hamp-Lyons (1990), and Horowitz (1991). Three experienced lecturers, who taught the students to write the argumentative essays, were requested to examine the tasks and select only four. They paid particular attention to the wordings and topics of the prompts. Finally, the four selected tasks covered the following topics:

1. Equality of chances for higher education for males and females,
2. Children's free time to be spent on fun or educational activities,
3. Advantages and disadvantages of mass media, and

4. Children starting school at seven or younger age.

Sample

The tasks were given to students (n=167) from six different faculties (Economy & Management, Health & Medicine, Design, Communication, Agriculture, and Ecology) in a public university in Malaysia. The students were mostly female (about 66%) and aged between 19 and 28 ($M = 21$, $SD = 1.3$). Different faculties and students with varying English proficiency levels were selected to obtain samples with diverse writing performance levels. The students provided information like their Malaysian University English Test (MUET) bands. Fifteen anchor papers were selected, three for each of the five performance levels in ASAW. Out of the remaining legible samples, a batch of 110 samples was randomly selected for the reliability and validity tests.

Five raters scored the same batch of samples to test the inter-rater reliability and economy of the scale and the raters' satisfaction with the scale. For all concurrent validity tests and intra-rater reliability tests, a minimum of two raters scored similar samples. The sample size in these tests ranged between 50 and 110. In educational correlation studies, a rough estimate of 30 samples is assumed to be sufficient (Creswell, 2007). Wong (1989) tested her instrument using a sample size of 50 for a similar but less complex purpose.

Raters

Female ESL lecturers (n = 5) with a minimum experience of 12 years in rating and master's or Ph.D. degrees in Teaching English as a Second Language (TESL) were trained to use ASAW. The number of the raters was equal to that of previous studies (Harland, 2003; Wong, 1989). Commonly in assessing essays in high stakes writing tests, two raters are recruited with a third rater re-assessing the essays scored discrepantly by the two raters (Hamp-Lyons, 1990). A higher number of raters was chosen to raise the probability of discrepancy among the raters and thus the accuracy of our measurement. Rater experience affects the reliability of scores (Cumming, 1990), so experienced raters were selected for this study. Additionally, as the raters were supposed to evaluate the scale, they had to have rating experience using similar instruments.

Rater Training

The raters were trained to use ASAW and its anchor papers. Views on rater training vary (Alderson et al., 1995; Shaw, 2002). ASAW and its anchor papers were presented to the raters. The descriptors of different levels were explained using the anchor papers. The raters individually rated five similar samples following ASAW and the anchor papers. The essays had been selected with roughly different levels of performance. The raters compared their scores with others' and discussed discrepancies. The consensus was assumed when a sample was rated at a similar level by all. The sample was reconsidered if a rater scored a level above or below the others' scores. Off-track raters explained their rating approach. Often they found it hard to draw a line between some dimensions, which caused inconsistencies. For example, as they explained the score they had assigned for the 'content' of a sample, the features they mentioned concerned 'form' rather than 'meaning.' Overall agreement was evident regarding the raters' total scores. A similar procedure was repeated for samples written in response to the four different topics. As the training session continued, the raters scored more consistently. Training stopped at this point.

At the end of the training, the raters previewed the questionnaire (Appendix 2). This was important because they had to state how long they took to rate each sample in the questionnaire. They would not record the time if they were unaware of this item. Next, each rater was given a similar batch of argumentative essays (n = 110), anchor papers, mark sheets, and questionnaires. Finally, they were given a week to remote-score the samples individually. A shorter period would cause rater fatigue, while a longer period would affect intra-rater consistency.

Instruments

The instruments included a questionnaire and four other writing scales, ESL-CP (Jacobs et al., 1981), PAS and AQS (Connor & Lauer, 1988) as well as Tests of Written English Scoring Guide, TWE-SG (Educational Testing Service, 2011). A combination of scales was used to account for all the

sub-scales of ASAW to test the concurrent validity of ASAW. The first reference scale was ESL-CP, an established generic analytic scale. It consists of the five sub-scales of 'content,' 'organization,' 'vocabulary,' 'language use,' and 'mechanics,' which correspond with all the sub-scales of ASAW, excluding 'overall effectiveness.' The other two scales were AQS and PAS, both genre-specific instruments. The scores assigned to the essays using these scales were tested for correlation with the ASAW 'content' sub-scale scores awarded to similar essays. The final instrument was TWE-SG, a holistic scale used for rating the writing section of paper-based TOEFL that often has argumentative topics. Brown (2003) tested TWE and TOEFL scores for their relationship and reported high correlations "ranging from 0.57 to 0.69 over 10 test administrations from 1993 to 1995" (pp. 237-238). Studies have supported the high validity of the scale (e.g., Frase et al., 1999; Hale et al., 1996). The instrument includes aspects of argumentative writing like organization, development, task fulfillment, appropriate and detailed support of ideas, cohesion, and coherence, facility in language use, syntactic variety, and appropriate word choice. Therefore, the scores assigned to the samples using this scale were tested for correlation with those assigned to similar samples using the 'overall effectiveness' sub-scale of ASAW and its 'total' scores.

The 'ASAW Evaluation Questionnaire' (Appendix 2) was developed to test the raters' satisfaction with ASAW based on four dimensions of Bachman and Palmer's

(1996) test usefulness, including reliability, validity, impact, and practicality. The questionnaire was a five-point scale Likert-style instrument with 13 items, followed by a short-answer question and a final open-ended question. Items 1 to 3 and 13 were related to the scale impact on the raters at a micro-level. Reliability was addressed by items 6 to 11, among which items 8 to 10 were also related to construct validity as it can be affected by the clarity of the rubrics. Items 4, 5, and 12 dealt with construct validity as well. Finally, item 14 focused on practicality, while item 15 covered all four dimensions.

Data Analysis

SPSS version 16 was used for statistical analyses. Descriptive statistical tests such as means and standard deviations were used. Bivariate correlation tests like Pearson and Spearman were also used to analyze the reliability and concurrent validity tests.

RESULTS AND DISCUSSION

The results are presented and discussed following the research questions in order.

Reliability

The scores collected from the five raters, who remote-scored 110 similar samples, were tested for their inter-rater reliability. In addition, intra-rater reliability was also tested with the help of two raters scoring the same samples at two different intervals.

Inter-rater Reliability

Table 1 shows the Pearson correlation coefficients for the scores assigned to different dimensions of students' argumentative writing performance by different pairs of raters using ASAW.

Table 1

Inter-rater reliability estimates of ASAW sub-scales (Pearson coefficients)

Raters	Content	Organization	Vocabulary	Language conventions	Overall Effectiveness	Total
1 and 2	0.79	0.73	0.84	0.85	0.79	0.84
1 and 3	0.71	0.78	0.80	0.79	0.07*	0.81
1 and 4	0.80	0.73	0.79	0.77	0.75	0.84
1 and 5	0.71	0.70	0.77	0.78	0.65	0.77
2 and 3	0.81	0.83	0.86	0.87	0.10*	0.87
2 and 4	0.78	0.84	0.85	0.86	0.81	0.88
2 and 5	0.82	0.83	0.85	0.91	0.74	0.87
3 and 4	0.77	0.77	0.81	0.83	0.15*	0.84
3 and 5	0.80	0.79	0.81	0.83	0.23*	0.85
4 and 5	0.73	0.79	0.79	0.84	0.71	0.82

*low correlations

According to Farhady et al.'s (2001) guideline, correlation coefficients below 0.50 are regarded as low, 0.50 to 0.75 as moderate, and 0.75 to 0.90 as high. Thus, based on this guideline, the scores indicated moderate to high (r = 0.7-0.9) inter-rater reliability for almost all the sub-scales and raters.

The inter-rater reliability scores showed negligible to low (r = 0.07-0.23) correlations between the scores of the third rater and the others for the sub-scale of 'overall effectiveness.' A follow-up interview with the rater revealed that she had been involved in scoring MUET essays while rating for this study. Therefore, it could be assumed that she scored inconsistently due to rater fatigue. However, her scores for other sub-scales were consistent, so fatigue could not be the real culprit. A more likely reason could be the contrast effect (Grote, 1996), which occurs when a rater scores two different batches of samples using different scales simultaneously or within a short period. Her exposure to the MUET scale and/or samples could have affected the rater's 'overall effectiveness' scores. Probable differences between the rubrics of the two scales may have caused this inconsistency. Another reason could be the 'overall effectiveness' sub-scale itself. An examination of the sub-scale indicates that it covers two different dimensions, including 'style' and 'task fulfilment,'

thus violating the important assumption of unidimensionality that should be met in developing instruments. In instrument development, separate dimensions of a complex construct should be evaluated, focusing on only one attribute at a time (McCoach et al., 2013). Combining the two irrelevant dimensions of 'style' and 'task fulfilment' under one sub-scale seems to have confused the rater.

Inter-rater reliability was also tested about four different topics. Table 2 shows the results of this test.

Table 2

Inter-rater reliability of total scores across topics (Pearson coefficients)

Rater	Topic 1	Topic 2	Topic 3	Topic 4
1 and 2	0.85	0.75	0.84	0.87
1 and 3	0.87	0.72	0.75	0.77
1 and 4	0.75	0.79	0.82	0.82
1 and 5	0.83	0.74	0.74	0.60
2 and 3	0.93	0.92	0.78	0.83
2 and 4	0.84	0.89	0.84	0.71
2 and 5	0.85	0.90	0.81	0.70
3 and 4	0.86	0.84	0.83	0.67
3 and 5	0.85	0.94	0.90	0.67
4 and 5	0.83	0.84	0.76	0.60

The 'total' scores that the raters assigned for the samples indicate moderate to high-reliability coefficients ($r = 0.60$-0.94) for the four topics. Thus, it can prove that the scale can help raters assign fairly reliable scores for essays prompted by varying topics.

Intra-rater Reliability

From the batch of 110 samples, 50 were randomly selected and given to the first and second-raters to be scored after six-week and nine-week intervals, respectively, to test the intra-rater reliability achieved by the raters using ASAW. Various intervals have been suggested in the literature ranging from two weeks (Rohde et al., 2020) to 10 weeks (Kayapınar, 2014). We did not opt for a small interval to allow enough time for a wash-out period. Instead, we tested intra-rater reliability at medium and large intervals of 6 and 9 weeks to ensure that the two raters would forget their first rating experiences. We also went for two different intervals to compare the two raters' reliability scores caused by the intervals. The scores assigned by the raters were tested for correlations with the scores they had previously given to similar samples. Table 3 shows the results of the intra-rater reliability test for each sub-scale.

Table 3

Intra-rater reliability with a time interval of six and nine weeks

Rater	Interval	Content	Organization	Vocabulary	Language conventions	Overall Effectiveness	Total
Rater 1	6 weeks	0.85	0.92	0.90	0.85	0.84	0.92
Rater 2	9 weeks	0.70	0.70	0.74	0.75	0.71	0.77

Based on Farhady et al.'s (2001) guideline, high (r = 0.844-0.92) and almost moderate (r = 0.69-0.77) correlations were found for the first and second-raters, respectively. Furthermore, as indicated by the findings of the previous studies (Kayapınar, 2014; Rohde et al., 2020), a longer period is deemed to reduce the intra-rater reliability (Kayapınar, 2014). Likewise, in the case of our study, the first rater's higher reliability scores suggest that time may negatively affect intra-rater reliability; the longer the interval between the two ratings, the lower the reliability. Admittedly, making such a conclusion based on the scores assigned by only two raters may be questionable. However, since the time interval works as a wash-out period that removes the carry-over effect of the first scoring experience, it sounds logical to argue that a lengthier period will put the rater and the scale in a more difficult position to achieve acceptable intra-rater reliability scores.

Overall, the few unimpressively moderate reliability scores obtained from some of the raters necessitate further refinement of ASAW. It seems particularly true for the 'overall effectiveness' sub-scale that indicated relatively lower reliability scores than other sub-scales.

Concurrent Validity

The scores awarded by the raters to the 110 samples were tested for their correlation with five related measures to test the concurrent validity of ASAW. They included the students' MUET band scores and the scores assigned to their essays using four other established writing scales, including AQS, PAS, ESL-CP, and TWE-SG.

MUET Band Scores

MUET is recognized as a well-established high-stakes testing system in Malaysia. Based on its bands, which indicate students' general proficiency in the English language, decisions are made for Malaysian students' academic future in universities. Therefore, Spearman's rho was used to analyze the correlation between the students' MUET bands and the scores assigned by the five raters to their written samples (Table 4).

Based on Guilford (1973) Rule of Thumb, (>0.20 as Negligible, 0.20-0.40 as Low, 0.40-0.70 as Moderate, 0.70-0.90 as High and 0.90 as Very high correlation strength), moderate ($r_S = 0.63$-0.69) to high ($r_S = 0.73$-0.79) and statistically significant ($p < .01$) correlations were found between the students' MUET bands and the scores assigned to their samples. According to Jacobs et al. (1981), a correlation of 60

Table 4

Correlation test between each rater's scores and students' MUET bands

	Correlation coefficient (r_S)	Significant value (p)
Mean and MUET bands	0.79	.000
Rater1 and MUET bands	0.64	.000
Rater2 and MUET bands	0.69	.000
Rater3 and MUET bands	0.73	.000
Rater4 and MUET bands	0.63	.000
Rater5 and MUET bands	0.74	.000

or above can provide "strong empirical support for the concurrent validity" (pp. 74-75). Therefore, the students' MUET bands strongly support the validity of the assigned scores using ASAW. It should, however, be noted that the students' MUET bands represent their proficiency level in all English language skills. Testing the correlation between their writing scores and MUET bands would not provide a very accurate measure of validity. Therefore, the results of ASAW were also tested for their correlation with those of other instruments that were specifically related to writing or argumentative writing.

AQS

After briefing the first rater on AQS, she used it to remote-score 100 samples selected from the previously scored batch using ASAW. Her scores were collected and tested for correlation with the mean content scores assigned by the five raters for the same samples. Based on the results of Pearson analysis, a moderate (r =0.62) and statistically significant ($p < .01$) correlation was found between the results of AQS and the 'content' sub-scale of ASAW. This

coefficient provides strong empirical support for the concurrent validity of ASAW (Jacobs et al., 1981).

PAS

The first rater was briefed on PAS before using it to remote-score the same batch of 100 samples. These scores were collected and analyzed using Pearson Product-Moment Correlations. A moderate (r = 0.52) and statistically significant ($p < .01$) correlation was found between the results of PAS and the 'content' sub-scale of ASAW. However, the value was below Jacobs et al.'s (1981) threshold (≤ 0.60). The reason could be that PAS evaluates essays based on their persuasive appeals. Thus, it includes rational, credibility, and affective appeals, while ASAW was developed based on the Pyramid of Argumentation (Nimehchisalem, 2010), in which the affective appeal was discarded.

Further analysis showed that in the entire batch of 100 essays, affective appeals occurred only 12 times (4%), as compared with the high frequency of rational (54%) and credibility (42%) appeals (Table 5).

Table 5

Occurrence of rational, credibility, and affective appeals in the samples (n = 100)

Appeal	Minimum	Maximum	Sum	Percentage (%)
Rational	0.00	3.00	144	54
Credibility	0.00	2.00	112	42
Affective	0.00	1.00	12	4
Total			268	100

This incidental finding confirms the difference between argumentative and persuasive modes. While persuasive texts may make frequent appeals to emotions, argumentative texts typically appeal to logic and character (Glenn et al., 2004). It can also be a reason for the lack of a strong correlation between ASAW and PAS scores.

ESL-CP

The first and second-raters were briefed on the ESL-CP before individually using it to remote-score a batch of 50 samples from the samples that they had previously scored using ASAW. The two raters' scores assigned following the ESL-CP sub-scales were recorded with moderate inter-rater reliability coefficients ($r = 0.51$-0.74).

All the scores assigned using ASAW sub-scales (excluding 'overall effectiveness') were tested for their correlation with the scores of their counterpart sub-scales in ESL-CP. Unlike ASAW, ESL-CP has two separate sub-scales for 'grammar' and 'mechanics.' Therefore, the mean scores of these two sub-scales were tested for their correlation with the 'language conventions' sub-scale in ASAW. Table 6 presents the results of Pearson's test of correlation between the sub-scales of the two instruments.

Table 6

Pearson test results between ESL-CP and ASAW scores

Scale and Sub-scales		Rater 1		Rater 2	
ASAW	ESL-CP	r	p	r	p
Content	Content	0.60	.000	0.60	.000
Organization	Organization	0.60	.000	0.60	.000
Language conventions	Grammar and mechanics mean scores	0.65	.000	0.62	.000
Vocabulary	Vocabulary	0.61	.000	0.62	.000
Total	Total	0.72	.000	0.67	.000

Based on Guilford's (1973) Rule of Thumb, the scores given by the raters to the similar batch of samples showed moderate ($r = 0.60\text{-}0.65$) correlations between the four sub-scales of ASAW and ESL-CP. The 'total' scores of the first rater indicated a high correlation with a coefficient of ($r = 0.719$), while the second-raters showed a moderate correlation of ($r = 0.66$). According to Jacobs et al.'s (1981) guideline, these coefficients empirically support the validity of ASAW scores. However, these correlation values are not very impressive, suggesting that there is room for improving the reliability and validity of ASAW.

TWE-SG

The first and second-raters were briefed on the TWE-SG. They used this scale to remote-score 50 of the 110 samples that they had scored using ASAW. The scores that the two raters assigned for the samples following TWE-SG were separately tested for correlation with the 'overall effectiveness' and 'total' scores assigned by each rater for the same samples using ASAW. Table 7 summarizes the results of Spearman's rho analysis for each rater's scores.

Table 7

Correlation test results for ASAW and TWE-SG scores

Rater		Correlation coefficient (r_S)	Significant value (p)
1	Total and TWE-SG	0.77	.000
2	Total and TWE-SG	0.74	.000
1	Overall effectiveness and TWE	0.73	.000
2	Overall effectiveness and TWE	0.66	.000

As the results in Table 7 indicate, coefficients of ($r_S = 0.77$ and 0.74) show high correlations between ASAW 'total' scores and TWE-SG scores given by both raters. The correlation between ASAW 'overall effectiveness' and TWE-SG scores was high for the first rater ($r_S = .73$) but moderate ($r_S = 0.66$) for the second. All the correlations were statistically significant ($p < .01$) and provided strong empirical support for concurrent validity of ASAW ($r_S > 0.6$).

According to the concurrent validity results, the scores awarded using ASAW indicated moderate and high correlations with those assigned using other related instruments. It may be argued that in the present concurrent validity tests, the reference instruments had been developed for different test settings and varying purposes. At the same time, some were generic (e.g., ESL-CP), others focused on different features. For example, PAS

evaluated emotional persuasive appeals that were not covered by ASAW, which resulted in moderate correlations (0.52) between the results of the two scales. Such variations lead to different descriptors, which may result in different scores and ultimately in low correlations. However, a higher correlation was expected from concurrent validity tests between AQS and ASAW 'content' sub-scale. The moderate correlation (r = 0.62) between the two instruments will lead most scale developers to doubt the validity of the new instrument.

However, it may be argued that these results are acceptable because the scale was not developed for high-stakes testing purposes.

Raters' Satisfaction

After working with ASAW, the raters evaluated its usefulness in a questionnaire (Appendix 2). The data were collected and analyzed to find out how they evaluated ASAW. Table 8 presents the results of this analysis.

Table 8

Raters' satisfaction with ASAW

Rater	Total score (upon 65)	Percentage (%)
1	62	95
2	59	91
3	41	63
4	50	77
5	26	40
Average	47.6	73

The raters had different views. At the same time, the first and second-raters found ASAW 'very highly' useful (91% & 95%), the other three rated it as a 'highly' (77%) or 'moderately' useful scale (40% & 60%). On average, the scale was rated as rather highly useful (73%). Additionally, analysis of the qualitative data elicited by the open-ended question (item 15) at the end of the questionnaire showed that almost all the raters agreed on:

1. re-wording the descriptors of the 'content' sub-scale as they believed terms like 'data' and 'warrant' might confuse novice raters.

2. separating 'overall effectiveness' into two separate sub-scales of 'style' and 'task fulfilment' as they were two separate writing features.

Refining ASAW based on these two suggestions may result in better evaluation results. Even though they had been trained and briefed on all the scale descriptors, the raters in this study may have been confused by the rather technical terms in the 'content'

sub-scale. In addition, as discussed earlier, it is important that each domain of an instrument must be unidimensional and focus on a single construct at a time.

Economy

Each rater stated how long it took her to score the whole batch of 110 samples using ASAW (Table 9).

Table 9

Time spent scoring essays

Rater	Overall evaluation time for 110 essays (hours)	Average evaluation time for each essay (minutes)	Essays per hour
1	18	9.8	6.1
2	6	3.3	18.3
3	7.5	4.1	14.7
4	7	3.8	15.7
5	12	6.5	9.2
Average	10.1	5.5	12.8

While the first rater was the slowest and the second was the fastest in scoring the samples, the other three had fairly reasonable ratings. On average, each rater took 5.5 minutes to rate a sample about 13 samples per hour. This time is about twice as much as the time spent by the raters in Wong (1989), in which scoring each sample only took an average of 2½ minutes. However, the samples in Wong's study were stories composed of only ten sentences, whereas in this study, some samples included argumentative essays of over 540 words. In Glasswell and Brown's (2003) study, an average scoring rate of about seven samples per hour was reported for rating samples, markedly lower than the average number of samples scored per hour (almost 13) using ASAW. Therefore, it can be concluded that ASAW is economical in terms of the time required to score papers.

CONCLUSION

The literature on ASAW shows it was developed based on multiple sources and methods. Developing rating scale descriptors based on the analysis of students' written samples has been recommended in the literature as an empirical method (Fulcher & Davidson, 2007). It reduces the problem of assigning unfairly low scores to learners who respond taking unusual perspectives (Odell, 1981) and helps evaluation of students' writing work best (Hamp-Lyons, 1990). Additionally, determining the evaluative criteria of the scale based on quantitative and qualitative data may contribute to its *a priori* validity.

The results of this study provide information on the reliability, concurrent validity, consequential validity, and economy of the instrument from *a posteriori* perspective. The results indicated moderate to high reliability and concurrent validity of the scores assigned using ASAW. The raters who used the scale indicated high average levels of satisfaction with it, although they did not consider it completely flawless. The scale also proved to be relatively economical.

IMPLICATIONS

The present study has both theoretical and practical contributions. From a theoretical perspective, the findings confirmed the accuracy of the Pyramid of Argumentation (Nimehchisalem, 2010) in discarding the emotional appeal. 'Argumentation' and 'persuasion' are commonly used interchangeably (e.g., Cohen, 1994). However, although the terms are similar, they are not synonymous (Hall & Birkerts, 2007). It has been argued that, unlike argumentation, persuasion involves appeals to emotion (Glenn et al., 2004). The analysis of the argumentative essays in this study showed that emotional appeals were rarely made. Our findings lead us to draw a line between the two terms. Therefore, discarding the emotional appeal on an argumentative scale seems appropriate. Indeed, its presence would have unfairly penalized the students who did not use it, decreasing the scale's construct validity. The theoretical framework based on which

ASAW was developed can be a useful model in assessing argumentative essays.

The results indicated the raters' overall satisfaction with ASAW. Due to the small sample size, further research is required on the instrument's usefulness before making any generalizations. However, it cannot be denied that as an analytic scale, ASAW can be regarded as a useful tool for diagnosing ESL students' difficulties in writing argumentative essays. It can provide predictive as well as retrospective information for assessing the effectiveness of their writing courses. It is of particular importance in the educational context of today with its increasing emphasis on accountability. As is the case in most parts of the world, in Malaysia, ESL writing is a problematic area of English language teaching (Pandian, 2006). Malaysian students often lack the essential writing skills to meet academic literacy requirements at university (Nambiar, 2007; Ramaiah, 1997), reporting high levels of ESL writing anxiety (Nor et al., 2005). Although Malaysian practitioners are aware of the advantages of approaches like the genre-based instruction of writing (Hajibah, 2004; Zuraidah & Melor, 2004), they indicate unacceptable levels of their learners' argumentative writing ability (Rashid & Chan, 2008). At least in part, this problem may be due to the unprofessional ESL writing assessment methods practiced in Malaysian universities (Kho, 2006; Tan et al., 2006). Impressionistic scoring is typically practiced for assessing students' writing in Malaysian universities (Mukundan & Ahour, 2009). Developing

instruments like ASAW is practically a step forward in professionalizing language instructors in assessing writing from a local perspective.

Finally, ASAW can help ESL writing researchers and teachers develop self-assessment and peer feedback checklists. After making some modifications to the scale, they can customize it for the learners in their teaching-learning context (e.g., Vasu et al., 2018). ASAW has already proved a useful model for developing self-assessment checklist developers (Vasu et al., 2020) by reducing the teacher's workload and promoting the student's self-regulation and learner autonomy. It can also serve as a useful model in developing checklists that help student writers provide feedback for their peers' argumentative essays.

LIMITATIONS

The reliability test results indicated that one of the raters' scores was markedly inconsistent with others'. The case highlights the importance of factors that can result in rating errors. No matter how rigorously a scale is developed, rating errors (Grote, 1996) and unsystematic administration can result unreliable results. In addition, it was found that the 'overall effectiveness' sub-scale is not unidimensional. Instead, it mixed 'style' and 'task fulfillment,' which resulted in one of the raters' very low inter-rater reliability. According to the developers of ASAW, in the first focus group study, 'style' and 'task fulfillment' were two separate sub-scales (Nimehchisalem & Mukundan, 2012). The two sub-scales

collapsed after the focus group reconvened for two reasons: giving a holistic look to ASAW and enhancing its economy (Nimehchisalem, 2010). However, based on the present study's findings, keeping the two dimensions separate seems necessary.

More research in a broader group of stakeholders on the consequential validity of the instrument also seems necessary. The sub-scale of 'overall effectiveness' need further revision and trial. Rater training and rating experience seem to contribute to scores and the rating process (Barkaoui, 2010). Testing the scale with the help of novice or untrained raters may result in more useful findings. As mentioned earlier, in the development process of ASAW, multivariate analysis methods such as Exploratory Factor Analysis (EFA) were used, the results of which have already been published by Nimehchisalem and Mukundan (2011). More studies on the ASAW which adopt item response theory (IRT) (also referred to as latent trait theory) can have more illuminating results. Likewise, further research that focuses on cognitive processes used by raters while employing ASAW and how it influenced their decision-making involved in this process could result in interesting findings.

ACKNOWLEDGEMENT

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74, https://doi.org/10.1080/15434300903464418

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework.* http://www.nocheating.org/Media/Research/pdf/RR-99-03-Breland.pdf

Brennan, R., Kim, J., Wenz-Gross, M., & Siperstein, G. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review*, *71*(2), 173-216. https://doi.org/10.17763/haer.71.2.v51n6503372t4578

Brown, G. T. L., Glasswell, K., & Harland, D. (2004) Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing 9*, 105-121. https://doi.org/10.1016/j.asw.2004.07.001

Brown, H. D. (2003). *Language assessment principles and classroom practices*. Longman.

Cohen, A. D. (1994). *Assessing language ability in the classroom*. Heinle & Heinle Publishers.

Connor, U., & Lauer, J. (1988). Cross-cultural variation in persuasive student writing. In A.C. Purves (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 206-227)*.* Sage.

Cooper, C. R. (1999). What we know about genres, and how it can help us assign and evaluate writing. In C. R. Cooper & L. Odell, (Eds.), *Evaluating writing: The role of teacher's knowledge about text, learning, and culture* (pp. 23-52). National Council of Teachers of English (NCTE).

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Sage.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). American Council on Education.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51. https://doi.org/10.1177/026553229000700104

Educational Testing Service. (2011). *Writing scoring guide*. http://www.ets.org/toefl/pbt/scores/writing_score_guide/

Farhady, H., Jafarpur, A., & Birjandi, P. (2001). *Testing language skills: From theory to practice.* SAMT.

Frase, R., Faletti, J., Ginther, A., & Grant, L. A. (1999). *Computer analysis of the TOEFL test of written English* (TOEFL Research Report #64). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01791.x

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Rutledge. https://doi.org/10.4324/9780203449066

Glasswell, K., & Brown, G. T. L. (2003). *Accuracy in the scoring of writing: Study in large-scale scoring of asTTle writing assessments* (asTTle Technical Report 26). University of Auckland/Ministry of Education.

Glasswell, K., Parr, J., & Aikman, M. (2001). *Development of the asTTle writing assessment rubrics for scoring extended writing tasks* (Technical Report 6, *Project asTTle*). University of Auckland/Ministry of Education.

Glenn, C., Miller, R. K., Webb, S. S., Gary, L., & Hodge, J. C. (2004). *Hodges' harbrace handbook* (15th ed.). Thompson Heinle.

Grote, R. D. (1996). *The complete guide to performance appraisal.* AMACOM Books.

Guilford, J. P. (1973). *Fundamental statistics in psychology and education* (5th ed.). McGraw-Hill.

Hajibah, O. (2004). Genre-based Instruction for ESP. *The English Teacher, 33*, 13-29.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs* (TPOEFL Research Report #54). Educational Testing Service.

Hall, D., & Birkerts, S. (2007). *Writing well*. Longman.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge University Press.

Hamp-Lyons, L. (1991). Pre-text: Task-related influences on the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 87-107). Ablex. https://doi.org/10.1017/CBO9781139524551.009

Harland, D. (2003). *Using asTTle persuasive writing: A case study of teaching argument writing* (asTTle Technical Report, #29). University of Auckland/Ministry of Education.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 71-85). Ablex.

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V. F., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Newbury House Publishers.

Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research, 57*, 113-136 http://dx.doi.org/10.14689/ejer.2014.57.2

Kho, S. J. (2006). *Assessment criteria in a holistic scoring scale as a pedagogical tool in teaching English language argumentative writing in a Malaysian secondary school* [Unpublished Master's thesis]. Universiti Teknologi Malaysia.

Kinneavy, J. A. (1971). *A theory of discourse: The aims of discourse*. Prentice Hall.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain: School and corporate applications.* Springer. https://doi.org/10.1007/978-1-4614-7135-6

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). Macmillan.

Mukundan, J., & Ahour, T. (2009). Perceptions of Malaysian school and university ESL instructors on writing assessment. *Journal Sastra Inggris, 9*(1), 1-21.

Nambiar, R. M. K. (2007). Enhancing academic literacy among tertiary learners: A Malaysian experience. *3L Journal of Language Teaching, Linguistics and Literature, 13*, 1-21.

Nimehchisalem, V. (2010). *Developing an analytic scale for argumentative writing of students in a Malaysian public university* [Unpublished Doctoral dissertation]. Universiti Putra Malaysia.

Nimehchisalem, V. (2018). Pyramid of argumentation: Towards an integrated model for teaching and assessing ESL writing. *Language & Communication, 5*(2), 185-200.

Nimehchisalem, V., & Mukundan, J. (2011). Determining the evaluative criteria of an argumentative writing scale. *English Language Teaching, 4*(1), 58-69. https://doi.org/10.5539/elt.v4n1p58

Nimehchisalem, V., & Mukundan, J. (2013). Development of the content subscale of the Analytic Scale of Argumentative Writing

(ASAW). *Pertanika Journal of Social Sciences & Humanities, 21*(1), 85-104.

Nimehchisalem, V., Mukundan, J., & Shameem, R. G. (2012). Developing an argumentative writing scale. *Pertanika Journal of Social Sciences & Humanities*, *20*(S), 185-204.

Nor, S. M. D., Nuraihan, M. D., & Noor L. A. K. (2005). Second language writing anxiety: Cause or effect? *A Journal of the Malaysian English Language Association (MELTA), 1*, 1-19.

Odell, L. (1981). Defining and assessing competence in writing. In C. R. Cooper, (Ed.), *The nature and measurement of competency in English* (pp. 95-138). National Council of Teachers of English.

Pandian, A. (2006). What works in the classroom? Promoting literacy practices in English. *3L Journal* of *Language Teaching, Linguistics and Literature, 11*, 1-25.

Ramaiah, M. (1997). *Reciprocal teaching in enhancing the reading ability of ESL students at the tertiary level* [Unpublished Doctoral thesis]. University of Malaya.

Rashid, S. M., & Chan, H. H. (2008). Exploring the interplay of mode of discourse and proficiency level in ESL writing performance: Implications for testing. *The English Teacher, 37*, 105-122.

Reid, M. J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191-210). Cambridge University Press. https://doi.org/10.1080/09638288.2020.1776774

Rohde, A., McCracken, M., Worrall, L., Farrell, A., O'Halloran, R., Godecke, E., David, M., & Doi, S. A. (2020). Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test. *Disability and rehabilitation*, 1-9. https://doi.org/10.1080/09638288.2020.1776774

Shaw, S. (2002). The effect of standardization on rater judgment and inter-rater reliability. *Cambridge ESOL Research Notes 8*, 13-17.

Tan, B. H., Emerson, L., & White, C. (2006). Reforming ESL writing instruction in tertiary education: The writing center approach. *The English Teacher, 34*, 1-14.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Vasu, K., Nimehchisalem, V., Fung, Y. M., & Rashid, S. M. (2018). The usefulness and effectiveness of argumentative writing self-assessment checklist in undergraduate writing classrooms. *International Journal of Academic Research in Business and Social Sciences, 8*(4), 202-219. https://doi.org/10.6007/IJARBSS/v8-i4/4008

Vasu, K., Yong M. F., Nimehchisalem, V., & Rashid, S. M. (2020). Self-regulated learning development in undergraduate ESL writing classrooms: Teacher feedback vs self-assessment. *RELC Journal, 51*(3). https://doi.org/10.1177/0033688220957782

Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall International.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave MacMillan.

Wong, H. (1989). *The development of a qualitative writing scale.* University Kebangsaan Malaysia Press.

Yeh, S. S. (1998). Validation of a scheme for assessing argumentative writing of middle school students. *Assessing Writing, 5*(1), 123-150. https://doi.org/10.1016/S1075-2935(99)80009-9

Zuraidah, A., & Melor M. Y. (2004). An ESL writing course: Unraveling students' needs and concerns. *The English Teacher, 33*, 120-132.

APPENDICES

Appendix 1

Analytic Scale of Argumentative Writing (ASAW)

Score	1. Content	Grade (level)
15-20	Effectively introduces the claim(s), maturely provides an in-depth or extensive account of relevant data supporting the claim(s), backs the warrants, accounts for rebuttals, and may employ qualifiers	A (Excellent)
12-14	Presents a reasonably mature and extensive account of relevant claims and data but at times lacks adequate backing	B (Competent)
10-11	Presents relevant claims and data, but the data sound immature, and are not well-elaborated	C (Modest)
8-9	Presents claims, data, warrants and backings, some of which may be irrelevant	D (Basic)
0-7	No response Or only makes a number of claims, some of which may be irrelevant	F (Very limited)

Score	2. Organization	Grade (level)
15-20	Well-organized introduction/narration/division, body and conclusion; sentences skillfully linked; an internal logic is clearly showing writer's purpose and flow of ideas	A (Excellent)
12-14	Reasonably well-arranged introduction, confirmation, and conclusion; sentences connected reasonably well; sometimes hard to follow the line of thought because of the gaps between a few ideas	B (Competent)
10-11	Introduction/conclusion: brief/lacking; despite certain redundant ideas, easy to follow writer's line of thought and purpose; sentences linked well but cases of wrong connections evident	C (Modest)
8-9	No introduction/conclusion; evidence of some basic form of cohesion but in case of complicated ideas, lack of cohesion; despite a few incoherent sentences, a simple pattern of thought evident	D (Basic)
0-7	Lacking an introduction/conclusion; no/vain attempts to create cohesion; OR no response	F (Very limited)

Score	3. Vocabulary	Grade (level)
15-20	Appropriate use of simple-complex/technical words, phrases, collocations, idioms, or figures of speech; few incorrect forms; skillful use of synonyms/antonyms to avoid repetition	A (Excellent)
12-14	Occasional incorrect word forms, phrases, or collocations; mostly using simple words; using synonyms/antonyms to avoid repetition but still a few repeated words	B (Competent)
10-11	Incorrect word forms, phrases, or collocations in almost every sentence, sometimes even lacking simple words to communicate, OR repeating the same words throughout the essay	C (Modest)
8-9	Incorrect word forms, phrases, or collocations in almost all sentences	D (Basic)
0-7	No response or a collection of irrelevant words	F (Very limited)

Score	4. Language conventions	Grade (level)
15-20	Few negligible slips; a variety of simple-complex structures; form getting meaning across very skillfully, very skillful control over spelling, capitalization, and punctuation	A (Excellent)
12-14	Occasional errors; mostly simple structures; form still getting meaning across, occasional spelling, capitalization, or punctuation problems not blurring the meaning	B (Competent)
10-11	Almost one error every other sentence; form blurring meaning sometimes, some spelling, capitalization, or punctuation problems blurring meaning, spelling, capitalization, or punctuation problems in almost all sentences blurring the meaning	C (Modest)
8-9	A collection of garbled sentences and fragments, confusing rather than communicating	D (Basic)
0-7	No response/fragments; spelling, capitalization/punctuation problems in almost all the essay	F (Very limited)

Score	5. Overall effectiveness	Grade (level)
15-20	Very skillful and effective presentation and justification of arguments through a highly engaging, correct, clear, appropriate and/or ornate style; task requirements skillfully fulfilled; written well over the word limit	A (Excellent)
12-14	Effectively presenting and justifying arguments through a reasonably engaging, correct, clear, and appropriate style; task still fulfilled reasonably well; written over/to the word limit	B (Competent)
10-11	A reasonable ability to present arguments but through a simple, fairly correct, clear, and appropriate style, task requirements are almost fulfilled; written around the word limit	C (Modest)
8-9	Lacking a reasonable ability in presenting arguments through a monotonous, usually incorrect, unclear, and inappropriate style; task partially fulfilled; written below the word limit	D (Basic)
0-7	No ability to present arguments; incorrect, unclear, and inappropriate style; a task not fulfilled; written far below the word limit	F (Very limited)

ASAW Score Convertor

ASAW Scores	University Mark	University Grade	University Value
16-20	80-100	A	4.00
15	75-79	A-	3.75
14	70-74	B+	3.50
13	65-69	B	3.00
12	60-64	B-	2.75
11	55-59	C+	2.50
10	50-54	C	2.00
9.5	47-49	C-	1.75
9	44-46	D+	1.50
8	40-43	D	1.00
0-7	0-39	F	0

Appendix 2

Analytic Scale of Argumentative Writing Evaluation Questionnaire

This questionnaire has been developed to evaluate the Analytic Scale of Argumentative Writing based on your judgment of its quality. Assess the scale by marking the numerical values next to each statement below that best describe your evaluation of it:

1. Strongly disagree
2. Disagree
3. Unsure
4. Agree
5. Strongly agree

The questionnaire also consists of three open-ended questions at the end (Questions 14-16) that you are requested to answer.

	Item	1	2	3	4	5	Comments
1.	I found it easy to work with the scale.						
2.	I will use this scale to correct my own students' written works.						
3.	I recommend using this scale with my colleagues.						
4.	The scale fully covers the aspects of argumentative writing skills.						
5.	The scale assesses an adequate scope of writing construct.						
6.	The scores produced by the scale distinguish learners' levels.						
7.	The scale helped me draw a clear line between the essays that seemed to be of different levels.						
8.	All the terms in the scale are clear and easy to understand.						
9.	The sample scripts helped me get a grip of the different levels of performance.						
10.	The scoring guideline is clear and leaves no concept vague.						
11.	Overall, the scale sounds like a reliable instrument.						
12.	Weighting of different aspects of writing is fair.						
13.	Overall, I am satisfied with this scale.						
14.	On average, it took me ……… minutes to score a single essay.						
15.	I think the scale can be improved by						

SOCIAL SCIENCES & HUMANITIES

Towards a CEFR Framework for Workplace Communication: Students' Perceptions of the Sub-Skills, Use and Importance of Language Productive Skills (LPS)

Ahmad Mazli Muhammad[1]*, Maisarah Ahmad Kamil[2] and Zachariah Aidin Druckman[3]

Department of English Language and Linguistics, Akademi Pengajian Bahasa, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

ABSTRACT

The ever-changing demands of the workforce due to current trends have led to the need for universities to equip their graduates with the necessary soft skills to increase their employability. As a result, the implementation of CEFR in language curricula was emphasised to address this matter. However, research on how CEFR could be implemented into a university's workplace communication course is severely lacking. Moreover, there is room to further enhance existing CEFR frameworks for workplace communication. Thus, this preliminary study was conducted to investigate students' perceptions of the use and importance of language productive skills (LPS) at the workplace towards developing a CEFR framework for workplace communication. The study adopted the quantitative approach through questionnaires to gauge students' perceptions of the use and importance of LPS at the workplace. A total of 354 students from various faculties under the clusters of science and technology, business and management, and social sciences and humanities participated in the study. The responses were analysed using the Statistical Package for Social Sciences (SPSS). The study's findings show that, generally, students' perceptions regarding the use and importance of speaking skills in the workplace are congruent to the CEFR scale for formal discussions. However, the use and importance of writing skills do not match the current available scale under CEFR to cater

*E-mail address*es:
a.mazli@uitm.edu.my (Ahmad Mazli Muhammad)
maisarah@uitm.edu.my (Maisarah Ahmad Kamil)
zachariah@uitm.edu.my (Zachariah Aidin Druckman)
*Corresponding author

to workplace communication. Thus, future research calls for curriculum developers to identify relevant descriptors needed for written workplace communication.

INTRODUCTION

Graduate employability and the increasing need to set higher standards in university curricula has been well acknowledged by the Ministry of Education in Malaysia. However, past studies have shown that the English proficiency level of new graduates in Malaysia is a high concern, particularly regarding poor communication skills (Agus et al., 2011). The current situation is severe enough that universities in Malaysia have been subject to criticism in producing graduates with a low level of English proficiency, which has made it difficult for the students to market themselves to join companies and businesses (Dzulkifly, 2018). Even more concerning is that industries in Malaysia have also made it clear that they would not hire graduates who do not meet the minimum level of language proficiency required (Sarudin et al., 2013).

In 2003, the English Language Standards and Quality Council (ELSQC) was established in Malaysia, which led to the implementation of the Common European Framework of Reference for Languages (CEFR) to boost Malaysian education to international standards (Hazita Azman, 2016, as in Uri & Aziz, 2018).

This initiative was part of the plan under the Malaysian Education Blueprint 2013–2025, which highlighted poor English proficiency as one of the top five issues faced by Malaysian graduates, which needed to be given considerable attention as deliberated further in the second shift (Malaysian Ministry of Blueprint, 2013). The adoption of CEFR into the education system, however, has been gradual. For instance, CEFR was adopted in phases whereby the first phase (from 2013 to 2015) focused on teachers' levels of English proficiency. The second phase (2016) sought to match the education level from pre-school to teacher education against the CEFR standards, while the third phase concerns ELSQC's role to evaluate, review and revise the implementation of CEFR (Foley, 2019). Thus, as CEFR has been gradually adopted into the design of courses and assessments, with its prevalence becoming clearer in recent years, the implementation of the standards in Malaysian schools and universities is still difficult to gauge.

CEFR is a set of scales that are used to describe users as Basic (A1, A2), Independent (B1, B2) and Proficient (C1, C2). It is distinguished by its 'can do' design which describes the extent to which language users can demonstrate their abilities rather than focus on the deficiency of their skills. It is the most widely adopted language proficiency framework worldwide, and its use is relevant for the design and development of language policies, curricula, and assessments in many parts of the world (Foley, 2019). The CEFR framework was recently updated in 2018, signalling new

and is still undergoing much research and progress. However, it should be emphasised that the framework was not designed as a standardising tool; rather, it is a tool that can be used to facilitate curriculum design and development and does not focus on what practitioners need to do or even how to do it (Council of Europe, 2001). Thus, in the context of countries' courses and examinations, the learning and assessments designed may be guided by CEFR but must ultimately be based on what the learners should do in the target language in their context (Foley, 2019).

Over the past several years, much research has been conducted on the design and development of courses that align with the proficiency standards of CEFR. According to Harsch and Seyferth (2020), one challenge faced by language course providers is shifting from institution and educator-defined tests aligned to current education standards to tests aligned to an internationally recognised framework. However, in designing courses that match the current education standards and an internationally recognised framework, there is also a dire need to align such course designs to the learners' current needs and the industry. Thus, there is a need for such standards to reflect the industry's current needs and practices.

LITERATURE REVIEW

The utilisation of CEFR as a proficiency scale for curriculum development has not escaped criticism in current research. One significant criticism raised by Barni (2015), for instance, was highlighting that the use of CEFR has led policymakers to use the proficiency level to impose gatekeeping strategies without conducting a thorough needs analysis. This form of needs analysis for curriculum development, especially pertaining to understanding and meeting the needs of the industry, has been implied in the Malaysian Education Blueprint 2015-2025, as the blueprint emphasises the need for universities to work with the industry for better curriculum design and delivery (Mustafa, 2019). Furthermore, according to the Malaysian Qualifications Agency (MQA), as outlined in the Programme Standards Language (Malaysian Qualifications Agency, 2018), higher education providers are obligated to conduct regular curriculum reviews by engaging professional bodies, government agencies and the industry.

According to Hutchinson and Waters (1987), for an effective course design to take place, there is a need to focus on identifying the needs of learners and the needs of the industry. It is important, as the aim of a language course should be to uncover the competence level and how a person can acquire that competence. Thus, there is a need first to engage the learners to understand their perspectives and thoughts of the current curriculum, what they foresee may be useful in the future, and where they currently stand, as this will help inform the university of the changes that may be necessary to be done on the existing curriculum. Thus, the first step to the learning-centred approach to course design is Hutchinson and Waters (1987) in Figure 1.
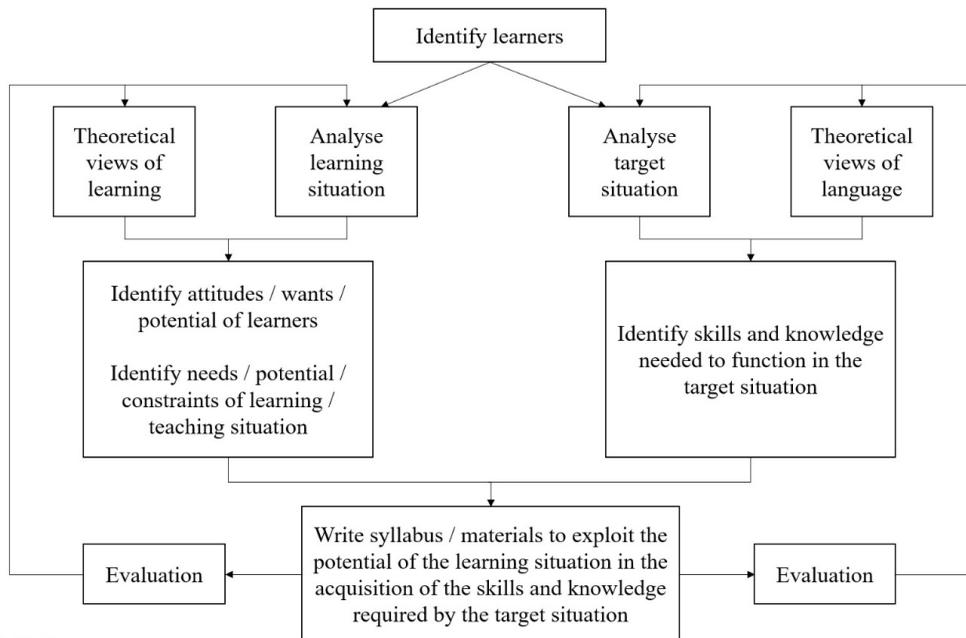
Figure 1. A learning-centred approach to course design (Hutchinson & Waters, 1987, p. 74)

Based on Figure 1, one of the first steps to a learning-centred approach to course design is to understand the views of learning, the learning situation, as well as the attitudes, wants, and potential of the learners, along with possible constraints in the learning or teaching situation. In addition, it highlights the crucial role that the learners play in the curriculum design, which has not been fully addressed in the Programme Standards Language set by the MQA as the programme standards only emphasised the need to engage professional bodies, government agencies, and the industry (Malaysian Qualifications Agency, 2018).

Numerous needs analyses have been conducted to understand better the language and communication needs of employers in Malaysia to address the challenge of language proficiency affecting graduate employability. Past studies have looked at the importance of the English language for employment (Sarudin et al., 2013; Tajuddin et al., 2015; Zainuddin et al., 2019) as well as specific needs of the industry (Hee & Zainal, 2018; Isnin et al., 2018; Perinpasingam et al., 2015). Past needs analyses have looked into the skills and subskills required to communicate well in the context of workplace and professional communication.

However, while many past studies focused on the needs of employers, very few studies have looked at the perspectives of students in particular to understand their viewpoints and challenges, which is a criticism that has been given by

Hutchinson and Waters (1987) on the practice of conducting a needs analysis. For instance, Tajuddin (2015), who conducted a qualitative study, found that for speaking skills in the professional context, the main requirement is the ability for graduates to contribute to productive and appropriate verbal interactions. On the other hand, for writing, the main requirement is to contribute to the effective execution of tasks at work and make the workflow efficient. However, this study was conducted via interviews with three stakeholders: employers from Malaysian companies, representatives from a couple of ministries in Malaysia, and lecturers from three universities.

According to Hutchinson and Waters (1987), for universities to design a curriculum that can meet the needs of both

students and the industry, there is a high need to analyse the needs of students in light of the target situations where such required skills will be used. Thus, this study was conducted to understand the learners' perspectives on the importance and perceived use of language productive skills in the workplace. Additionally, this study takes a step further to compare the stated skills against the current CEFR scales for speaking and writing as a preliminary study towards the development of a CEFR framework for workplace communication.

In the context of this study, the CEFR scale that is considered most relevant to workplace communication is the CEFR speaking scale for formal discussion and meetings, as depicted in Table 1.

Table 1

CEFR Speaking Scale for Formal Discussions and Meetings (Council of Europe, 2001)

	FORMAL DISCUSSIONS AND MEETINGS
C2	Can hold his/her own in a formal discussion of complex issues, putting an articulate and persuasive argument at no disadvantage to native speakers.
C1	Can easily keep up with the debate, even on abstract, complex, unfamiliar topics. Can argue a formal position convincingly, responding to questions and comments and answering complex lines of counterargument fluently, spontaneously and appropriately.
B2	Can keep up with an animated discussion, identifying arguments supporting and opposing points of view accurately. Can express his/her ideas and opinions with precision, present and respond to complex lines of argument convincingly.

Table 1 *(Continued)*

	FORMAL DISCUSSIONS AND MEETINGS
B2	Can participate actively in routine and non-routine formal discussion. Can follow the discussion on matters related to his/her field, understand in detail the points given prominence by the speaker. Can contribute, account for and sustain his/her opinion, evaluate alternative proposals and make and respond to hypotheses.
B1	Can follow much of what is said related to his/her field, provided interlocutors avoid very idiomatic usage and articulate clearly. Can put over a point of view clearly, but has difficulty engaging in debate. Can take part in a routine formal discussion of familiar subjects conducted in a clearly articulated speech in the standard dialect and involves the exchange of factual information, receiving instructions or the discussion of solutions to practical problems.
A2	Can generally follow topic changes in formal discussion related to his/her field, which is conducted slowly and clearly. Can exchange relevant information and give his/her opinion on practical problems when asked directly, provided he/she receives some help with formulation and can ask for repetition of key points if necessary. Can say what he/she thinks about things when addressed directly in a formal meeting, provided he/she can ask for repetition of key points if necessary.
A1	No descriptor is available.

One point of interest that should be noted here is that there does not seem to be an existing CEFR scale for written communication in the context of formal or workplace/professional communication. Thus, the following research objectives were formed, and the research questions were constructed as a preliminary step to close this identified gap.

Research Objectives

1. To identify students' perceptions of the importance of language productive skills for employability.

2. To identify students' perceptions of the most important language productive sub-skills needed at the workplace.

3. To evaluate the sufficiency of the CEFR framework to test the identified language productive skills and subskills.

Research Questions

1. What are students' perceptions regarding the importance of language productive skills for employability?

2. What are students' perceptions of the most important language productive sub-skills needed at the workplace?
3. How accommodating is the current CEFR framework in testing the identified language productive skills and subskills?

MATERIALS AND METHODS

The present study utilises the quantitative approach utilising survey questionnaires to identify the relationship between the independent and dependent variables (Labaree, 2009). The focus of the quantitative approach the application of scientific methods in the collection of data, which constitutes the possibility of generalisation based on the samples (Daniel, 2016). The questionnaire was adopted and adapted from the syllabus of a course called English for Professional Interaction offered at a Malaysian public university. The questionnaire items were formulated based on the course content encompassing forms of communication, language functions for interpersonal communication and workplace interaction, and considerations for professional interaction (Akademi Pengajian Bahasa, 2016).

From that, the study adopts the descriptive research design, which involves making detailed descriptions of the phenomena being studied (Singh et al., 2015, p. 111). As for the population and samples concerned, the population refers to Bachelor Degree students of Malaysia where 354 samples were selected via simple random sampling where 86.2% are from Public Universities (UA) and 13.8%

from Private Institutions (13.8%). These respondents range from Year 1, Year 2, Year 3, Year 4, Year 5, and above. Furthermore, the respondents' fields of study are separated into three different fields – science and technology, social sciences and humanities, and business and administration.

Regarding their working experience, 61.3% of the respondents have had working experience, while the remaining 38.7% do not. Those who have had working experience claimed to have worked between five months or less to more than two years in a variety of working fields, specifically oil and gas, retail, self-employed, food and beverages, corporate, recruitment, human resources, education, fitness and sports, film, performing arts, building, property, engineering, medical and health, photography, information technology, hotel and tourism, accountancy and finance, laboratories, delivery services, customer service, call centres, attachment, and manufacturing.

An online survey questionnaire was self-administered to the samples via Google Forms comprising nominal, ordinal and mainly Likert scales (Singh et al., 2009). The application of the Likert scale is to measure the respondents' attitudes in terms of their agreement or disagreement based on the items (Albaum, 1997). Therefore, it is essential in analysing the data for inferential statistics (Singh et al., 2009). The data was then collected and proceeded for analysis. Inferential statistics were utilised for the present study, specifically frequency statistics, descriptive statistics, independent samples t-tests, and the one-way analysis of variance (ANOVA). The findings were then compared to two CEFR scales that seemed

to be the most suitable for the language productive skills for formal communication in the context of the workplace, which is the Formal Discussion (Meetings) scale and the Overall Written Interaction scale from the document "Common European Framework of Reference for Languages: Learning, teaching, assessment" (Council of Europe, 2001).

RESULTS

Research Question 1–What are the students' perceptions regarding the importance of language productive skills for employability

Table 2 shows the descriptive statistics on the students' perceptions regarding the importance of language productive skills

(LPS) for employability, and their perception of the university's curriculum in preparing them with the skills. For example, the mean score for item "Language productive skills are important for future employability" recorded M=4.64 (SD=0.557). In contrast, for item "The university curriculum prepares students to attain sufficient language productive skills," recorded M=4.64 (SD=0.841). Furthermore, the students were asked about their confidence in the sufficiency of their LPS for the workplace in item "I am confident that my language productive skills are sufficient for the workplace," which recorded a mean score of M=3.72, SD=0.763.

Table 2

Importance of language productive skills for employability

	M	SD
Language productive skills are important for future employability	4.64	.557
The university curriculum prepares students to attain sufficient language productive skills.	3.98	.841
I am confident that my language productive skills are sufficient for the workplace.	3.72	.763

1 – Strongly Disagree, 2 – Disagree, 3 – Slightly Agree, 4 – Agree, 5 – Strongly Agree

Additionally, independent samples t-tests were conducted to test the mean differences with all three items in Table 3b based on the respondents' educational institutions – public universities (UA) and private institutions (US); and their working experience. The results are as follows:

Table 3a

Mean Comparisons between UA and US

	Edu. Ins.	M	SD
Language productive skills are important for future employability	UA	4.64	0.562
	US	4.61	0.533
The university curriculum prepares students to attain sufficient language productive skills.	UA	3.98	0.843
	US	3.94	0.841
I am confident that my language productive skills are sufficient for the workplace.	UA	3.71	0.767
	US	3.78	0.743

Table 3b

Independent Samples T-Test (Institutions)

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Language Productive Skills are Important for Future Employability	Equal variances assumed	0.017	0.895	0.354	352	0.724
	Equal variances not assumed			0.368	66.355	0.714
The university curriculum prepares students to attain sufficient language productive skills.	Equal variances assumed	0.276	0.600	0.163	352	0.871
	Equal variances not assumed			0.163	64.485	0.871
I am confident that my language productive skills are sufficient for the workplace.	Equal variances assumed	1.045	.307	-0.572	352	0.567
	Equal variances not assumed			-0.586	65.536	0.560

Based on the results from Table 3a, it appears that more UA students agree that LPS are important for future employability and that the university curriculum prepares students to attain sufficient LPS. On the other hand, more US students are confident that their LPS are sufficient for the workplace. However, based on the independent samples t-test in Table 3b, there was no significant difference between the variables ($p=0.05$).

Table 4a

Mean Comparisons between Working Experience

	Have you had any working experience	M	SD
Language Productive Skills are important for future employability	Yes	4.62	0.565
	No	4.66	0.546
The university curriculum prepares students to attain sufficient language productive skills.	Yes	4.00	0.825
	No	3.95	0.869
I am confident that my language productive skills are sufficient for the workplace.	Yes	3.83	0.722
	No	3.54	0.795

Table 4b

Independent Samples T-Test (Working Experience)

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Language Productive Skills are important for future employability	Equal variances assumed	0.650	0.421	-0.692	352	0.490
	Equal variances not assumed			-0.697	296.735	0.486
The university curriculum prepares students to attain sufficient language productive skills.	Equal variances assumed	2.011	0.157	0.506	352	0.613
	Equal variances not assumed			0.500	278.160	0.617

Table 4b *(Continued)*

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
I am confident that my language productive skills are sufficient for the workplace.	Equal variances assumed	9.299	0.002	3.529	352	0.000
	Equal variances not assumed			3.453	268.644	0.001

Concerning Table 4a, the students with no working experience agree that LPS are important for future employability. Furthermore, the students with working experience agree that the university curriculum prepares them to attain sufficient LPS and are more confident that their LPS are sufficient for the workplace. Table 4b shows the independent samples t-test between the variables; there is a significant difference between the mean scores of the students with working experience and without working experience for the item "I am confident that my language productive skills are sufficient for the workplace," (p=0.05).

The mean differences were also compared between the respondents' field of study (FoS), specifically, science and technology (ST), social sciences and humanities (SH), and business and administration (BA) and also based on their years of study (YoS). ANOVA was conducted, where the results are as follow:

Table 5a

Mean Comparisons between Fields of Study

		N	M	SD
Language productive skills are important for future employability	ST	90	4.51	0.604
	SH	148	4.73	0.489
	BA	116	4.62	0.585
The university curriculum prepares students to attain sufficient language productive skills.	ST	90	3.93	0.790
	SH	148	3.96	0.864
	BA	116	4.03	0.854
I am confident that my language productive skills are sufficient for the workplace.	ST	90	3.57	0.704
	SH	148	3.82	0.738
	BA	116	3.71	0.824

Table 5b

ANOVA (FoS)

		df	F	Sig.
Language productive skills are important for future employability	Between Groups	2	4.477	0.012
	Within Groups	351		
The university curriculum prepares students to attain sufficient language productive skills.	Between Groups	2	0.423	0.656
	Within Groups	351		
I am confident that my language productive skills are sufficient for the workplace.	Between Groups	2	3.075	0.047
	Within Groups	351		

Table 5c

Multiple Comparisons (FoS)

Dependent Variable	(I) FoS	(J) FoS	Sig.
Language productive skills are important for future employability	ST	SH	0.003
		BA	0.159
	SH	ST	0.003
		BA	0.112
	BA	ST	0.159
		SH	0.112
The university curriculum prepares students to attain sufficient language productive skills.	ST	SH	0.817
		BA	0.393
	SH	ST	0.817
		BA	0.473
	BA	ST	0.393
		SH	0.473
I am confident that my language productive skills are sufficient for the workplace.	ST	SH	0.014
		BA	0.189
	SH	ST	0.014
		BA	0.240
	BA	ST	0.189
		SH	0.240

In comparing the mean differences between the FoS as showcased in Table 5a, more SH students agree that LPS are important for future employability and are confident that their LPS are sufficient for the workplace. However, more BA students agree that the university curriculum prepares them to attain sufficient LPS. Albeit the overall ANOVA results in Table 5b which indicate no significant differences in all three items, based on Table 5c, there is a significant difference in the agreement that LPS are important for future employability between the ST and SH students (p=0.05).

Table 6a

Mean Comparisons between Years of Study

		M	SD
Language productive skills are important for future employability	Y1	4.58	0.591
	Y2	4.60	0.547
	Y3	4.76	0.471
	Y4	4.92	0.272
	Y5	4.50	0.905
The university curriculum prepares students to attain sufficient language productive skills.	Y1	4.06	0.720
	Y2	3.93	0.837
	Y3	4.00	0.991
	Y4	3.88	0.864
	Y5	3.83	1.193
I am confident that my language productive skills are sufficient for the workplace.	Y1	3.78	0.753
	Y2	3.66	0.745
	Y3	3.71	0.773
	Y4	3.81	0.749
	Y5	3.58	1.084

Table 6b

ANOVA (YoS)

		df	F	Sig.
Language Productive Skills are important for future employability	Between Groups	4	3.162	0.014
	Within Groups	349		
The university curriculum prepares students to attain sufficient language productive skills.	Between Groups	4	0.572	0.683
	Within Groups	349		
I am confident that my language productive skills are sufficient for the workplace.	Between Groups	4	0.549	0.700
	Within Groups	349		

Table 6c

Multiple Comparisons (YoS)

	(I) YoS	(J) YoS	Sig.	(I) YoS	(J) YoS	Sig.	(I) YoS	(J) YoS	Sig.	(I) YoS	(J) YoS	Sig.	(I) YoS	(J) YoS	Sig.
Language Productive Skills are Important for Future Employability	Y1	Y2	0.731	Y2	Y1	0.731	Y3	Y1	0.040	Y4	Y1	0.040	Y5	Y1	0.648
		Y3	0.040		Y3	0.066		Y2	0.066		Y2	0.066		Y2	0.547
		Y4	0.004		Y4	0.006		Y4	0.207		Y3	0.207		Y3	0.140
		Y5	0.648		Y5	0.547		Y5	0.140		Y5	0.028		Y4	0.028
The university curriculum prepares students to attain sufficient language productive skills.	Y1	Y2	0.216	Y2	Y1	0.216	Y3	Y1	0.661	Y4	Y1	0.661	Y5	Y1	0.377
		Y3	0.661		Y3	0.588		Y2	0.588		Y2	0.588		Y2	0.708
		Y4	0.340		Y4	0.807		Y4	0.562		Y3	0.562		Y3	0.534
		Y5	0.377		Y5	0.708		Y5	0.534		Y5	0.862		Y4	0.862
I am confident that my language productive skills are sufficient for the workplace.	Y1	Y2	0.229	Y2	Y1	0.229	Y3	Y1	0.554	Y4	Y1	0.554	Y5	Y1	0.398
		Y3	0.554		Y3	0.722		Y2	0.722		Y2	0.722		Y2	0.725
		Y4	0.866		Y4	0.381		Y4	0.577		Y3	0.577		Y3	0.611
		Y5	0.398		Y5	0.725		Y5	0.611		Y5	0.402		Y4	0.402

Based on the mean differences between the YoS as reported in Table 6a, the Y4 students are in the highest agreement that LPS are important for future employability and also the most confident that their LPS are sufficient for the workplace. Aside from that, the Y3 students are in the highest agreement that the university curriculum prepares them to attain sufficient LPS. Based on the ANOVA in Table 6b overall the mean differences are not significant but based on Table 6c, the mean scores between Y1 and Y4 in "language productive skills are important for future employability," are significant (r=0.05).

Research Question 2 – What are the students' perceptions of the most important language productive sub-skills needed at the workplace?

Regarding Table 7, the most important written communication sub-skill perceived by the students are writing reports (91.2%, N=323), followed by writing external emails (79.4%, N=281) and writing internal emails (75.7%, N=268). On the other hand, the least important sub-skill according to the students would be online chatting (42.4%, N=150), writing on company social media sites/websites (54.5%, N=193) and writing memos (59%, N=209). The other items recorded frequency statistics between 59% (N=209) to 69.2% (N=245).

According to the data in Table 8, the students perceived that presentations (89.8%, N=318), meetings (85.3%, N=302), and interviews (75.1%, N=266) to be so. As for the least important sub-skill, the students perceived that teleconferences (48%, N=170), dialogues (51.1%, N=181), and video conferencing (51.4%, N=182) fall under. As for the other items, the perceptions of importance were between 58.5% (N=207) and 74.9% (N=265).

Table 7

Students' perceptions on the most important written communication LPS at the workplace

Item	N	%
Writing reports	323	91.2
Writing internal emails	268	75.7
Writing external emails	281	79.4
Producing minutes of meeting	222	62.7
Preparing presentation slides	245	69.2
Writing memos	209	59
Writing business letters	216	61
Online chatting	150	42.4
Writing on company social media sites / websites	193	54.5
Writing proposals	240	67.8

Table 8

Students' perceptions on the most important spoken communication LPS at the workplace

Item	Frequency	%
Presentations	318	89.8
Idea pitching / product pitching	254	71.8
Teleconferences	170	48
Video conferencing	182	51.4
Internal phone calls	221	62.4
External phone calls	265	74.9
Meetings	302	85.3
Interviews	266	75.1
Dialogue	181	51.1
Round table discussion	249	70.3
Making appointments	207	58.5

Research Question 3 – What are the language productive skills needed by the students for the workplace?

Table 9 and Table 10 describe the findings on the students' needs on LPS regarding workplace communication. Based on Table 9, most of the students claimed that the written communication LPS needed is clear, concise, and complete writing (86.4%, N=306) followed by formatting documents (85.9%, N=304) and the usage of appropriate words/jargon (84.2%, N=298). Coherent writing is the least written communication LPS needed, with only 59.3% (N=210) claiming so. The other two items, sentence structure and grammar, recorded frequency statistics of 80.5% (N=285) and 76.6% (N=271), respectively.

Table 9

Written communication LPS needed by students

Item	N	%
Grammar	271	76.6
Sentence Structure	285	80.5
Usage of appropriate words / jargons	298	84.2
Format of document	304	85.9
Coherent writing	210	59.3
Clear, concise and complete writing	306	86.4

On the other hand, Table 10 shows the spoken communication LPS needed by the students with the highest skill needed is speaking confidently (92.1%, N=325), along with negotiation skills (76.8%, N=271) and speaking tone (74.2%, N=262). On the other hand, the least required skill needed by the students is articulation (41.4%, N=146), subsequently voice projection (58.6%, N=207) and pitch and volume (59.5%, N=210). The other skills, persuasion skills, voice clarity, and pronunciation, recorded demand of 65.7% (N=232), 67.4% (N=238) and 71.7% (N=253), respectively.

Table 10

Spoken Communication LPS Needed by Students

Item	N	%
Persuasion	232	65.7
Negotiation	271	76.8
Speaking confidently	325	92.1
Pronunciation	253	71.7
Articulation	146	41.4
Voice projection	207	58.6
Pitch and volume	210	59.5
Tone	262	74.2
Clarity	238	67.4

DISCUSSION

From the research conducted, we have identified that all the respondents generally agreed that the LPS is important in the workplace. There exists no difference between UA and US. Generally, students in the fourth year of their studies had a higher agreement that LPS is important for the workplace. An assumption could be made that because the students in their fourth year are closer to their industrial attachment and graduating, they have come to a higher realisation of the importance of LPS for workplace communication. Interestingly, social science students have the highest agreement that LPS is important for future employability. Also, students with working experience reported that they are more confident in their LPS as sufficient for the workplace.

In terms of the language forms and functions that were considered as important, the findings were divided into spoken and written communication. For spoken communication, the respondents believed that presentations, meetings and interviews were the most important spoken forms of workplace communication. They reported confidence, negotiation and intonation as the most important skills. In the CEFR

scale for formal discussion (meetings), students could achieve the C1 or B2 level if they can "keep up with the debate, even on abstract, complex unfamiliar topics", "keep up with an animated discussion", "argue a formal position convincingly, responding to questions and comments and answering complex lines of counter argument fluently, spontaneously and appropriately", and "express his/her ideas and opinions with precision, present and respond to complex lines of argument convincingly".

The data of this study are congruent to the literature of CEFR. Therefore, it assists a curriculum developer to design a syllabus, content and assessment for future language courses for spoken communication.

In contrast, the respondents believed that the most important forms are reports, external emails and internal emails for written communication. To do this, they believed that the most important skills are knowing the usage of appropriate words and jargon, formatting, and using clear and concise writing. When viewing the CEFR scales for writing, there did not seem to be a clear scale that could be used for written workplace communication. For the CEFR scale of overall written interaction, students could achieve the C1 or B2 level if they can "express him/herself with clarity and precision, relating to the addressee flexibly and effectively" and "express news and views effectively in writing, and relate to those of others".

The data in this study is not congruent to current CEFR literature because the current scale does not seem to comprehensively capture the necessarily written skillsets for workplace communication (Tables 8 and 9). Thus, curriculum developers must identify the CEFR scales to determine the LPS needed for written workplace communication.

CONCLUSIONS

Overall, this research was conducted as a needs analysis to identify students' perceptions of the use and importance of language productive skills (LPS) for workplace communication. The study was conducted as a preliminary study towards the development of a CEFR scale for workplace communication that can better reflect the needs of the industry to address Malaysian students' language proficiency and increase their employability. The findings from the study found that the perceptions for speaking skills generally match the scale available for CEFR's formal spoken communication. However, it did not match any available CEFR scale for written communication. This study has several implications. Firstly, for future curriculum development of language courses, this paper's findings help universities design relevant language proficiency/EOP courses. Secondly, it allows teaching practitioners to make informed decisions on the content of their language classes and courses. Thirdly, this study could form a framework for a CEFR-aligned scale for workplace communication in universities.

ACKNOWLEDGEMENT

REFERENCES

Agus, A., Awang, A. H., Yussof, I., & Makhbul, I. Y. M. (2011). The gap analysis of graduate employees' work skills in Malaysia. *Proceedings of Business and Information*, *8*, 1-15.

Akademi Pengajian Bahasa. (2016). *Course Information for ELC650 (English for Professional Interaction)*. Universiti Teknologi MARA.

Albaum, G. (1997). The likert scale revisited. *International Journal of Market Research, 39*(2), 1-21. https://doi.org/10.1177/147078539703900202

Barni, M. (2015). In the name of the CEFR: Individuals and standards. In B. Spolsky, O. Inbar-Lourie & M. Tannenbaum (Eds), *Challenges of language education and policy: Making space for people* (pp. 40-52). Routledge.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Council of Europe.

Daniel, E. (2016). The usefulness of qualitative and quantitative approaches and methods in researching problem-solving ability in Science education curriculum. *Journal of Education and Practice, 7*(15), 91-100.

Dzulkifly, D. (2018, November 8). Employers blame poor communication, language skills for unemployed Bumi grads. *Malay Mail*. https://www.malaymail.com/news/malaysia/2018/11/08/employers-blame-poor-communication-language-skills-for-unemployed-bumi-grad/1691146

Foley, J. A. (2019). Issues on assessment using CEFR in the region. *Language Education and Acquisition Research Network Journal*, *12*(2), 28-48.

Harsch, C., & Seyferth, S. (2020). Marrying achievement with proficiency - Developing and validating a local CEFR-based writing checklist. *Assessing Writing, 43*, Article 100433. https://doi.org/10.1016/j.asw.2019.100433

Hee, S. C., & Zainal, A. Z. (2018). Analysing the English communication needs of service technicians in the pest control industry. *Pertanika Journal of Social Sciences & Humanities, 26*(2), 1079-1095.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centred approach.* Cambridge University Press. https://doi.org/10.1017/CBO9780511733031

Isnin, S. F., Mustapha, R., & Othman, W. M. (2018). engineering students' perspectives on the need of a new module in technical report writing at Polytechnic in Malaysia. *Journal of Engineering Science and Technology, February* (2018), 31-38.

Labaree, R. V. (2009). *Research guides - The methodology*. USC Libraries. Retrieved February 3, 2020, from https://libguides.usc.edu/writingguide/quantitative

Malaysian Ministry of Education. (2013). *Malaysia Education Blueprint 2013-2025*. Kementerian Pendidikan Malaysia.

Malaysian Qualifications Agency. (2018). *Programme standards: Language*. MQA.

Mustafa, Z. (2019, January 23). Importance of academia-industry linkages. *New Straits Times*. https://www.nst.com.my/education/2019/01/453582/importance-academia-industry-linkages

Perinpasingam, P. T. S., Arumugam, N., Thayalan, X., & Maniam, M. (2015). Needs analysis on

the importance of English language skills for workplace: Trainee architects. *International Review of Basic and Applied Sciences, 3*(10), 129-137.

Sarudin, I., Noor, Z. M., Zubairi, A. M., Ahmad, T. B. T., & Nordin, M. S. (2013). Needs assessment of workplace English and Malaysian graduates' English language competency. *World Applied Sciences Journal, 21*(Special Issue of Studies in Language Teaching and Learning), 88-94. https://doi.org/10.5829/idosi.wasj.2013.21.sltl.2141

Singh, P, Fook, C. Y., & Sidhu, G. K. (2015). *A comprehensive guide to writing a research proposal*. Venton Publishing (M) Sdn. Bhd.

Singh, P., Ghani, P. A., & Hoon, T. S. (2009). *Quantitative data analysis for novice researchers*. Primera Publishing.

Tajuddin, A. J. A. (2015). *A Malaysian professional communication skills in English Framework for English for Occupational Purposes Courses* [Doctoral dissertation, University of Nottingham]. Nottingham eTheses. http://eprints.nottingham.ac.uk/id/eprint/28747

Uri, N. F. M., & Aziz, M. S. A. (2018). Implementation of CEFR in Malaysia: Teachers' awareness and the Challenges. *3L The Southeast Asian Journal of English Language Studies, 24*(3), 168-183. https://doi.org/10.17576/3l-2018-2403-13

Zainuddin, S. Z. B., Pillai, S., Dumanig, F. P., & Phillip, A. (2019). English language and graduate employability. *Education + Training, 61*(1), 79-93. https://doi.org/10.1108/et-06-2017-0089

Malaysian ESL Teachers' Practice of Written Feedback on Students' Writing

Khairil Azwar Razali¹*, Zainurin Abdul Rahman², Ismail Sheikh Ahmad² and Joharry Othman²

¹*Kulliyyah of Languages & Management, International Islamic University Malaysia, Pagoh Edu Hub, 84600 Johor, Malaysia*
²*Kulliyyah of Education, International Islamic University Malaysia, P.O.Box 10, 50728 Kuala Lumpur, Malaysia*

ABSTRACT

Feedback to students' writing plays an important role as a scaffolding technique to help the students to improve their writing skills. With the introduction of school-based assessment and the Common European Framework of Reference (CEFR) into the new Standards-based English Language Curriculum (SBELC), teachers are expected to adapt the process writing approach in their classroom, where feedback is at the core of the process writing approach. This present study aims to explore Malaysian ESL teachers' practice of written feedback in their writing classrooms. Two sample essays were used in this study. The sample essays were written by a Form Three student of a secondary school in Kuantan, Pahang, and a Form Five student from a secondary school in Manjung, Perak. The sample essays were sent to all secondary schools in Pahang, and teachers who teach the English Language at the schools were asked to mark the essay as how they would normally mark their students' essays. The participants of this study were selected using purposive sampling. A total of 89 student sample essays with the teachers' marking were returned, and the teachers' feedback were analysed. This study found that most of the participants mark their students' essays comprehensively and implicitly. However, some of the respondents did not give any feedback at all, and even if they did, the feedback would be retracted from the marking rubric. It has also been found that the respondents of this present study did not utilise comments on goals to work towards or specific activities for improvement. This

paper further discusses the findings in view of the assessment of learning (AfL) and gives recommendations for future practice.

Keywords: ESL writing, teaching writing, writing assessment, written corrective feedback

INTRODUCTION

Within ESL classrooms, teachers' written corrective feedback has always been under scrutiny by academics, as an inconclusive debate is still going on since the publication of Truscott (1996) that sparked the debate. Teachers and researchers have been studying all aspects of teachers' feedback to students' writings since then. However, the results are still inadequate as to whether such practice could help students develop their writing. Realising this, future research on teachers' feedback needs to move from whether it is effective to focus on what type of feedback is effective (Shelly, 2014). Moreover, teachers need to be innovative in providing feedback to students' writings (Lee, 2014).

The teaching of writing within Malaysian ESL classrooms is governed by the curriculum specifications and syllabuses set by the Malaysian Ministry of Education. The Education Ministry advocates the process writing approach; thus, as stipulated in the English Language curriculum, the learning outcomes match the process writing approach (Abdullah & Sidek, 2012).

With the introduction of school-based assessment (SBA) and the Common European Framework of Reference (CEFR), the new Standards-based English Language Curriculum (SBELC) has been developed to align the pedagogies in Malaysian schools to that of CEFR (Kementerian Pendidikan Malaysia, 2017a). For example, under the new SBELC, students are expected to "produce a plan or a draft of two paragraphs or more and modify this appropriately either in response to feedback or independently" (Kementerian Pendidikan Malaysia, 2017b, p. 38). That is in line with the process writing approach, which is embedded within the formative assessment.

The same element, process writing, has also been highlighted in the Curriculum Specifications for English Language Form 4, where teachers need to apply process writing skills, which include "making an outline, … writing out 1st draft, revising and editing the draft…, rewriting 2nd draft, proof-reading draft, … and writing out the final draft" (Malaysian Ministry of Education, 2003, p. 18).

However, a study done by Maarof et al. (2011) has found that students are not allowed to revise their essays, as the teachers have not utilised the process writing approach. The study was done in five secondary schools in southern Malaysia, where 150 Form Five students answered a survey on students' perceptions of teacher and peer feedback in enhancing ESL students' writing. Maarof et al. (2011) mentioned that students do not produce multiple drafts of their essays "because of time constraints, the large number of students in a classroom, absence of the practice of process writing and students' lack of motivation." (p. 29). Further to this,

Nesamalar et al. (2001) claim that Malaysian students have writing skills deficiencies.

In a study done by Gurnam et al. (2011), it has been found that only 68% of the students received feedback immediately after each assessment. This finding indicates that the conception of formative assessment that the Ministry of Education champions is not being practised in schools.

There has been no study done on teacher practice of written feedback within Malaysian ESL classrooms. Previous studies have not looked into teachers' practices in providing feedback to students' writings. It is imperative to align teachers' practices to the formative assessment framework that is part of the school-based assessment that has been introduced in the national curriculum. The objective of this study is to find out the practices of written feedback of English Language teachers in Pahang in their writing classroom, and more specifically, this study tries to answer the research question "What is ESL teachers' current practice of written feedback in the writing classroom?"

Feedback in ESL Writing Classroom

According to Ramaprasad (1983, p. 4), feedback is "the information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way." Given the definition, feedback could come in two forms: corrective feedback and general comments about the work.

Hyland and Hyland (2006) claimed that feedback in an ESL writing classroom functions in two ways, firstly, as a key element of the students' growing control over writing skills, and secondly, as teachers' scaffolding technique. Summative feedback, designed to evaluate writing as a product, is generally replaced with formative feedback, which helps students develop their writing skills. The process approach in providing feedback to students' writing, that is formative feedback, encourages teachers to support students' development in writing through multiple drafts by providing feedback during the writing process itself, rather than at the end of the writing process.

According to Hyland and Hyland (2006), feedback in ESL writing classroom could be divided into:

1. Written feedback
2. Teacher-student conference
3. Peer feedback and
4. Computer-mediated feedback

Written Feedback in the ESL Classroom

Feedback on students' writing is a critical, non-negotiable aspect of writing instruction, in which teachers help students shape their composition and writing skills (Ferris & Hedgcock, 2005). Research on teacher feedback has been done extensively; nonetheless, the result is somewhat inconclusive (Ferris, 2012). The argument on the effectiveness of feedback to students' writing started with a paper by Truscott (1996), where it was argued that previous research failed to show positive results of teachers' written feedback to students' writing development. Truscott further argued that such practice is harmful because

it takes energy and attention away from more pressing issues, such as developing students' ideas in writing courses. Finally, by supporting Krashen's Monitor Hypotheses, Truscott claimed that comprehensible input is sufficient for L2 acquisition. Students should be exposed to extensive experience with the target language through various reading and writing exercises.

The first response to Truscott (1996) was written by Ferris (1999) where she claimed that corrective feedback does help in language learning. Ferris' challenge led to more research done in the area up until today. Chandler (2003), one of the important studies, found that the grammar accuracy of students who received corrective feedback improved in L2 over time compared to the control group who did not receive any corrective feedback.

In studying the effectiveness of corrective feedback of different types, Bitchener (2008), in a two-month study, found that students who received corrective feedback of any type performed better compared to those who did not receive any corrective feedback. Furthermore, a further study on the same participants showed that the treatment group who received corrective feedback improved their writing accuracy. Thus, it clearly shows the positive effects of corrective feedback.

One of the most common types of written feedback is corrective feedback. Lee (2005) explains four written corrective feedback methods, divided into two categories: Comprehensive vs Selective and Explicit vs Implicit. Although providing correct grammatical errors is one of the most popular techniques among many language teachers, various types of corrective feedback have been recommended as it is considered more effective and successful than simply relying on a single method (Corpuz, 2011).

The comprehensive written corrective feedback approach is made when the teacher corrects all students' writing errors, irrespective of their error category. Comprehensive written corrective feedback could help students notice errors made and new features of the target language as postulated in Krashen (1992) Noticing Hypothesis. By noticing, effective language learning could be promoted. Nevertheless, Ellis et al. (2006) claimed that given the limited capacity of students processing ability, students might be overwhelmed; thus, comprehensive written corrective feedback may not be as effective as it should be.

On the other hand, the selective written corrective feedback approach targets specific grammatical errors only, leaving all other errors uncorrected. Ellis (2009) claimed that selective written corrective feedback might be more effective than comprehensive written corrective feedback as students can examine multiple corrections of a single error. Thus, students obtain a richer understanding of what is wrong in their writing and opportunities to acquire the correct form.

Explicit written corrective feedback is the type of feedback where the L2 teacher

directly provides the correct forms or structures to show explicitly the error in the students' writing. In the research done by Ellis et al. (2006), it is found that explicit written corrective feedback is more effective for treating errors in verb tenses.

Implicit written corrective feedback is where the teacher shows that an error is made by underlining, marginal description, circling or correction codes. Correction codes implicitly provide corrections using symbols and abbreviations to inform students of an error and the kind of error made. Lee (1997) found that students favour implicit written corrective feedback compared to explicit written corrective feedback. In earlier research by Lalande (1982), participants showed a reduction of errors in writing when implicit written corrective feedback is used.

Over the years, improvement-oriented feedback has emerged and is said to be more favourable than the achievement-oriented feedback. According to Dinnen and Collopy (2009), achievement-oriented feedback would give suggestions on improving the students' work, as compared to achievement-oriented feedback, where the emphasis is given on whether the work has achieved the desired standards. Cho et al. (2006), in a research done on perceived usefulness of comments, found that improvement-oriented feedback to be more effective. In a more recent study, Wu and Schunn (2020) found that students would respond to feedback that offered specific revisions recommendations and often better understood the problem that occurred in their work.

METHODOLOGY

For this study, the case study approach was used. Case study offers insight into regularities or recognisable patterns of the unique individual, or group of people, that could be used in understanding the phenomenon more accurately (Fraenkel & Wallen, 2010). The objective of this study is to find out the practices of written feedback of English Language teachers in Pahang in their writing classroom, and more specifically, this study tries to answer the research question "What is ESL teachers' current practice of written feedback in writing classroom?"

Two sample essays were used for data collection. A Form 3 student from a secondary school in Kuantan, Pahang, wrote the first sample essay. The student was said to be an average student, where he would normally score a B or C in his English Language tests and exams. However, later in the same year, this student sat for PT3 and scored a B when this study took place. The writing task was taken from a module for PT3, which was developed by Hamidi (2015). The writing task is on recount where it follows the format as stipulated in PT3. In the task, students were asked to write a letter to a friend about the incident that happened during his/her birthday party. In the task, salutation, the first paragraph and the last paragraph are given. The sample consists of 143 words, written in two paragraphs.

The second student sample essay was written by a Form 5 student from a secondary school in Manjung, Perak. The student attended a tuition class held by the

author of the module (Kamaruddin, 2016). The student was said to be an average student where she would normally score B in her English Language tests and exams. However, in the same year, this study was held, the student sat for her SPM and scored B+ for English Language. The task of the essay was postulated by Kamaruddin (2016). The question follows the format of Section B, Paper 1 SPM, asking students to respond to several options in continuous writing. For this study, the student wrote an essay entitled "The Most Embarrassing Moment of My Life." The essay consists of 396 words, written in seven paragraphs. In order to retain the authenticity of both samples, photocopied copies of the students' handwritten essays were used. Respondents of this study were expected to give feedback to the essay in written form.

The respondents of this study consist of English Language teachers who teach in Pahang. They were selected using a purposive sampling method. Teachers who teach English Language in secondary schools in the state of Pahang were approached and asked to participate voluntarily in this study. A cover letter explaining the study's objectives was sent together with the sample essays and the consent form to be signed by the participants should they agree to participate in this study.

A total of 89 sample essays were returned to the researcher. Out of which, 42 essays were PT3, and 47 essays were SPM essays. From Table 1 and Table 2 below, the majority of the teachers who participated in this study were Language Teachers (62 teachers), and this was followed by Heads of Panel (14 teachers). Interestingly, there were six non-optionist teachers participated in this study. Generally, non-optionist teachers are not trained to be English Language teachers, but they were trained to teach other subjects. Schools with an insufficient number of English Language teachers often assign teachers of other subjects to teach English. It is also common for other subjects. Out of the 89 participants, 43 teachers teach at rural area schools, while there were 46 participants from urban schools. From the demographic data collected, about 72% (n = 64) of the participants teach at *Sekolah Menengah Kebangsaan* (SMK–National Secondary School), 14 teachers were from *Sekolah Berasrama Penuh* (SBP–Boarding Schools), followed by eight teachers from *Sekolah Menengah Kebangsaan Agama* (SMKA–Islamic National Secondary School), and three teachers from *Sekolah Menengah Jenis Kebangsaan* (SMJK–National Type Secondary School). Most of the participants had four to five years of pre-service training, with about 43% (n = 38), while the majority had been teaching between five to nine years (24.7%, n= 22). Out of the 89 participants who participated in this study, only 11 teachers have master's degrees. Most of the participants are females (n = 75), and only 14 teachers are males.

Although the respondents were not marking their own students' essays, they have been reminded to mark the sample essays like they would normally do in their

Table 1

Profile of the Participants (PT3)

Role	School location	School category	Pre-service training (year)	Teaching experience (year)	Education level	Sex
Head of Department: 1 Head of Panel: 4 Language Teacher: 31 Media Teacher: 1 Non-optionist: 3 Senior Assistant: 2	Rural = 22 Urban = 20	SBP: 7 SMJK : 2 SMK: 29 SMKA: 4	0 – 1: 12 2 – 3: 8 4 - 5: 17 ≥ 6: 5	0 – 4: 9 5 – 9: 7 10 – 14: 5 15 – 19: 8 20 – 24: 8 25 – 29: 1 ≥ 30: 4	Bachelor's degree: 38 Master's degree: 4	Female: 32 Male: 10

*SBP = *Sekolah Berasrama Penuh* (Boarding School), SMJK = *Sekolah Menengah Jenis Kebangsaan* (National Type Secondary School), SMK = *Sekolah Menengah Kebangsaan* (National Secondary School), SMKA = *Sekolah Menengah Kebangsaan Agama* (Islamic National Secondary School)

Table 2

Profile of the Participants (SPM)

Role	School location	School category	Pre-service training (year)	Teaching experience (year)	Education level	Sex
Head of Department: 3 Head of Panel: 10 Language Teacher: 31 Non-optionist: 3	Rural = 21 Urban = 26	SBP: 7 SMJK : 1 SMK: 35 SMKA: 4	0 – 1: 5 2 – 3: 9 4 – 5: 21 ≥ 6: 12	0 – 4: 7 5 – 9: 15 10 – 14: 7 15 – 19: 10 20 – 24: 3 25 – 29: 4 ≥ 30: 1	Bachelor's degree: 40 Master's degree: 7	Female: 43 Male: 4

*SBP = *Sekolah Berasrama Penuh* (Boarding School), SMJK = *Sekolah Menengah Jenis Kebangsaan* (National Type Secondary School), SMK = *Sekolah Menengah Kebangsaan* (National Secondary School), SMKA = *Sekolah Menengah Kebangsaan Agama* (Islamic National Secondary School)

classrooms as if they are marking their students' essays. Moreover, the same essays were used in this study so that the feedback given by the teachers are comparable, as opposed to if they were to mark different essays. Finally, the original handwriting of the students was also retained to ensure the authenticity of the sample essays. Thus, it is in line with the design of a case study.

Data Analysis

For data analysis of the students' sample essay, teachers' responses to the sample were analysed in two stages: written corrective feedback and written feedback or comments.

In analysing the teachers' written corrective feedback on the sample essay, their written corrective feedback was first categorised. According to Lee (1997), teachers' written corrective feedback could be categorised into four main groups, which are 1) Selective, 2) Comprehensive, 3) Direct, and 4) Indirect. Furthermore, all these four groups could be overlapping, where a teacher's marking could be selective and indirect when the teacher chooses certain features of language that he/she wants to mark. For example, the teacher can put a symbol on the error or at the right margin of the paper without giving any correct answer.

In the second stage, teachers' comments and remarks were analysed using a checklist that was developed for this study. The list was adapted from several earlier works by Wiliam by Nyquist (2003), Nicol and

Macfarlene-Dick (2006), and Juwah et al. (2004). The checklist is as follows:

1. score/grade
2. stating students' current learning state
3. goals to work towards
4. correct answers
5. explanation of the correct answers
6. suggestions for improvement
7. specific activities for improvement
8. facilitates self-reflection
9. encourages positive motivation and self-esteem, and
10. encourages teacher and peer dialogue.

FINDINGS

The findings of this study are divided into two parts: the PT3 sample essay and the SPM sample essay.

Findings of PT3 Sample Essay Analysis

A total of 42 PT3 sample essays were returned to the researcher. All were marked using a comprehensive marking style, with no specific errors marked and/or corrected. From this number, 32 teachers marked Implicitly, which is not correcting the errors, while nine teachers marked Explicitly, which is correcting the errors committed by the student. In addition, there is one sample essay marked using Impression marking style, which gives marks without making any mark on the sample essay. Table 3 below summarises the analysis of the written corrective feedback (WCF).

Table 3

Analysis of WCF (PT3)

Type of WCF	n	Percentage, %	N
Explicit	9	22%	41
Implicit	32	78%	
Selective	0	0%	41
Comprehensive	41	100%	
*1 with impression marking			

From the sample essays, written corrective feedback ranges from zero to 25 on the sample essays. Most of the teachers (*n* = 18) corrected between six to ten errors on the sample essays. It is followed by 11–15 corrective feedback (*n* = 9), followed by zero to five (*n* = 8), five teachers gave 16–20 corrective feedback, and two teachers gave 21–25 corrective feedback. Table 4 depicts the corrective feedback count for PT3.

Table 4

Corrective feedback count (PT3)

CF Count	Frequency	Percentage, %
0–5	8	19
6–10	18	42.9
11–15	9	21.4
16–20	5	11.9
21–25	2	4.8
TOTAL	42	100

In terms of comments, 17 teachers did not comment on the student's sample essay, twelve others wrote only one comment, seven teachers wrote two comments, three teachers wrote three comments, and two teachers wrote four comments. Table 5 illustrates the number of comments per script for PT3.

Table 5

Number of comments per script (PT3)

No. of Comment	Frequency	Percentage, %
0	17	41.4
1	12	29.3
2	7	17.1
3	3	7.3
4	2	4.9
5	-	0
TOTAL	41	100

All the feedback could be distributed into eleven types, where the highest number of teachers (n = 10) gave the correct answers. Nine teachers gave scores and/or grades, and the same number of teachers identified the students' current learning state. Often enough, this is taken from the marking rubric prepared by the Examination Board. For example, teacher #77 wrote '*task fulfilled*' and '*some mistakes in grammar and spelling*'. Teacher #88 also made

remarks about the student's current learning by listing four comments, which were 1) *Task is fulfilled*; 2) *Ideas are sufficiently developed*; 3) *Vocabulary is sufficient but lacks precision*; and 4) *Interest is sufficiently aroused*. Again, these kinds of remarks could be found in the marking rubric.

Six teachers encouraged positive motivation and self-esteem. For example, teacher #83 wrote '*very good writing,*' and teacher #192 wrote '*good try!*' Four teachers wrote suggestions for improvement ('*use sentence connectors*' and '*some of the sentences could be merged, so that it'll*

be longer + complete with some details.'), four teachers explained the correct answers, and two teachers encouraged teacher dialogue ('*come and see me*'). Two teachers facilitate self-reflection ('*why did you serve the cake when you realised the cake was salty beforehand?*'). Only one teacher commented on goals to work towards, and another teacher's comment fell under 'Other'. None of the teachers suggested specific activities for improvement. Table 6 below depicts the distribution of comments according to feedback type for PT3.

Table 6

Distribution of comments according to feedback type (PT3)

No	Type of Feedback	Examples of Feedback	Frequency
1.	Score/grade	Mark range: (full mark is 15)	9
		Mark n	
		7 1	
		8 5	
		10 2	
		12 1	
2.	Correct answer	*To went* been corrected *to go*:	10
3.	Explanation of the correct answer	"*to + <u>base word</u>, e.g., to go*"	4
4.	State students' current learning state	*Task is fulfilled.* *Ideas are sufficiently developed.* *Vocabulary is sufficient but lacks precision.* *Interest is sufficiently aroused.*	9
5.	Goals to work towards	nil	1
6.	Suggestions for improvement	'*use sentence connectors*' '*some of the sentences could be merged, so that it'll be longer + complete with some details.*'	4
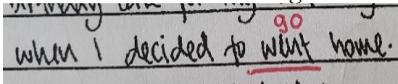
Table 6 *(Continued)*

No	Type of Feedback	Type of Feedback	Frequency
7.	Specific activities for improvement	nil	0
8.	Facilitates self-reflection	'*why did you serve the cake when you realised the cake was salty beforehand?*'	2
9.	Encourage positive motivation & self-esteem	'very good writing' 'good try!'	6
10.	Encourage teacher & peer dialogue	'*come and see me*'	2
11.	Other comment	'do your correction'	1

Findings of SPM Sample Essays Analysis

Table 7

Analysis of WCF (SPM)

Type of WCF	*n*	Percentage, %	N
Explicit	3	6.4	
Implicit	36	76.6	47
Explicit + Implicit	8	17.0	
Selective	0	0	47
Comprehensive	47	100	

Out of the 47 sample essays received, all of them were marked using the Comprehensive style. From this, 36 were marked Implicitly, while three were marked Explicitly. However, eight teachers marked the sample essays with a combination of Explicit and Implicit styles. Table 7 above illustrates the analysis of WCF for SPM samples.

In terms of corrective feedback, the lowest count was 20, while the highest was 77. For example, one teacher gave 20 corrective feedback on the sample essay, and only one teacher gave 77. A total of nine teachers gave corrective feedback within the 46–50 range, followed by eight teachers who gave 36–40 corrective feedback, and seven teachers who gave 41–45 corrective feedback. Finally, four teachers gave 56–60 corrective feedback, while one teacher gave 61–65, 66–70, and 71–75 corrective feedback. Table 8 illustrates the corrective feedback count for SPM.

Table 8

Corrective feedback count (SPM)

CF Count	Frequency	Percentage, %
16–20	1	2.1
21–25	4	8.5
26–30	1	2.1
31–35	3	6.5
36–40	8	17.1
41–45	7	14.9
46–50	9	19.1
51–55	6	12.8
56–60	4	8.5
61 – 65	1	2.1
66–70	1	2.1
71–75	1	2.1
76–80	1	2.1
TOTAL	47	100

Table 9

Number of comments per script (SPM)

No. of Feedback	Frequency	Percentage, %
0	18	38.3
1	9	19.1
2	10	21.3
3	7	14.9
4	1	2.1
5	2	4.3
TOTAL	47	100

In terms of the comment, 18 teachers did not give any comment at all, while nine only gave one comment, ten teachers wrote two comments, seven teachers wrote three comments, one teacher wrote four comments, and two teachers wrote five comments on the student sample essays. Table 9 depicts this information.

The feedback could be distributed to eleven types, where 15 of the teachers wrote scores/grades on the essay, while 13 others encouraged positive motivation and self-esteem. Examples of positive motivation and self-esteem are "*very interesting!*" and "*good try,*" written by Teacher #16. In contrast, Teacher #352 wrote, "*Don't stop writing. I can see your potential—just need to polish it,*" and a smiley accompanied this remark at the end.

Eleven teachers gave the correct answers and suggestions for improvement. One of the respondents, for example, listed four suggestions for improvement, namely 1) *Please be careful with the tense you use*; 2) *Just stick to simple past tense that will minimise your errors*; 3) *Try to use sophisticated words/phrases to enhance the accuracy of your sentences*, and 4) *Please read your essay before submitting it as it helps you a lot in detecting errors/missing words*.

Two teachers explained the correct answers, and two others commented 'Other.' However, none of the teachers commented on goals to work towards, gave specific activities for improvement, or encouraged teacher and peer dialogue. Table 10 below illustrates the distribution of comments according to feedback type for SPM.

DISCUSSION

The findings from the student sample essays indicate that most teachers in this present study mark their students' essays

Table 10

Distribution of comments according to feedback type (SPM)

Type of Feedback	Examples of Feedback			Frequency
Score/grade	Grade	Score	n	15
	Satisfactory (C5 – C6)	26 – 30	6	
	Passable (D7)	21 – 25	4	
	Unsatisfactory (E8)	16 – 20	3	
	Poor (F9)	15 – 0	2	
Correct answer	The word *took* has been corrected as *to take*.			11
Explanation of the correct answer	Past tense			2
	was looked	X telea dilihatkan??		
	looked	√ telah melihat		
	was looking	√ sedang melihat (lepas)		
State students' current learning state	"*Errors in wrong usage of prepositions, articles and determiners somehow hinder the reading*"			10
Goals to work towards	nil			-
Suggestions for improvement	*Just stick to simple past tense that will minimise your errors.* *Try to use sophisticated words/phrases to enhance the accuracy of your sentences.* *Please read your essay before submitting it as it helps you a lot in detecting errors/missing words.*			11
Specific activities for improvement	nil			-
Facilitates self-reflection	nil			-
Encourage positive motivation & self-esteem	"Very interesting!" "Good try" "Don't stop writing. I can see your potential – just need to polish it"			13
Encourage teacher & peer dialogue	nil			-
Other comments	"Please do the correction!" "short [sic] than required number – write longer please!"			2
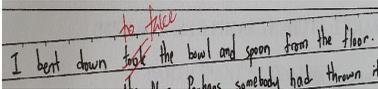
comprehensively and implicitly. That means teachers would mark almost all errors they could locate on the essays, but the corrected forms are not provided. This result is concurrent with the findings from Lee (2008) where she found that most the teachers' feedback on students' writings was focusing on the students' errors. It is in line with a long-held belief as mentioned by Lalande (1982), specifically on comprehensive error correction, where he wrote, "unless all errors are identified, the faulty linguistic structures, rather than the correct ones, may become ingrained in the students' interlanguage system" (p. 140). However, more recent literature suggests that comprehensive error correction may overwhelm the student, as their limited processing ability may not digest the amount of WCF provided by their teachers (Ellis et al., 2006), so comprehensive marking may not be as effective as teachers hope. Moreover, such practice is unclear, inconsistent, and overemphasised the negative (Fregeau, 1999; Cohen & Cavalcanti, 1990). Williams (2003) argues that correction of errors allows passive action among the students who would rewrite the corrected form without knowing the nature of their errors. Therefore, this practice is ineffective in promoting learning among the students.

The findings of the teachers' marking on the student sample essays came back with some peculiarities. First, there was one teacher who used impression marking on the PT3 sample essay. Impression marking, as mentioned by Baird et al. (2004), is based upon a general impression of the essay by the examiners. Impression marking is not designed to correct or edit a piece of writing or even to diagnose its weakness, but rather is a set of procedures for assigning a value to the writing according to a list of previously established criteria (Charney, 1984). Baird et al. (2004) claimed that there are problems related to impression marking: the reliability and validity of the marks awarded through this procedure. In their attempt to rectify this issue, they have carried out an experiment using the theory of community of practice and found that neither use of exemplar essays nor discussion between examiners demonstrated an improvement in marking reliability. Because of this, the effectiveness of general impression marking has been questioned. In the context of PT3 and SPM, the general impression is used.

However, it has not been a practice among teachers because they are still required to check for errors to justify their marks for the essay. It is particularly true for SPM level essays. For example, in order to award band D7 (21–25 marks) for SPM Paper 1 (Continuous Writing), teachers need to identify "many mistakes in grammar but the meaning is still clear—patches of accurate language use occur." Without marking the student's essay, it is arguable how teachers can justify the marks they give. Moreover, the teachers in this present study were asked to mark the essay "as they would normally do in the classroom;" hence if the teacher uses impression marking without making any marking on the essay, the students may not be able to know what is wrong with the essays that they

deserve such mark. Nyquist (2003) labelled feedback that gives only the score or grade as 'weaker feedback only.' This kind of feedback is not in line with the suggestions in implementing AfL as suggested by Black et al. (2003), where score or grade only may not enable students to improve the essay nor the following essay. Moreover, Black and Wiliam (1998) noted that such marking and grading practices emphasise competition, not the student's improvement.

It has also been noted that teachers in this present study employed a mix of explicit and implicit marking ($n = 8$). These teachers would mark some errors and give the correct forms while leaving some errors marked but not corrected. It means that teachers still employ comprehensive error correction marking. Teachers believe that language accuracy is an important focus in their feedback, and this echoes the results of previous studies (Amrhein & Nassaji, 2010; Hyland, 2003; Lee, 2008). As mentioned above, comprehensive error correction marking may not necessarily promote students' learning, even though many teachers practice it (Lee, 2009) and students prefer it (Salteh & Sadeghi, 2015). In the study done by Lee (2009), it was found that 94.1% of the teachers in that study focused on correcting error forms, while they believe that there are more to writing besides grammar accuracy, such as delivering good ideas. The study done by Salteh and Sadeghi (2015) reveals that 77% of the students in that study prefer indiscriminate correction of all errors in their essays. The present study may not be able to reveal students' preference for written corrective feedback,

but the findings echo the previous studies on the same issue.

Even though many students prefer correcting all errors, as mentioned above, Salteh and Sadeghi (2015) also noted some issues related to comprehensive error marking. In their study, Salteh and Sadeghi noted that 23% of the students felt frustrated when receiving their essays filled with red marks. The same frustration by students was highlighted more than twenty years ago by Reid (1998). Moreover, Lee (2004) caution that marking all errors in the students' essays could enslave the teachers, as mentioned by many earlier studies (Enginarlar, 1993; Ferris, 2002; Mantello, 1997). Hence, it can be said that teachers in the present study still practice what has been mentioned as not thoroughly effective in the earlier studies.

In terms of teachers' comments, Lunsford (1997) stated that three well-thought-out comments per essay is optimum, given that students would act on those comments. Ferris (2006) in a study found that students utilized the teachers' feedback in their revision, and this refuted earlier studies done by Cohen and Robbins (1976), Truscott (1996), and Zamel (1985). However, it is a concern for those teachers who did not comment on the sample essays, besides marking the errors committed on the essays. If this is their common practice in the classroom, students may not get much help from these teachers. Feedback, at its basis, should tell the students their current state of learning, the goal they need to achieve, and how to achieve the goal (Black, 1999). Comprehensive error correction, without any other feedback, cannot even be

categorised into any typology of feedback, either by Nyquist (2003) or Black (1999), let alone those criteria of good feedback (Juwah et al., 2004; Nicol & Macfarlene-Dick, 2006). Irons (2008) listed no feedback as lousy feedback; therefore, it should be avoided.

As Lee (2009) mentioned, teachers acknowledged that students would ignore their other feedback if they wrote scores or grades on their essays. However, it is arguable that such grade or score is necessary to state the students' current learning state, as defined by Black (1999). By knowing their grade or score for a particular essay, the students will know their level of attainment, allowing them to work towards the goal of obtaining a grade of A.

CONCLUSION AND RECOMMENDATIONS FOR FUTURE PRACTICE

The findings of this present study prove that there is a need to include written feedback into teacher training courses, especially for English Language teachers. There is little emphasis on written feedback on students' writings within teacher training courses, especially in Malaysia. Besides the courses on Theories of Assessment, teacher trainees should also be taught on how to give feedback to their students' work. An emphasis on process writing should also be included. It is to match with the current school-based assessment system that takes place in Malaysian schools now.

Another suggestion that could be made is to put more emphasis on giving feedback

to students' work. There is no mention of how teachers should mark and give feedback to the students' writings in the English Language syllabus for both primary and secondary schools. According to the English Language Curriculum Specification for Form 1 (Ministry of Education, 2003), under the subtopic Evaluation, "After every lesson, teachers are encouraged to assess their set of learners through simple questioning techniques or some other exercise so that they can pace their lessons in accordance with learners' progress" (p. 5). However, the simplistic instruction on evaluation is not enough to give the right ideas to the teachers on how to give feedback to their students' work, let alone on giving feedback on the writings.

Perhaps, we should learn from our neighbouring countries, which elaborate further on assessing the students' work. Take Singapore for example, in their English Language Syllabus Primary and Secondary (Ministry of Education Singapore, 2010), teachers are guided in planning assessment through a framework in the syllabus (see page 123, Singapore English Language Syllabus, 2010). On the other hand, Hong Kong's English Language Syllabus comes with Curriculum and Assessment Guide (Hong Kong Education Bureau, 2007). In the guide, thorough explanation is given not only on the curriculum, but also on teaching and learning process, as well as on carrying out assessments in schools. In terms of writing, teachers are guided on how to carry out process writing in the classrooms (see p. 116, Hong Kong Education Bureau, 2007),

and teachers are also reminded of timely feedback on the students' work. These comparisons show a need for the Malaysian Education Ministry to relook into our current syllabus. At the same time, universities and teacher training colleges also need to restructure their teaching courses so that effective written feedback practices can be taught to pre-service teachers.

Teachers also need to change their marking style from comprehensive to selective marking. They have been complaining about time constraints they face in schools, and perhaps selective marking would make their feedback practice easier. By focusing on certain aspects of grammatical items, teachers are not burdened to go through word by word to find errors in the students' writings. On the other hand, students may find it less intimidating to see fewer red marks on their essays. Selective marking could also help the students to stay focused when they are revising their essays. It could be done if teachers could link their written corrective feedback systematically with their grammar instruction in the classrooms.

Finally, teachers' written feedback should adhere to good feedback as proposed by earlier literature. Concerning written feedback, teachers must remember that comprehensive WCF may not always be the best. Besides taking up so much of the teachers' time, it can also overwhelm the students. Therefore, teachers need to be selective in marking errors. Teachers should tie the writing task to a certain grammatical aspect during the writing lesson. Teachers also need to remember that written commentary is not the only option. Student-teacher conferences should also be utilised to clarify their problems in completing the writing task. Moreover, both positive and negative feedback are equally important in supporting the students' learning, but they must always be linked to the task at hand, or the feedback would be meaningless. In terms of feedback timing, there is no fast rule as to when it is the best time to provide written feedback. Whether the feedback is immediate or delayed, it would be useless unless the students can revise their essays and raise their grades. Nevertheless, teachers need to consider the nature of the task and the ability of the students. As Mathan (2003) claimed, immediate feedback would be most beneficial for the student's learning if the task is difficult, but delayed feedback may be better if the task is easy. As such, delayed feedback may promote the transfer of learning better, such as in concept-formation tasks, while immediate feedback may be more efficient for procedural skills (Corbett & Anderson, 2001).

LIMITATIONS OF THE STUDY

This study is not without its limitations. Firstly, this study is limited in scope where it is to look only into the written feedback given by the teachers to students' writings. It is acknowledged that AfL covers four main components, namely questioning, feedback through marking, peer- and self-assessment, and the formative use of summative test. However, only the second component, i.e.,

feedback through marking, is examined. The other components are not being examined, although it is to be made aware that they may play a vital role in teachers' conception of feedback and their classrooms practices.

Secondly, as the participants of this study are teachers who teach in the state of Pahang, the findings of this study may not be generalised to the general population of Malaysian ESL teachers. It is because it may be almost impossible to collect data from each ESL teacher in Malaysia. Consequently, this study selected its respondents carefully so that they represent teachers from an array of different educational, experience and cultural backgrounds, to some extent, mirror the entire population of Malaysian ESL teachers.

Thirdly, since the researcher is the instrument in this study, bias may also affect the study results. Therefore, cross-checking with other raters was carried out to reduce the effect of researchers' biases. It includes cross-checking the reliability of the transcriptions before the coding process is done, the validity of the codes of the transcriptions, and overall data analysis.

ACKNOWLEDGEMENT

REFERENCES

Abdullah, H., & Sidek, H. M. (2012). L2 writing feedback: Alignment of instructional planning and implementation. *Journal of Language Studies*, *15*, 15-17.

Amrhein, H. R., & Nassaji, N. (2010). Written corrective feedback: What do students and teachers prefer and why? *Canadian Journal of Applied Linguistics, 13*(2), 95-127.

Baird, J. A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in education: Principles, Policy & Practice*, *11*(3), 331-348. https://doi.org/10.1080/0969594042000304627

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of second language writing*, *17*(2), 102-118. https://doi.org/10.1016/j.jslw.2007.11.004

Black, P. (1999). Assessment learning theories and testing system. In P. Murphy (Ed.), *Learners, learning and assessment* (pp. 118-134). Paul Chapman Publishing.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-68. https://doi.org/10.1080/0969595980050102

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). The nature and value of formative assessment for learning. *Improving Schools, 6*(3), 7-22. https://doi.org/10.1177/136548020300600304

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*(3), 267-296. https://doi.org/10.1016/s1060-3743(03)00038-9

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer

reviewers and subject matter experts. *Written communication*, *23*(3), 260-294. https://doi.org/10.1177/0741088306289261

Cohen, A. D., & Cavalcanti, M. C. (1990). Feedback on compositions: Teacher and student verbal reports. In Kroll, B. (Ed.), *Second language writing* (pp. 155-177). Cambridge University Press. https://doi.org/10.1017/cbo9781139524551.015

Cohen, A. D., & Robbins, M. (1976). Toward assessing interlanguage performance: The relationship between selected errors, learners' characteristics, and learners' expectations. *Language Learning, 26*(1), 45-66. https://doi.org/10.1111/j.1467-1770.1976.tb00259.x

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 245-252). ACM Publishing. https://doi.org/10.1145/365024.365111

Corpuz, V. (2011). *Error correction in second language writing: Teachers' beliefs, practices and students' preferences* [Unpublished Master's dissertation]. Queensland University of Technology.

Dinnen, J. L., & Collopy, R. (2009). An analysis of feedback given to strong and weak student writers. *Reading Horizons: A Journal of Literacy and Language Arts*, *49*(3), 239-256.

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal*, *63*(2), 97-107. https://doi.org/10.1093/elt/ccn023

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339-368. https://doi.org/10.1017/s0272263106060141

Enginarlar, H. (1993). Student response to teacher feedback in EFL writing. *System*, *21*(2), 193-204. https://doi.org/10.1016/0346-251x(93)90041-e

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*(1), 1-11. https://doi.org/10.1016/s1060-3743(99)80110-6

Ferris, D. R. (2002). *Treatment of error in second language student writing*. The University of Michigan Press.

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short-and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81-104). Cambridge University Press https://doi.org/10.1017/cbo9781139524742.007

Ferris, D. R. (2012). Written corrective feedback in second language acquisition and writing studies. *Language Teaching, 45*(4), 446-459. https://doi.org/10.1017/s0261444812000250

Ferris, D., & Hedgcock, J. (2005). *Teaching ESL composition: Purpose, process and practice*. Lawrence Erlbaum Associates.

Fraenkel, J., & Wallen, R. N. E. (2010). *How to design and evaluate research in education*. McGraw-Hill.

Fregeau, L. A. (1999). Preparing ESL students for college writing: Two case studies. *The Internet TESL Journal, 5*(10), 1-3.

Gurnam K. S., Chan Y. F. & Sarjit, K. S. (2011). Students' reactions to School-based oral assessment: Bridging the gap in Malaysia. *Asian EFL Journal, 13*(4), 300-327.

Hamidi, M. F. (2015). *Writing a recount for PT3* [Unpublished module]. Pejabat Pelajaran Daerah Kemaman, Kementerian Pelajaran Malaysia.

Hong Kong Education Bureau (2007). *English language: Curriculum & assessment guide (Secondary 4-6)*. Curriculum Development Council and the Hong Kong Examinations and Assessment Authority.

Hyland, K. (2003). Genre-based pedagogies: A social response to process. *Journal of second language writing*, *12*(1), 17-29. https://doi.org/10.1016/s1060-3743(02)00124-8

Hyland, K., & Hyland, F. (Eds.) (2006). *Feedback in second language writing: Context and issues*. Cambridge University Press. https://doi.org/10.1017/cbo9781139524742

Irons, A. (2008). *Enhancing learning through formative assessment and feedback.* Routledge. https://doi.org/10.4324/9780203934333

Juwah, C., Macfarlane-Dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). *Enhancing student learning through effective formative feedback*. The Higher Education Academy Generic Centre.

Kamaruddin, K. (2016). *SPM paper one (Section B: Continuous writing) practice* [Unpublished module]. Pejabat Pelajaran Manjung, Kementerian Pelajaran Malaysia.

Kementerian Pendidikan Malaysia (2017a). *Dokumen standard kurikulum dan pentaksiran Bahasa Inggeris Tingkatan 1* [Form 1 English Curriculum and Assessment Standard Document]. https://sumberpendidikan.com/2017/07/18/dskp-bahasa-inggeris-tingkatan-1-kssm/

Kementerian Pendidikan Malaysia (2017b). *Dokumen standard kurikulum dan pentaksiran Bahasa Inggeris Tingkatan 2* [Form 2 English Curriculum and Assessment Standard Document]. https://sumberpendidikan.com/2018/02/01/dskp-cefr-bahasa-inggeris-kssm-tingkatan-2/

Krashen, S. (1992). The input hypothesis: An update. In J. E. Alatis (Ed.), *Linguistics and language pedagogy: The state of the art* (pp. 409-431). Georgetown University Press.

Lalande, J. F. (1982). Reducing composition errors: An experiment. *Modern Language Journal, 66*(2), 140-149. https://doi.org/10.2307/326382

Lee, I. (1997). ESL learners' performance in error correction in writing: Some implications for teaching. *System*, *25*(4), 456-477. https://doi.org/10.1016/S0346-251X(97)00045-6

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing*, *13*(4), 285-312. https://doi.org/10.1016/j.jslw.2004.08.001

Lee, I. (2005). Error correction in the L2 writing classroom: What do students think? *TESL Canada Journal, 22*(2), 1-16. https://doi.org/10.18806/tesl.v22i2.84

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing*, *17*(2), 69-85. https://doi.org/10.1016/j.jslw.2007.10.001

Lee, I. (2009). Ten mismatch between teachers' beliefs and written feedback practice. *ELT Journal, 63*(1), 13-22. https://doi.org/10.1093/elt/ccn010

Lee, I. (2014). Teachers' reflection on implementation of innovative feedback approaches in EFL writing. *English Teaching*, *69*(1). 23-40.

Lunsford, R. F. (1997). When less is more: Principles for responding in the disciplines. *New Directions for Teaching and Learning*, *1997*(69), 91-104. https://doi.org/10.1002/tl.6908

Maarof, N., Yamat, H., & Li, K. L. (2011). Role of teacher, peer and teacher-peer feedback in enhancing ESL students' writing. *World Applied Sciences Journal*, *15*(1), 35-29.

Malaysian Ministry of Education (2003). *English language curriculum specification for Form 4*. Curriculum Development Centre, Ministry of Education.

Mantello, M. (1997). Error correction in the L2 classroom. *Canadian Modern Language Review*, *54*(1), 127-131. https://doi.org/10.3138/cmlr.54.1.127

Mathan, S. (2003). *Recasting the feedback debate: Benefits of tutoring error detection and correction skills* [Doctoral dissertation, Carnegie Mellon University]. ProQuest Information and Learning Company. https://www.proquest.com/openview/cddcc4dafe8cd89adc25a9874777cd91/1?pq-origsite=gscholar&cbl=18750&diss=y

Ministry of Education Singapore (2010). *English language syllabus primary & secondary (Express/Normal[Academic]).* Curriculum Planning & Development Division.

Nesamalar, C., Saratha, S., & Teh, S. C. (2001). *ELT methodology: Principles and practice*. Penerbit Fajar Bakti.

Nicol, D. J., & Macfarlene-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Nyquist, J. B. (2003). *The benefits of reconstruing feedback as a larger system of formative assessment: A meta-analysis* [Unpublished Master's thesis]. Vanderblit University.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science, 28*, 4-13. https://doi.org/10.1002/bs.3830280103

Reid, J. (1998). "Eye" learners and "ear" learners: Identifying the language needs of international students and U.S. resident writers. In P. Byrd & J. M. Reid (Eds.), *Grammar in the composition classroom: Essays on teaching ESL for college-bound students* (pp. 3–17). Heinle & Heinle.

Salteh, M. A., & Sadeghi, K. (2015). Teachers' and students' attitudes towards error correction in L2 writing. *The Journal of Asia TEFL, 12*(3), 1-31. https://doi.org/10.18823/asiatefl.2015.12.3.1.1

Shelly, A. (2014). *Dynamic written corrective feedback: Achieving manageability*. Brigham Young university. http://scholarsarchieve.byu.edu/etd/4109

Sidhu, G. K., Chan Y. F., & Sidhu, S. K. (2011). Students' reactions to School-based oral assessment: Bridging the gap in Malaysia. *Asian EFL Journal, 13*(4), 300-327.

Singaporean Ministry of Education (2010). *English language syllabus.* Ministry of Education.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Williams, J. G. (2003). Providing feedback on ESL students' written assignments. *The Internet TESL Journal, 9*(10), 1-5.

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, *60*, Article 101826. https://doi.org/10.1016/j.cedpsych.2019.101826

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19*, 79-101. https://doi.org/10.2307/3586773

Students' and Teachers' Perceptions on the Impacts of Service-Learning in a Language Course

Giang Thi Thu Bui* and Hanh Thi Minh Nguyen

Faculty of Foreign Languages, Hanoi University of Industry, 298 Cau Dien Street, Bac Tu Liem District, Hanoi, Vietnam

ABSTRACT

Service-learning, also known as community-based learning, is considered a pedagogical tool in various disciplines at different levels, including tertiary education. It has proved its significant effects on social and academic aspects. Different institutions have incorporated service-learning in their language curricula in language education to create better exposure to the target language for learners. However, the research on the application of service-learning components in language learning has still been limited in Asia-Pacific countries, especially in Vietnam, where learners have fewer chances of serving a native community by using a target language like English. This paper reports on teachers' and students' perspectives on the impacts of a service-learning project incorporated in an English speaking course at a public university in Vietnam. The study used questionnaires for 117 second-year English-majored students, three focus-group discussions with 16 surveyed students, and semi-structured individual interviews with four subject teachers. The results reflect students' positive feedback on language competence, social awareness, personal traits and soft skills. The interviews with teachers reveal the progressive changes in students' presentation skills, self-confidence, and critical thinking skills. However, the main difficulties hindering the implementation of the project are the insufficient understanding of project procedures, the passive way of thinking, lack of financial support, and the limited exposure to English in authentic environments. This paper also brings about some practical implications for language teachers and researchers in similar educational contexts.

Keywords: English major**,** impacts, language course, perceptions, service-learning

INTRODUCTION

Service-learning, known as community-based learning, is considered a pedagogical tool in a wide range of disciplines at different levels. This teaching strategy provides students with opportunities to apply their academic knowledge to serving a community. Since its first introduction in the United States in the 1960s, service-learning has proved its merit in making the subject more alive than what happens in a classroom setting (Brown & Purmensky, 2014). With its experiential, goal-oriented, communicative, and interpersonal nature, service-learning gives students hands-on experience and a chance to address community needs, beneficial to the community and important to them (Minor, 2001).

With its indispensable effects on bridging the theory and practice, service-learning has been incorporated in language curricula in many countries in the world, including the United States, Germany, Lybia (Suwaed, 2018), Ecuador (Brown & Purmensky, 2014), and Australia (Pazmino, 2017). It is more advantageous as, in language education, students can use their language as a tool to do the service. Therefore, they will have opportunities to read, speak about the topics, participate in discussions and write reflections using the target language (Minor, 2001). With a service component in the syllabus of a language course in their study, Brown and Purmensky (2014) found students' positive perceptions about the relationship between service-learning and the development in their linguistic and cultural competency. Service-learning is also proved to enhance students' soft skills, promote their values and self-worth, and help orient their future career choices (Burgo, 2016; Jouët-Pastré & Braga, 2006; Suwaed, 2018).

In Vietnam, service-learning also receives profound attention with the development of service-learning organizations which engage volunteers in community-based activities. Moreover, in realizing of the great impact this learning strategy brings, some educational institutions have incorporated a service component in their curricula (Nguyen et al., 2012). However, most of these courses are non-linguistic, and thus studies on the relationship between service-learning and language education have hardly been found. Moreover, much of the current literature on service-learning pays particular attention to the benefits of this pedagogical tool from students who directly take part in the activity, not the teachers.

This study, therefore, aims at investigating the impact of a service component in a particular language course on primarily students' language development and, secondarily, their social awareness from both students' and teachers' perspectives. The research is conducted by incorporating a service-learning project in a speaking skill course at a public university with second-year English-majored students and the course lecturers. Thus, the study highlights teachers' and students' subjective perceptions on how service-learning project has benefited students' language learning and their social background knowledge.

Moreover, the findings are supposed to reveal some challenges during the implementation of the project and some suggestions from both teachers and students to maximize the efficiency of the project. Finally, the study seeks answers to the following questions to achieve these objectives:

Question 1. What are the linguistic and social benefits the service-learning project brings about?

Question 2. What difficulties do students and teachers encounter during the implementation of the project?

Question 3. What do teachers and students suggest to maximize the efficiency of the project implementation?

This paper is expected to contribute to the literature on the relationship between service-learning and linguistic as well as cultural competence in Vietnam. Furthermore, findings from this study help encourage practitioners and teachers to integrate service-learning activities in language curricula or language syllabi. An innovative teaching method can promote students' academic excellence and their sense of responsibility to the community to meet the requirements of global human resources in the 21st century.

LITERATURE REVIEW

Definitions of Service-Learning

A literature review on service-learning reflects numerous ideas on how different authors define the term. According to Jacoby (1996), service-learning is a form of experiential education that involves students in design activities that address human and community needs and promote student learning and development. In other words, this is an opportunity for students to apply their academic knowledge in real life through a designed activity that allows them to reflect on their practical experience. This definition is close to those of Heuser (1999) and Minor (2001), who defines service-learning as the combination of community service and academic course work. It is to say that students will partake in some kinds of activities in their communities while applying their academic knowledge and skills to meet identified needs of those communities. Pazmino (2017) considers service-learning as the immersion in the community of the target language as a service process with the goal of learning, which, reversely, involves service.

In other words, service-learning is one approach to learning which allows learners to be exposed to the target language by serving the community. O'Connor (2012) provides different examples of service-learning projects, which include students serving as conversation partners, volunteering as interpreters at local hospitals, clinics, schools, or social service agencies; tutoring or mentoring Spanish-speaking children and adolescents; facilitating Spanish story hours and other cultural activities for children in libraries and community centers; assisting in after-school programs for children of all ages, elaborating publicity materials to raise funds for associations, and working in HIV prevention programs. To carry out those projects, students prepare for placements in class, participate in community activities,

and reflect on the experience and how it might have enhanced their language, cultural and social skills. These examples highlight the various contexts in which students can use their academic knowledge to contribute to the community in the form of service. However, it is not easy for English learners in Vietnam to find a community whose native language is English. Therefore, it would be more challenging to have exposure to the target language in an English-speaking community with specific needs to address. Thus, service-learning means that students are allowed to participate in a community service which to some extent requires them to use English during the implementation of a specific project.

Components of Service-Learning

As highlighted in the definition, desspite having different shapes, service-learning, in general, has two components: some community service and related classroom instruction (Minor, 2001). It means the service that students participate in must be related to the academic knowledge or skills they learn in class. Warschauer and Cook (1999) also agree that service-learning consists of two parts, yet besides the participation in activities that both benefit the learners and the community (reciprocity), they mention reflection as an integral part which requires learners to reflect on the benefits occurring from collaborative discussion and on the experience. Thus, it coincides with the idea suggested by Heuser (1999) that participation in community and reflection on the participation and the

connection of that experience to class-based knowledge are the main components of a service-learning project. Munz et al. (2018) also emphasize reflection as one of ten best practices of integrating service-learning in the public speaking course. Accordingly, students' reflections can be combined with their assessments of any course assignment that incorporated the service. At least they should be assigned writing or speaking tasks to reflect on how the service component helps them achieve the course learning outcomes. Pazmino (2019) mentions the service-learning model that connects these components in five steps of (1) exploration, (2) clarification, (3) realization, (4) activation, and (5) internalization. The three first steps help students understand the nature of service-learning and guide them to the ideas of their service-learning project, while the two last focus on their experience and reflection on their participation in the project.

Benefits of Service-Learning

There has been widespread research into the benefits of service-learning for students, which generally supports the claim that this approach positively affects linguistic and cultural skills (Brown & Purmensky, 2014; Burgo, 2016; Heuser, 1999; Jouët-Pastré & Braga, 2006; Pazmino, 2017). Furthermore, service-learning projects also create great opportunities for students to promote their values, self-worth, confidence, and motivation in using the target language, develop new perspectives, improve the relationship with the community (Burgo,

2016; Minor, 2001; O'Connor, 2012; Pazmino, 2017; Suwaed, 2018). This innovative approach, moreover, has been claimed to raise students' awareness of social problems, increase content knowledge about social issues, encourage active learning and creativity, develop communication skills, high-order thinking skills, teamwork and help reevaluate their career plans (Beckman, 1997; Jouët-Pastré & Braga, 2006; O'Connor, 2012; Pazmino, 2017; Suwaed, 2018; Warschauer & Cook, 1999).

Challenges in Service-Learning

Suwaed (2018) and Pazmino (2017) mention timing as the first challenge to implementing a service-learning component in a course. Most service-learning projects are carried out in a limited time, which should be extended to maximize the efficacy of the activities. Suwaed (2018) stresses students' lack of confidence in teaching in a children's teaching project. In contrast, Pazmino (2017) emphasizes the lack of stimulants for the conversations between participants groups and the excessive number of participants in the project to help members of a Latin American community expand their knowledge of English language and enable them to connect with Australian people and culture. Heuser (1999) raises an ethical issue on how those short-termed activities can be performed in a manner that is not patronizing or disrespectful and carried out to promote content and language learning. In a word, more time and support are needed due to both linguistic and cultural challenges. However, these challenges arise in language courses where students communicate in the target language in an authentic setting with native speakers. Moreover, the fact is that service-learning projects usually involve non-native English-speaking students performing service when they study abroad (Brown & Purmensky, 2014). This research on service-learning was carried in Vietnam, where authentic settings with native speakers are not always available for participants. Therefore, there may be other difficulties that arise in this different context.

METHODS

Setting and Participants

The study was carried out at a public university in Vietnam. The course under investigation is the third among five English-majored students speaking courses, designed based on the CDIO approach to meet social needs. Accordingly, the course is designed with three learning outcomes, two of which can give a presentation and develop arguments relatively effectively. The service-learning project is part of the assessment in this course which aims to achieve those outcomes. Specifically, the project was carried out from week 2 to week 12 of the 15-week course. Students were required to work in groups and develop a project to apply English to benefit the community. The project procedures were explained on the first day of the course. All the project documents were then sent to students. Two weeks after the procedures of the project were introduced, students were required to submit their project plan, which

outlines the community's problem and the possible solutions. Every three weeks, students had to send a written report to teachers for feedback and support. Finally, in week 13, students presented their project in class following the project requirements for the presentation. This study, therefore, was carried out after they finished the course.

The participants were 117 second-year English-majored students and all four teachers of the investigated speaking course at a public university in Hanoi, Vietnam.

Data Collection Instruments

The study employed quantitative and qualitative methods through survey questionnaires, focus-group discussions with students, and semi-structured interviews with teachers.

Survey Questionnaires. The questionnaire includes 12 questions which are categorized into three parts. The first part (questions 1–6) explores students' experience and understanding of service-learning activities. The second part (questions 8–11) focuses on the project's benefits to students. The last part (question 12) uncovers the difficulties they encountered during the project.

Focus-group Discussions. The discussions, guided by five questions, aim to acquire additional information on how students view the project's impacts on their learning and other social or cultural aspects, clarify the difficulties they mentioned in the survey, and

elicit their suggestions for further project implementation.

Semi-structured Interviews. The interviews with subject teachers are based on six questions which can be classified into four main points. The first question centers around the impacts of the service-learning project on students' academic performance, language skills, personal traits and other possible aspects. The second and the third questions aim at exploring the obstacles hindering both teachers and students during the implementation of the project. The two next questions, namely questions four and five, focus on what the four teachers did to support their students during the project and things students consulted about. The last question addresses teachers' suggestions for better improvement of the project in the upcoming semesters.

Data Collection and Analysis Procedures

The survey questionnaires were first delivered online to collect data to 175 second-year English-majored students one week after finishing the course. The time limit for completing the questionnaires was two weeks since the delivery. After two weeks, 117 students responded to the questionnaires. The data were then collected for analysis. After the survey, the focus group discussions were carried out with 16 students who provided their contact details for further study in their questionnaires. These students were divided into two groups

of 5 and one group of 6 for each discussion. At the same time, four subject teachers were interviewed to collect more data. Each teacher interview and student focus-group discussion were conducted within 45–60 minutes in places convenient to participants. Each student taking part in the in-depth interviews was coded from S1 to S16 and each teacher from T1 to T4 to ensure the confidentiality of the research.

The data collected were then coded and analyzed in three themes: the understanding of participants about service-learning activity, the benefits it brings about, and the difficulties participants encountered during the implementation of the project.

RESULTS

Students' Understanding of and Experience with Service-Learning before Implementing the Project

The first part of the questionnaire consists of six questions to examine students' understanding and experience with service-learning before carrying out the project. These questions aim to determine whether or not students have participated in activities that benefit the community whether or not they have used knowledge or skills, especially English learned in school, to benefit the community. Other questions are to explore their understanding of the size and scope of a community. The answers to these questions may show how their understanding and experience with service-learning affect how they implement the project.

The first two questions revealed a different understanding of the size of a community and what it is. The answers were varied, yet most of them viewed a community comprising at least three people who do the same activity (26.7%) or live in the same local area (25.9%). A small number thought that these are people who study in the same class (15.5%), befriends (13.8%), or colleagues (6.9%). It means that students can locate the community that they may work within the project. Regarding the experience with community service, half of the students reported they participated in some of these activities. While 60.3% confirmed their knowledge and skills learned at school benefit people in the community before carrying out the service-learning project, 58.1% said they never used English to benefit people in their community. It is to say that they had experience with using their target language in community service. When asked about the types of activities that can benefit the community, most chose to volunteer (60.3%). Others fell for recycling, charity, and tutoring at low rates (18.1%, 10.3%, and 8.6%, respectively). These answers show that students may have more experience with voluntary activities in which they used their knowledge, skills, and English to support the community.

Students' Perspectives on the Impacts of the Project

Benefits of the Service-Learning Project. The second part of the questionnaire is to determine whether or not and to what

extent the service-learning project has any positive or negative impact on students' academic and social or cultural awareness. As shown in Figure 1, it is noted that 56.9% of students confirmed that the experience in doing this project had a very positive or positive effect on their study, 35.3% chose neutral, and the percentages for negative and very negative are minimal.
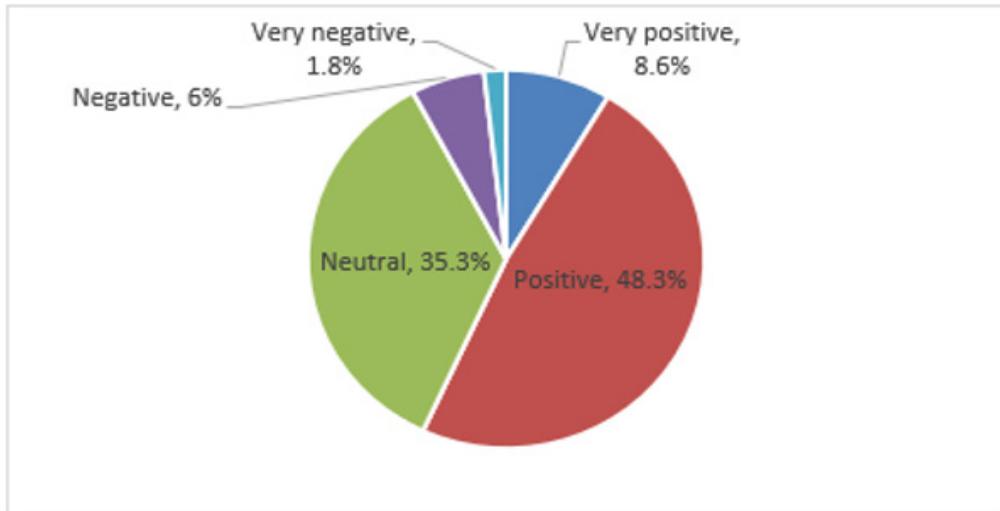
Figure 1. Students' general view of the impacts of the project

The next question focuses on the impact of the service-learning project on participants' academic performance. Again, a high percentage of 91.1% felt an improvement of academic knowledge and skills while the rest was for gains in their subject grades. Students in the focus group discussion back it up. In particular, all interviewed students mentioned a considerable enhancement in vocabulary due to their exposure to many reading materials they worked on for the project. One student (S8 in group 2) even emphasized that her lexis level increased, and she got to know higher-level vocabulary thanks to her readings:

"As our teacher encouraged us to use more vocabulary at B2 level, I now know more words at this level which I did not pay attention before the project. In order to find the solutions to the problem of the community, we had to read a lot, so we came across many interesting words and phrases. Our group shared vocabulary about environment and learnt interesting structures to describe the process of making a thing."

Besides academic performance, students also confirmed the positive changes in personality traits due to their participation in
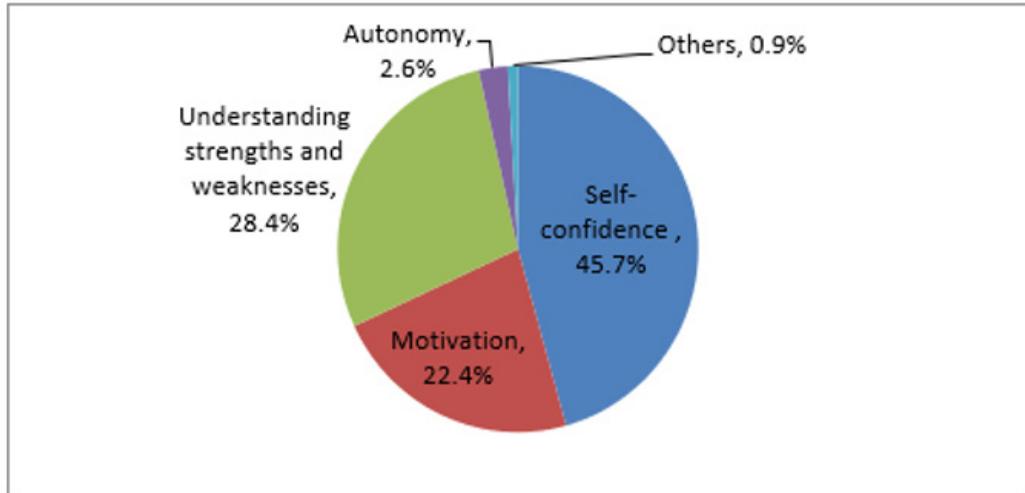
Figure 2. Students' views of the impacts on personal traits

the project. Figure 2 indicates that nearly half of the participants said their self-confidence improved (45.7%). Understanding personal strengths and weaknesses and enhancing motivation in learning English comes next with 28.4% and 22.4%, respectively.

The next question explores the impacts of the service-learning project on students' skills which are shown in the following Table 1.

Table 1

Students' views of the impacts on skills

Skills	Percentage (%)
Life skills (e.g., communication, negotiation, problem solving, critical thinking….)	40.9
Ability to work and learn in groups	33.9
Occupational skills (e.g. searching for information, synthesizing, translating, presenting)	22.6
IT skills	2.6

The results show life skills such as communication, negotiation, problem-solving, and critical thinking as the most improved (40.9%). The ability to work and learn in groups comes next with 33.9%.

In the focus group discussions, some students admitted that as they had to work in groups to complete the project, they had to collaborate and cooperate with others despite their different personalities. As a result,

their teammates' understanding improved, and their relationships strengthened. For example, S14 from group 3 comments:

"Sometimes our opinions contradicted and we argued for the solution. And finally, we tried to understand the other's ideas and agreed on a common solution."

Besides, 22.6% agreed that there was an improvement in occupational skills, including searching for information, synthesizing, translating, and presenting. Data from the discussions showed that students were more aware of making a more effective presentation through group discussions about presenting their project on the presentation day in the most compelling way. S5 cited this advantage in group 1:

"While we were discussing how to present our project, we showed each other what and how to talk about each part of the presentation. When one mentioned a new word or even a new idea, we asked for clarification which means we could learn from each other. Besides, we also taught each other about pronunciation and how to show confidence while speaking."

Other students also asserted that the appropriate preparation time for the project and their active rehearsal before the presentation day improved their presentation skills. In addition, it is noticeable that among students' skills shown in Table 1, IT skills were reported with the lowest improvement (2.6%).

Concerning the changes to social and professional aspects, in their answers to question 11, 35.3 % confirmed more awareness of social issues, 25.9% reported an expansion of their social network, 15.5% said their sense of social responsibility to the public was increased. In addition, others reported more understanding of social/cultural differences (14.7%).

Difficulties Encountered during the Implementation of The Service-Learning Project. The last question focuses on the obstacles hindering the implementation of the project. The answers were varied, yet the most selected in the questionnaires was insufficient understanding of the steps to carry out the project and the difficulty in finding the possible solutions to the identified community's problems (18.3% and 17.4%, respectively). Also, in focus group discussions, students admitted that instead of consulting their teachers, they asked some other members in their class. Thus, they continued to carry out the project in the way they thought it should be conducted, which caused misunderstanding and affected their project's progress. In addition, other students added an obstacle related to the scope of the solution for their identified community's problem.

As the problem they chose to address was littering, and the proposed solution was recycling, they could not identify the focus of the solution and decided to make a wide range of recycled products which

caused them a heavy workload in some first weeks of the project. In addition, 12.2% reported that they lacked financial support, which caused different views among group members to deploy the next steps for their project. One respondent in focus group 3 highlighted the need for budget spending on making the products impressive with color-printed posters and buying materials and some equipment needed to make the final products such as stationaries or decorations. More interestingly, the focus-group discussions revealed trouble in the use of English. In particular, students S10, S12 from group 3 commented that their group had trouble in making the most use of English during the project process as they did not have excessive exposure to English:

"We were encouraged to make the most use of English in every stage of the project so that we could have more exposure to English. The community we chose were not native speakers, therefore we ourselves had to create our own opportunities to use English during the whole project, not only in the presentation."

Teachers' Perspectives on the Impacts of the Project

Four subject teachers were invited to the semi-structured interviews to explore how the service-learning project impacted their students' learning and other possible aspects such as social or cultural understandings. The interviews also revealed teachers' assistance towards students and addressed the difficulties they and their students encountered during the project and their suggestions for better implementation.

In terms of the benefits, all four teachers commented that their students showed better confidence and presentation skills on the showcase day compared to how they did in other class activities, as teacher T1 explained below:

"My students got more confident on the showcase day. Moreover, the way they presented was easier to follow and more attractive. This is rewarding as normally many of them were quite shy in class activities."

They also confirmed a wider range of words used in the regular reports and the presentations. The closer relationship between teacher and students resulting from the frequent consultancy and feedback and good cooperation among group members shown in the reports and presentations were also mentioned by teachers T1 and T3.

Teacher T2, in addition, mentioned a significant change in students' way of thinking. Through consultancy and feedback, her students showed greater autonomy, adaptation, and improved critical thinking skills, which was demonstrated in teacher T2's comments:

"As students received my feedback, they became more autonomous, active in finding other direction for their project. They even knew how to self-evaluate their work, ask appropriate

questions and report the results of their work."

With regard to the difficulties, teachers T2, T3, and T4 emphasized the issue of making students more critical about their work. Many students showed passive thinking through their reports and vague answers to teachers' guiding questions. They revealed their weaknesses in processing, selecting, and evaluating information needed for their project. Therefore, the ideas of the community's problems and the possible solutions were vague or general, and they got stuck in figuring out what to do in the following steps. It could explain why all four subject teachers admitted that students consulted them the most about the issues happening in the community and relevant solutions to them.

As a result, the four interviewed teachers also emphasized their strong support during the project to assist students in overcoming the barriers. All teachers frequently gave very detailed and precise feedback to each group based on students' group reports. The common problems in doing the project, such as report writing, group work distribution, were synthesized and explained by the teachers in front of the class. Another challenge was related to the assessment of the project. Teachers T2 and T4 found it challenging to individualize the assessment. As the project was the result of the whole process with an individual contribution to group work, it was challenging to make a fair assessment based on their efforts

during the ten weeks of the project and each student's contribution.

To limit the difficulties for the subsequent implementation of the service-learning project, teachers suggested making clear for students how specific their project should be and how critical they should be about the project process. A model sample of the project well-completed by former students was suggested to be introduced on the first day of the course. This sample would also be analyzed and commented on to figure out the strengths and weaknesses of current students' better projects. In terms of the assessment, it was recommended that after the completion of the project, each student needs to write a reflection in which they would report about what they did, learned, and how much they contributed as the project progressed, which would then be the reference for the final assessment. Besides, teachers' strong support during the project's implementation should be maintained. The written reports are suggested to change into oral reports in class to create more opportunities for students to speak English.

DISCUSSION

It was clear from the results that the service-learning project had considerably positive effects on students' academic performance, personal traits, soft skills, and social awareness. These findings partly match Chiva-Bartoll et al. (2020), whose research affirmed that the application of a service-learning program had a positive impact on prosocial behavior and perceived academic

learning of students with a reciprocal relationship. Furthermore, it was shown in the survey, focus-group discussions, and interviews that students had gained English language knowledge (mainly vocabulary) and skills since they had to work with various materials and discuss to get ideas for their projects. Besides, life skills, such as communication, negotiation, problem-solving, and critical thinking skills, were improved due to the project. Students' thinking became critical, and they tended to be more flexible to reach the solutions as the project progressed, thanks to teachers' comments and guidance. Moreover, the students' occupational skills such as translating, synthesizing, presenting have changed positively.

Since students had to search for the information and do readings, they had to translate, synthesize and decide how to present the information in the most effective way for their project in the showcase. In terms of the personal traits, confidence and motivation in learning English were on top due to their active engagement in making the way out and considerable efforts for rehearsing their presentations. Therefore, the more they indulged in English, the more interesting they found it is. Furthermore, the more they practiced, the more confidently they presented. Finally, many students confirmed in the survey that they became more aware of social issues and had more relationships upon completing the project. These results align with Capella-Peris et al. (2020), who found out that applying

a service-learning program stimulates academic learning and many social aspects.

Regarding this impact, teachers revealed that most groups chose to deal with the problems of littering or disposal of plastic bags/bottles—things related to environmental protection. Therefore, students might better understand environmental issues or other social aspects by identifying the community's problems. It is aligns with students' understanding of the nature of a community and community-benefited activities raised in the first part of the survey. Many students defined community as people doing the same activity or living in the same local area and mostly chose to volunteer as a beneficial activity. They may prefer things like environmental problems in local areas rather than other fields such as education or business.

As shown in the survey and focus-group discussions, most students encountered difficulty understanding the project steps, especially when they came to the stage of finding feasible solutions to the identified community's problems. Teachers also confirmed that students were not critical enough to figure out their teachers' comments and feedback. It seemed hard for them to think out of the box and redirect their projects without teachers' guidance. The lack of communication, negotiation and teamwork might significantly hinder the project process. Another interesting fact revealed was the limited exposure to English in its authentic environment.

Commonly, service-learning activities are undertaken in a target language community. However, this project was implemented in an environment where English native speakers are mainly tourists or foreigners working in companies that are hard to contact. It is a noted point that has not been widely mentioned in previous studies on service-learning. Therefore, the non-target language environments should be taken into consideration as the project commences. In addition, insufficient financial resources were another concern. Due to the expenses on materials, equipment, or decorations needed to complete the final products, students had to reconsider how they continued their project. As for teachers, it was challenging to individualize the assessment, which shows students' efforts during the whole process.

IMPLICATIONS

The study revealed positive perceptions of students and teachers about the impacts of service-learning projects on students' learning and social awareness. However, in order to ensure better implementation of the service-learning project, some following issues should be taken into considerations. Firstly, before students start to do their own project, a model sample of a complete project describing each step in details, from the identification of the community's problem to the feasible solutions for it, should be introduced and analyzed to students to make sure they have thorough understanding of what they are expected to do. Secondly, during the process of the project, as teachers

give comments and feedback based on students' regular reports, they should guide them in a way that can promote their critical thinking and cognition, making them more active, evaluative, and flexible in later stages of the project so that the project they do is still meaningful, cost-effective and encourages students themselves to create their opportunities to English exposure in many non-target language environments as in Vietnam. Furthermore, developing critical thinking should not be limited to the project in this speaking course but any language learning activity. Thirdly, to ensure a more precise individual assessment, it is suggested that each student should write a reflection on the whole project process describing what they did, how they contributed, and what they learned. It will then be served as the reference together with other criteria to assess the results for each student.

CONCLUSION

This study attempted to investigate the impacts of a service-learning project perceived by teachers and students in a speaking course. Based on students' survey questionnaire, focus group discussions, and teachers' semi-structured interviews, the findings showed that the service-learning project positively influenced students' academic performance, personal traits, soft skills and social awareness. The advantages they got due to the completion of project such as the boost in confidence and motivation, the enhancement of vocabulary level, communication, negotiation, presentation, and especially critical thinking skills are

not only beneficial in learning English speaking but also in other language skills. The changes in social awareness are also essential to students in the era of integration and globalization. Finally, the implications on the clarity of project procedures, teachers' comments and feedback promoting students' critical thinking, and students' reflections will hopefully improve the subsequent implementation of the service-learning project.

ACKNOWLEDGEMENT

REFERENCES

Beckman, M. (1997). Learning in action courses that complement community service. *College Teaching*, *45*(2), 72-75. https://doi.org/10.1080/87567559709596195

Brown, A. V., & Purmensky, K. (2014). Spanish L2 students' perceptions of service-learning: A case study from Ecuador. *International Journal of Research on Service-Learning and Community Engagement*, *2*(1), 78-94. https://doi.org/10.37333/001c.002001007

Burgo, C. (2016). Service-learning for students of Spanish: Promoting civic engagement and social justice through an exchange tutoring service. *Revista de Lingüística y Lenguas Aplicadas*, *11*, 11-18. https://doi.org/10.4995/rlyla.2016.4133

Capella-Peris, C., Gil-Gómez, J., & Chiva-Bartoll, Ò. (2020). Innovative analysis of service-learning effects in physical education: A mixed-methods approach. *Journal of Teaching in Physical Education, 39*(1), 102-110. https://doi.org/10.1123/jtpe.2019-0030

Chiva-Bartoll, O., Ruiz-Montero, P. J., Capella-Peris, C., & Salvador-García, C. (2020). Effects of service-learning on physical education teacher education students' subjective happiness, prosocial behaviour and professional learning. *Frontiers in Psychology, 11*, Article 331. https://doi.org/10.3389/fpsyg.2020.00331

Heuser, L. (1999). Service-learning as a pedagogy to promote the content, cross-cultural, and language-learning of ESL students. *TESL Canada Journal, 17*(1), 54-71. https://doi.org/10.18806/tesl.v17i1.880

Jacoby, B. (1996). Service-learning in today's higher education. In *Service-learning in higher education* (pp. 3-25). Jossey-Bass Publishers.

Jouët-Pastré, C., & Braga, L. J. (2006). Service-learning in language education: Bridging the university-community divide. *Community Psychology Connection Special Section. The Community Psychologist*, *39*, 32-34.

Minor, J. L. (2001). Using service-learning as part of an ESL program. *The Internet TESL Journal*, *7*(4).

Munz, E. A., Gatchet, R. D., & Meier, M. R. (2018). Integrating service-learning in the public speaking course. *Journal of Communication Pedagogy*, *1*(1), 115-123. https://doi.org/10.31446/jcp.2018.19

Nguyen, T. H., Phung, T. P., & Dong, T. B. T. (2012). *Enhancing students' civil responsibility and hands-on experience through service learning. University of Science – Vietnam National University Ho Chi Minh City*. https://vietnamservicelearning.wordpress.com/2012/08/26/haoec-tap-pha/

O'Connor, A. (2012). Beyond the four walls: Community-based learning and languages. *The Language Learning Journal*, *40*(3), 307-320. https://doi.org/10.1080/09571736.2011.585402

Pazmino, M. F. A. (2017). *How service-learning improves language-learning outcomes for Spanish language students* (Doctoral dissertation, The University of Melbourne). Minerva Access. http://hdl.handle.net/11343/208854

Pazmino, M. F. A. (2019). Service learning education improvement for foreign language learners. *Journal of Language Teaching and Research*, *10*(6), 1173-1180. http://dx.doi.org/10.17507/jltr.1006.05

Suwaed, H. (2018). The path less taken: Incorporating service-learning in the English language curricula. *IAFOR Journal of Language Learning*, *4*(1), 25-40. https://doi.org/10.22492/ijll.4.1.02

Warschauer, M., & Cook, J. (1999). Service learning and technology in TESOL. *PROSPECT-ADELAIDE-*, *14*(3), 32-39.

Mapping IIUM Students' English Language Writing Proficiency to CEFR

Mohd. Khairul Abu Sufi* and Engku Haliza Engku Ibrahim

English Language Division, Centre for Languages and Pre-University Academic Development, International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia

ABSTRACT

The goal of this research was to determine the English language writing skills of the International Islamic University of Malaysia (IIUM) graduating undergraduates based on two linguistic levels; the IIUM English Proficiency Test (EPT) and the Common European Framework of Reference for Languages (CEFR), based on two descriptors: IIUM's EPT descriptor, and CEFR's 'General linguistic range' descriptor (Council of Europe, 2001, p 110). The EPT results show that the majority (48 %) of the English language writing skills of IIUM undergraduates were in Band 6, while Band 5.5, Band 5 and Band 4, respectively, ranked 31.3%, 4.9% and 0.4%. On the other hand, Band 8, Band 7 and Band 6.5 were attained by 0.2%, 2.4% and 12.2% of undergraduates. A significant proportion of IIUM undergraduates (46.3 %) were found to be at Level B2 + when assessed against the CEFR scale for the 'General linguistic range.' At the same time, 2.2%, 12.4%, 33.9 and 5.2% of students were at Level C2, C1, B2, B1 +, respectively. The results also showed that a mere 0.2% was at Level B1. This research also showed that despite some inaccuracies and improper uses, IIUM students could write effectively and understand and use reasonably complicated language, particularly in familiar situations. Based on the CEFR scale for 'General linguistic range', the results of this study show that IIUM graduating undergraduates could clearly articulate themselves in their writing.

Keywords: CEFR, English placement test, second language, writing proficiency

INTRODUCTION

At institutions of higher learning in Malaysia, such as at the International Islamic University Malaysia (IIUM), being adept in the English language reassures students that they are well prepared to engage in their academic pursuits. In IIUM, entry into the faculties is often determined by students' score in a proficiency test such as the internationally acclaimed TOEFL (The Test of English as a Foreign Language) and IELTS (International English Language Testing System), or the in-house administered EPT (English Proficiency Test) (http://www.iium.edu.my) as students do not necessarily sit for the Malaysian University English Test (MUET) upon entry into the institution due to the international nature of the university.

This language policy is congruent with research that state that a certain level of proficiency in the language is a requirement for effective involvement in academic studies (Deygers et al., 2017; Singh, 2016). Accordingly, remedial English classes are usually offered to those who do not meet the minimum entrance requirement in the proficiency test. In contrast, those who do would advance to their respected faculties at the International Islamic University Malaysia. The practice of offering remedial English classes for less proficient students can also be observed in other public institutions of higher learning in Malaysia, such as Universiti Kebangsaan Malaysia (UKM), Universiti Putra Malaysia (UPM) (Ming & Alias, 2007) and Universiti Tun

Hussein Onn Malaysia (UTHM) (Noor & Kadir, 2007).

In Malaysia recently, the importance of being proficient in the language has resurfaced (Ali, 2013; Mohamed, 2008; Tan & Miller, 2007). Ali (2013), in her study, exemplifies this by attesting that in the more remote parts of the country, English does not play a significant role in the daily lives of the people. The language is generally heard and spoken only in English language classes and lessons in learning institutions. Disparities in opportunity and motivation to learn and use English between urban and rural learners have affected Malaysia's educational outcome (Tan & Miller, 2007). A published MUET result analysis for the 2007/8 university intake surprisingly revealed that a large fraction (73%) of the test takers fell within the Bands of 1 (extremely limited user), 2 (limited user) and 3 (modest user), even after eight years of its introduction (MUET) into the educational system. Such revelation indicates that students admitted to Malaysian public institutions of higher learning possessed an alarmingly low level of proficiency in the English language (Mohamed, 2008).

In light of this matter, the then Malaysian Prime Minister, while tabling the country's budget in 2014, under Measure 8: Enhancing Graduate Employability, Point No. 95 (ii), announced that a certain level of English language proficiency must be met as a graduation requirement from public universities in Malaysia (Ministry of Finance, 2014). It was also asserted that

this proficiency would be measured through the results of MUET administered by the Malaysian Examination Council. A band ranging from 3 to 5 (based on disciplines) must be attained for undergraduates to be conferred their degree. This concern is well established given the growth of the Malaysian economy in the era of globalisation, and being able to use the English language fluently is deemed essential for occupational purposes in Malaysia (Shakir, 2009).

The rationale behind the implementation of the new language requirement policy is in line with the Malaysian government's strategy to enrich the English language proficiency further and to equip Malaysian undergraduates in meeting and facing the challenges of globalisation (Ganapathy, 2015; Lee, 2015; Llurda, 2013; Samuel & Bakar, 2008; Shakir, 2009; Tajuddin, 2015). The outcome of such a scheme would result in heightened confidence of graduates and at the same time, prepare them for the workforce upon successful completion of their academic programmes at institutions of higher learning. As such, all Malaysian public institutions of higher education must abide by the new English language policy set forth by the Malaysian Government documented as "English Language Education Reform in Malaysia: The Roadmap 2015-2025" by the year 2025 (Don, 2015). The roadmap also stipulates adopting the Common European Framework of Reference for Languages (CEFR) into the Malaysian education system in profiling students' English language proficiency.

Although the IIUM stipulates EPT Band 6 as the language requirement to commence learning, EPT results can only be comprehended by officials at the IIUM. Thus, a mapping of the EPT's writing score bands to the CEFR 'General linguistic range' illustrative descriptor will need to be established to demonstrate students' linguistic profile in writing. This will also complement the EPT results enabling parties outside the scope of IIUM to understand and decipher EPT scores on a common scale of reference providing a universal overview of the IIUM students' English language proficiency level. In so doing, a gap in research can be filled as this study attempted to examine IIUM final year students' English language writing ability and expected that an alignment of the students' EPT writing bands could be made to the internationally recognised CEFR proficiency scales.

Three (3) research questions were formulated for this research:

1. How do IIUM students perform in the scale of writing for EPT?
2. How do IIUM students perform in the scale of writing through CEFR?
3. Is there a relationship between IIUM students' writing performance in the EPT and CEFR?

LITERATURE REVIEW

Common European Framework of Reference for Languages

The Common European Framework of Reference for Languages (CEFR) plays a vital role in language education and policy within Europe and worldwide. The

framework has become significant for language testers and examination boards worldwide as it assists language planners to define language proficiency levels and to decipher them into meaningful language credentials. For many language testers, it has become imperative for their exams to align with CEFR (Gyllstad et al., 2014; Harsch & Hartig, 2015; Nunan, 2014; Taylor & Jones, 2006). The Council of Europe has endeavoured to facilitate this by providing a toolkit of resources, including a draft pilot Manual for relating language examinations to the CEFR and a technical reference supplement (Council of Europe, 2020).

Based on theories of language competence (Finch, 2009), CEFR also aims to enhance transparency and mutual recognition of qualifications by providing an explicit set of objectives, content and methods as well as giving objective criteria for describing language proficiency (Council of Europe, 2001). It can be said that the impartial standards for describing language proficiency simplify the mutual recognition of qualifications extended in different learning contexts, ensuing the facilitation of European mobility. Furthermore, in describing levels of language proficiency, the framework intends to be user-friendly and accessible to practitioners, helping users consider the meaning of competence in their particular teaching context (Finch, 2009).

Apart from cataloguing one's language proficiency, the CEFR provides a common basis for elaborating language syllabuses, curriculum guidelines, examinations, and textbooks across Europe (Finch, 2009).

It also designates what language learners have to do and what knowledge and skills they have to develop. Also stressed in the framework is the appropriate language to be used based on cultural contexts and communication. The CEFR also defines levels of proficiency, which allow learners' progress to be measured at each stage of learning and on a life-long basis. In language testing, the CEFR has gradually been adopted and is known today as an important instrument. Indeed, Little (2007) states that the impact of the CEFR on language testing by far outweighs its impact on curriculum design and pedagogy.

In the Malaysian context, the decision to utilise the CEFR as a gauge in identifying one's language fluency, according to Don (2015), lies within the fact that the framework has had careful considerations in its development and that the targets of proficiency postulated by the CEFR are somewhat realistic. Moreover, backed by numerous research, the CEFR fits into the need of Malaysia in establishing an English language standard that is universally acknowledged. Also, the nature of the framework (CEFR) is such that users are free to customise and adjust the requirements and define what one needs to achieve to be put into any one of the bands of the CEFR proficiency scale (Council of Europe, 2020).

Previous CEFR Mapping Studies

Studies in mapping CEFR to individual tests have been conducted in different contexts inside and outside the European region. An example of one such study exists in

the educational context of Thailand where an attempt to map the Chulalongkorn University Test of English (CU-TEP) to the CEFR through a standard-setting procedure (Wudthayagorn, 2018). The study reported that students were able to receive scores based on both CU-TEP and CEFR standards.

Another study outside Europe was conducted in Taiwan, where the CEFR was also adopted into the educational system. In this study, it was reported that the reading component of the General English Proficiency Test (GEPT) needed an alignment to CEFR, and hence, a mapping through a standard-setting session was conducted to establish the association between GEPT and CEFR (Wu & Wu, 2007). Results of the mapping study indicated a congruency between the GEPT reading test and CEFR as the degree of abstractness of the texts increases as the GEPT level rises, as does the vocabulary used, similar to the conventions of CEFR.

In the European context, a mapping study was also conducted to contextualise the Dutch foreign language examinations to CEFR (Noijons & Kuijper, 2006). By employing various methods, including familiarisation, specification, standardisation and validation, the study revealed that it was possible to map the Dutch foreign language examination to CEFR through the prescribed methods.

Another more recent study was conducted by the Educational Testing Service (ETS), where there was a need to map the TOEFL iBT test scores to CEFR. In response to the feedback from university administrators, they indicated that most universities in Europe now utilised CEFR levels for admission and called for the TOEFL iBT scores to be mapped to CEFR. The study reported that a standard-setting session was also used to establish the link between TOEFL iBT scores and CEFR.

Similar to this study, many nations have conducted mapping studies to investigate the congruency of their language assessment to the stipulations of CEFR. In sum, it is safe to acknowledge that this study parallels many CEFR mapping studies that have been conducted globally, as this study aimed to establish a connection between the in-house administered EPT test scores and the CEFR.

Writing

Without a doubt, writing is reflected as a vital and the most cognitive of all language skills. Written language is viewed as totally different from spoken language both in its form and use, although its basis depends on the language's same linguistic feature (Weigle, 2002). The aim of being able to write fluently goes beyond the ability to present information in written form. Weigle (2002) affirms that the ultimate goal of being able to write, for a student, is to be able to "participate fully in many aspects of society beyond school, and for some, to pursue careers that involve extensive writing" (p. 4).

As writing is not an easy skill to learn, it is usually tested to measure one's ability and performance in a language course (Al Asmari, 2013). Ansarimoghaddam and Tan (2014) define writing as a "highly complex

and demanding task" (p. 7), while (Shah et al., 2011) affirm that writers who can write fluently are usually able to grasp the grammatical rules. On the other hand, Ivanic (2004) defines writing as a set of social practices involving different patterns based on participation, gender preferences, network of support and collaboration. In addition, Ivanic elaborates that writing and reading are interconnected to each other.

It is undeniable that tertiary level students further develop their writing skills. According to Raoofi et al. (2017), such development is crucial as writing is pertinent to students' academic advancement. Cummings (1990) is also of the opinion that writing may result in positive outcomes for students. Cummings also believes that writing provides learners with a record of their products in which they (learners) can reflect, correct and monitor, unlike other language skills such as listening and speaking.

In becoming proficient writers, several aspects influence a student's composing ability, such as vocabulary. Allen et al. (2016) posit that a student can compose better writing texts when the student's vocabulary size is large and that skilled writers can write longer compositions containing fewer grammatical and spelling errors. It is also noted that skilled writers tend to utilise low-frequency lexical items. As a result, their essays are usually longer, containing elements suggestive of more refined lexical, syntactical, and rhetorical properties.

Features of Non-native English Writing

In their study, Eckstein and Ferris (2017) also found that non-native writers usually exemplified less complicated compositions, which are also shorter and less impactful than native writers. In addition, previous empirical research has also shown that the development of a student's non-native writing skills, such as vocabulary and grammar, is observed to be uneven (Aryadoust, 2016).

Vedder and Benigno (ibid) also indicated that non-native writers tend to over employ a trivial quantity of generic verbs constructed collocations such as *be*, *have*, and *take* in addition to an overuse of lexical items that amplify, augment or extend a meaning such as *completely*, *highly*, and *very*. Also observed in the writing of non-native writers is an overextension of verbs that are non-restricted such as *make* and *do*, which usually results in incorrect combinations such as *'to make a favour'*. Again, these errors are induced from the native language or caused by other second languages transfer. Such findings also concur with Yoon (2016), who observes that the overuse and misuse of general verbs are prevalent in non-native writers' writing and such over and misuse do not necessarily indicate appropriate usage of the language.

Another study reported that linguistic knowledge non-native writers' ability to fluently write in English is usually determined by their capability to portray correct linguistic knowledge (Schoonen et al., 2003). This notion is supported by Yoon (2017) who asserts that linguistic complexity

is usually associated as variables which are dependent of second language writing and that students' educational background may contribute to their linguistic complexity.

A small amount of empirical research notes that the utilisation of cohesive devices among non-native writers signifies an increase in proficiency. Studies have also shown that the number of cohesive devices used usually correlate to the quality of anyone essay (Chiang, 2003; Jafarpur, 1991; Liu & Braine, 2005; Yang & Sun, 2012). A recent study conducted by Crossley et al. (2016) on a group of non-native students from Michigan State University finds that a high number of cohesion indices correspond to the overall essay quality of non-native writers.

IIUM EPT Writing Descriptor

At the IIUM, the EPT uses its own in house developed descriptor to assess students' writing. The analytical-type descriptor was developed to denote 12 different proficiencies of students writing (called a band), which ranges from the lowest value of zero (0), denoting an absence of mastery in writing to the highest value of nine (9) denoting an extremely strong mastery of their writing ability. Another feature of the IIUM EPT writing descriptor is that students' essays are evaluated against four categories when raters attempt to determine the writing band. These categories are as follows: (1) the students' ability to respond to the given task, (2) the students' ability to produce texts that are cohesive and coherent, (3) the students' ability to utilise a suitable range of vocabulary in their writing task, and (4) the students' ability to utilise an appropriate range of grammatical structure accurately. Each of these categories is carefully defined at their respective levels. Students will be awarded the band where descriptions of all or most of the four stated categories above correspond to the students' writing. As the IIUM EPT writing descriptor is a confidential internal document, the authors cannot provide a more detailed description of the instrument in this article.

CEFR Illustrative Descriptor for General Linguistic Range

According to CEFR (Council of Europe, 2001), research on linguistic universals has yet to yield directly applicable to language learning, teaching, and assessment. As such, the CEFR asserts that it can make statements for the "General linguistic range" illustrative descriptor (p.110) only to provide classificatory tools for some parameters and categories that may be useful for describing linguistic content serves as a basis for reflection. Additionally, the illustrative descriptor's attempt to distinguish the different abilities of language is a widely used one that reflects the need to consider the complexity of the language being used rather than the errors that learners commit. A more recent version of the published CEFR manual (Council of Europe, 2020) operationalised the illustrative descriptor for 'General linguistic range' as follows: (1) learners' linguistic proficiency ranges from A1 to C2, (2) learners produce language from either rehearsed phrases to a very broad

range of language conventions in order to express, emphasise, or differentiate their ideas concisely, and to eliminate any form of ambiguity, and (3) learners limitations in producing comprehensible language can range from frequent breakdown/misunderstanding in non-routine situations to having no restrictions of what they want to produce. For this research, the authors believe that the 'General linguistic range' illustrative descriptor scale closely matches the linguistic repertoires illuminated in the in-house developed IIUM EPT writing descriptor.

MATERIALS AND METHODS

Sampling

This study was conducted to investigate the English language writing proficiency of graduating undergraduates at the International Islamic University Malaysia based on two different scales of proficiency; (i) the writing proficiency scale of the IIUM administered English Proficiency Test (EPT), and (ii) the 'General linguistic range' illustrative descriptor of the Common European Framework of Reference for Languages (CEFR). In addition to this, the research aimed to investigate and explore the correlation between the graduating students' EPT examination writing bands and the CEFR 'General linguistic range' illustrative descriptor.

The study participants included 460 IIUM final year undergraduates at all Faculties of the Gombak campus and were selected using a random stratified sampling method. The researcher believed that

using such a sampling method would lead to an outreach of students from different faculties, thus representing the University's undergraduate final year population. In addition, the participants were also students studying an academic English language course offered by the Centre for Languages and Pre-University Academic Development (CELPAD).

Instrument

The research instrument comprises a writing test containing a single prompt requiring participants to respond in an essay-based format, grounded on a set of arranged criteria. The test was constructed based on the structure and emulation of the IIUM EPT. The test also demanded basic demographics from the respondents, which included students' names, matric number and year of study, faculty and major. Although the respondents to this study remain anonymous, a field requiring respondents to provide their name was included so that the class instructors could request the essay questionnaires for classroom activities should instructors wish to do so, a win-win situation for both the researcher and class instructors. On the other hand, the question prompt required respondents to answer a simple question concerning the English language and employability, suitable for students in their final year of study, as they were more mature in exploring the topics concerning after campus life. The notion behind the usage of a questionnaire in the form of an essay question mainly lies in meeting the objectives of this research,

which was to measure the proficiency level of graduating IIUM students based on the EPT bands and the CEFR illustrative descriptor for 'General linguistic range'.

Data Analysis

Since the nature of this research was to investigate the IIUM students' English written proficiency, a writing test was administered during the data collection process, and because a writing test was conducted, the scripts needed to be assessed and scored. A few examiners then assessed the papers after a sit-in session was conducted in order to standardise test scores. After all papers were examined and scores recorded, the data (test scores) were entered into the Statistical Package for the Social Sciences (SPSS) software to generate findings to the three research questions. For research questions 1 and 2, a simple statistical calculation was performed. The intended outcome was expressed by a mean score ($\bar{X}$) whereas for research question 3, the data were analysed for correlation through Spearman's rho (ρ). The result was conveyed via a coefficient correlational value.

RESULTS AND DISCUSSION

Research Question One

The first research question of this study attempted to identify the IIUM students' performance in the EPT scale of writing.

Table 1 displays the score range of the IIUM graduating undergraduates. As can be seen, it is clear that the minimum score achieved was 4 (representing Band 4) whilst the highest was 8 (representing band 8). However, the mean score achieved by respondents in this study was μ=5.872 and can be rounded up to 6 (representing Band 6). The result also displayed a standard deviation of 0.4402 between the lowest and highest score (band).

Table 1

EPT score range

N	Min band	Max band	Mean band	Std. Deviation
460	4	8	5.872	0.4402

Table 2, on the other hand, exhibits the distribution of the EPT scores (bands) of the respondents for this study. It was discovered that 171 (37.2%) of the respondents failed to achieve Band 6. Upon careful examination, it was discovered that only two respondents (.4%) achieved Band 4, while 23 (4.9%) achieved Band 5. On a more positive note, 146 (31.3%) respondents, making up one-third of the overall sample population, managed to attain Band 5.5–a score deemed 'acceptable'.

Table 2 also states that 289 of the respondents have, without doubt, surpassed the minimum passing score making up a passing rate of 62.8%. Nevertheless,

it is worth noting that most of this population only managed to secure a Band 6, the minimum band for entry into the undergraduate courses. On a more serious note, the remaining 37.2% have not reached the minimum EPT requirement as stipulated by IIUM. However, these respondents are completing their studies at the university.

Table 2

EPT score distribution

Band	Frequency	Valid Percent	Cumulative Percent
4	2	0.4	0.4
5	23	5.0	5.4
5.5	146	31.7	37.2
6	221	48.0	85.2
6.5	56	12.2	97.4
7	11	2.4	99.8
8	1	0.2	100

Research Question Two

The second research question to this study strives to investigate the IIUM students' performance in English language writing, based on the CEFR 'General linguistic range' illustrative descriptor. Therefore, the writing test papers were also scored based on the selected CEFR illustrative descriptor, and the results can be seen in Table 3.

Table 3

CEFR score range

N	Min band (CEFR)	Max band (CEFR)	Mean band	Std. Deviation
460	4 (B1)	9 (C2)	6.722	0.8329

Based on Table 3, the score ranged from 4 to a maximum of 9, representing Level B1- to C2, respectively. A reading of $\bar{X}$ = 6.722 was achieved for the mean score. When rounded up, it can be said that IIUM students' English language writing proficiency score is 7, signifying a CEFR Level of B2+ (strong vantage). In comparison to the test results based on the EPT Bands, a higher deviation of standard (0.8329) was seen when the writing papers were examined using the CEFR illustrative descriptor.

Table 4 illustrates a detailed distribution of the respondents' CEFR scores. It should be noted that the scores were coded

numerically for statistical analysis using SPSS, and therefore, each numerical item is representative of a specific CEFR level of proficiency: 1 (A1), 2 (A2), 3 (A2+), 4 (B1), 5 (B1+), 6 (B2), 7 (B2+), 8 (C1), and 9 (C2). Although an allowance was made to code all the CEFR proficiency levels numerically, only the levels from 4 (B1)–9 (C2) was relevant after the analysis was made.

From Table 4, it can be understood that only one respondent's English writing level of proficiency (0.2%) was recorded as being B1- whereas 23 (5%) of the respondents were assessed as being Level B1+ users. When clustered together, the number of respondents whose English language writing proficiency can generally be categorised as Level B1 was 24, making up 5.2% (the minority) of the total population. On the other hand, 156 (33.9%) of the respondents' proficiency was evaluated as Level B2- users and another 213 (46.3%) respondents' proficiency level was categorised as B2+. It brings about a total of 369 (80.2%) respondents whose English language writing proficiency can be universally categorised

as Level B2, making up the majority of the sample population. A total of 57 (12.4%) respondents' proficiency was rated as Level C1 users, whereas the remaining 10 (2.2%) were evaluated as having a proficiency level of C2, the highest level based on the CEFR illustrative descriptor.

To sum, it is safe to conclude that a large proportion of IIUM students' English language writing proficiency based on the CEFR 'General linguistic range' is Level B2+.

Research Question Three

This study's third and final research question is intended to investigate whether a relationship exists between the EPT bands and the CEFR illustrative descriptor for 'General linguistic range'.

For this purpose, a Spearman's correlation was run to assess the relationship between the EPT bands and the CEFR scale of proficiency using a substantial sample size of 460 undergraduate respondents who were in their graduating semester at the IIUM. From the statistical analysis, results in Table 5 indicate evidence to suggest good agreement and a strong positive correlation between the EPT Bands and the CEFR illustrative descriptor. In other words, there is a strong relationship between students' writing performance in the EPT and CEFR ($r_s = .874$). The results also illustrate that the relationship is statistically significant at $p = .000$.

Table 4

CEFR score distribution

Level	Frequency	Valid Percent	Cumulative Percent
B1	1	0.4	0.2
B1+	23	5.0	5.2
B2	156	33.9	39.1
B2+	213	46.3	85.4
C1	57	12.4	97.8
C2	10	2.2	100

Table 5

Correlation between EPT and CEFR

		EPT	CEFR
EPT	Correlation Coefficient	1.000	0.874**
	Sig. (2-tailed)	.	0.000
	N	460	460
CEFR	Correlation Coefficient	0.874**	1.000
	Sig. (2-tailed)	0.000	.
	N	460	460

**Correlation is significant at the 0.01 level (2-tailed)

IIUM Students Performance Based on EPT

Results indicate that most IIUM graduating undergraduates' English language writing proficiency based on EPT stands at Band 6 ($\bar{X}$ =5.872). According to the EPT rubric, this denotes that undergraduates at IIUM have an effective command of the language, although some imprecisions, incongruous usages and misapprehensions may be seen. Being a Band 6 also signifies that a student can comprehend a fairly sophisticated English language level, usually in a situation recognisable to them.

IIUM Students Performance Based on CEFR

The second research question to this study attempts to identify the IIUM English language writing proficiency based on the scale of the CEFR 'General linguistic range'. Grounded on the written examination results, IIUM students' level of proficiency, when measured against the CEFR 'General linguistic range', stands at Level B2+, characterised by the Council of Europe as "strong vantage" or "independent" users. However, based on the global proficiency scale, B2+ users still fall under the B2 (independent vantage users) portfolio. Learners at this level are generally described as being able to utilise a limited number of cohesive devices; link sentences together smoothly into clear, connected discourse; use a variety of linking words efficiently to mark the relationships between ideas; develop an argument systematically with appropriate highlighting of significant points as well as relevant supporting details (Council of Europe, 2020).

Correlation of EPT and CEFR Scores

The present study shows that the EPT writing bands have a strong positive relationship with the writing proficiency scale of the CEFR as the value of the correlation was found to be r_s = .874. As previously mentioned, the EPT writing test bands measure one writing proficiency on a scale from 0 (no attempt)–9 (native fluency), while the CEFR's descriptor for 'General linguistic range' further highlights

what students can execute whilst performing writing tasks and the proficiency is measured on a scale from A1 (beginner–breakthrough/basic level) to C2 (proficient user–mastery/proficiency).

As the strength of a correlation reflects how consistently values for each factor change, it can be deduced that the higher the score for the EPT, the CEFR levels would also increase. This study also discovered that the mean level of English writing proficiency of IIUM's graduating undergraduates stood at 6, while based on the CEFR writing proficiency scale, the students were at B2+. As such is the case, it is also safe to construe that a Band 6 (EPT) correlates to Level B2+ (CEFR).

As illuminated above, the discussion of the study's third research question corroborates that although the IIUM EPT writing descriptor is not developed based on the conventions of CEFR and its can-do statements, it does to a certain extent reflect students writing ability to communicate effectively despite IIUM's focus on preparing students to negotiate successful learning at the higher education level. About CEFR, its emphasis on communicative competence is exemplified in the IIUM students as the EPT, as shown in this research, corresponds to the principles of CEFR.

CONCLUSION

The study's findings have revealed that the majority (48%) of IIUM graduating undergraduates' level of English writing proficiency when measured using the IIUM English Proficiency Test's Band score stands at Band 6. The study also reveals that 37.2% scored lower where 0.4%, 4.9% and 31.3% scored Band 4, Band 5, and Band 5.5, respectively. However, the research results also indicated that 14.8% of the respondents managed to surpass the minimum language requirement for entry into the faculties as 12.2% scored Band 6.5, 2.4% scored Band 7, while the remaining 0.2% succeeded in scoring Band 8. As the study was conducted to investigate the English language writing proficiency of IIUM undergraduates, this can be interpreted to mean that most IIUM undergraduate students generally possess abilities to use the language effectively despite some inappropriacy, inaccuracy, and misunderstandings. In addition, these students can also use and understand fairly complex language, particularly in situations familiar to them. Such elucidations were made valid through the sample scripts scrutinised to identify what construes a Band 6 student at the IIUM.

In addition, results to this study have indicated that the English language writing proficiency level of the majority of IIUM undergraduate students, when set against the CEFR's 'General linguistic range' illustrative descriptor, stood at Level B2+. According to the CEFR framework, this level indicates that IIUM students are "strong vantage" language users while at the same time, independent (Council of Europe, 2020), also suggesting a transition into different and worthwhile zones of language development (McCarthy, 2013). Although placed at proficiency level B2+, the global descriptor specifies only the main

proficiency levels without consideration of the "strong" categorisations. With this in mind, it can be further deduced that B2 would best describe the proficiency level of IIUM undergraduates. The Council of Europe (2001) describes independent vantage language users as adept at producing vivid and comprehensive text on many subjects. It can explicate a viewpoint on an interesting issue by furnishing the advantages and disadvantages of various options. Also shown through the study results is that 5.2% of the respondents were assessed as universally being B1 users. The term 'universally' is used as the mentioned figure (5.2%) encompasses both B1 (0.2%) and B1+ (5.0%) users. Conversely, 14.6% of the respondents were rated to be C1 (12.4%) and C2 (2.2%), users and these graduating undergraduates' score is over the minimum graduation language requirement set forth in the English language roadmap.

Although results were encouraging, a small percentage of IIUM graduating undergraduates have not been able to maintain or improve their English language writing proficiency. In response to this revelation, research reveals a presence of variability and instability in proficiency, and because of this, it is challenging to define a person's level of proficiency at a specific moment in time (Lowie, 2012).

The study also reveals a statistically significant positive relationship between the scores of the EPT and CEFR as a reading of $r_s = .874$, $p = 0.000$ is obtained. The strength of the relationship ($r_s = 0.874$) is expected since both tests use the same construct, i.e. writing proficiency in English. Such disclosure signifies a similarity between the two measures of proficiency because the value designated in one construct of either EPT or CEFR will increase with the measurement of the other construct used in tandem.

RECOMMENDATIONS FOR FUTURE RESEARCH

Future researchers who are interested in conducting mapping studies of the writing proficiency levels of graduating IIUM undergraduates may be interested in aligning the EPT Band scores of the data interpretation writing task to the CEFR 'General linguistic range' illustrative descriptor. Such a research initiative would enable the IIUM to have a more detailed and comprehensive outlook of its students' linguistic profile in writing. In addition, outside the scope of IIUM EPT, recommendations include mapping an institution's own in-house developed language assessment to the CEFR so that a clearer overview of the extent to which it corresponds to CEFR, especially with the introduction of CEFR in the Malaysian educational system since 2015. Therefore, other researchers and institutions wanting to contextualise their learners along the CEFR scale could also benefit from this research.

ACKNOWLEDGEMENT

REFERENCES

Al Asmari, A. (2013). Investigation of writing strategies, writing apprehension, and writing achievement among Saudi EFL-major students. *International Education Studies*, *6*(11), 130-143. https://doi.org/10.5539/ies.v6n11p130

Ali, N. L. (2013). A changing paradigm in language planning: English-medium instruction policy at the tertiary level in Malaysia. *Current Issues in Language Planning*, *14*(1), 73-92. https://doi.org/10.1080/14664208.2013.775543

Allen, L. K., Snow, E. L., & Mcnamara, D. S. (2016). The narrative waltz: The role of flexibility in writing proficiency. *Journal of Educational Psychology*, *108*(7), 911-924. https://doi.org/10.1037/edu0000109

Ansarimoghaddam, S., & Tan, B. H. (2014). Undergraduates' experiences and attitudes of writing in L1 and English. *GEMA Online Journal of Language Studies*, *14*(1), 7-28. https://doi.org/10.17576/GEMA-2014-1401-02

Aryadoust, V. (2016). Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology*, *36*(10), 1742-1770. https://doi.org/10.1080/01443410.2014.950946

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, *31*(4), 471-484. https://doi.org/10.1016/j.system.2003.02.002

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages*. Cambridge University Press.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1-16. https://doi.org/10.1016/j.jslw.2016.01.003

Cummings, V. (1990). *Speech and writing: An analysis of expository texts compared by native and non-native speakers of English at The City University of New York* (Doctoral dissertation). Columbia University, New York.

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: Comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, *32*, 43-56. https://doi.org/10.1016/j.asw.2016.12.005

Don. Z. M. (Ed.). (2015). *English language education reform in Malaysia: The roadmap 2015-2025*. Ministry of Education Malaysia.

Eckstein, G., & Ferris, D.(2017). Comparing L1 and L2 texts and writers in first-year composition. *TESOL Quarterly, 52*(1), 137-162. https://doi.org/10.1002/tesq.376

Finch, A. E. (2009). Europass and the CEFR: Implications for language teaching in Korea. *English Language and Literature Teaching*, *15*(2), 71-92.

Ganapathy, M. (2015). *Boosting Malaysian graduates' employability via the new MUET policy*. National Higher Education Research Institute (IPPTN).

Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, *14*(August), 1-30. https://doi.org/10.1075/eurosla.14.01gyl

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, *12*(4), 333-362. https://doi.org/10.1080/15434303.2015.1092545

Ivanic, R. (2004). Discourses of writing and learning to write. *Language and Education, 18*(3), 220-245. https://doi.org/10.1080/09500780408666877

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System*, *19*(4), 459-465. https://doi.org/10.1016/0346-251X(91)90026-L

Lee, Z. E. (2015). *Colloquial Malaysian English (CMalE): A problem or a cool phenomenon?* (Published Masters Thesis). Universitat Jaume I, Spain.

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, *33*(4), 623-636. https://doi.org/10.1016/j.system.2005.02.002

Llurda, E. (2013). English-medium instruction at universities: Global challenges. *ELT Journal*, *67*(4), 497-500. https://doi.org/10.1093/elt/cct045

Lowie, W. (2012). The CEFR and the dynamics of second language learning: trends and challenges. *CercleS*, *2*(1), 17-34. https://doi.org/10.1515/cercles-2012-0002

McCarthy, M. (2013). Putting the CEFR to good use: Designing grammars based on learner-corpus evidence. *Language Teaching*, *June 2013*, 1-17. https://doi.org/10.1017/S0261444813000189

Ministry of Finance. (2014). *Budget 2015: People's economy* (Issue October 2014). Ministry of Finance.

Mohamed, M. (2008). Globalization and its impact on the medium of instruction in higher education in Malaysia. *International Education Studies*, *1*(1), 89-94. https://doi.org/10.5539/ies.v1n1p89

Ming, T. S., & Alias, A. (2007). Investigating readiness for autonomy: A comparison of Malaysian ESL undergraduates of three public universities. *Reflections on English Language Teaching*, *6*(1), 1-18.

Noijons, J., & Kuijper, H. (2010). Mapping the Dutch foreign language state examinations onto the CEFR. *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual, 33*, 247.

Noor, S. N. F. M., & Kadir, Z. A. (2007, June 14-16). Students' learning preferences of English for academic purposes -A KUiTTHO affair. In *The Second Biennial International Conference on Teaching and Learning of English in Asia : Exploring New Frontiers (TELiA2)* (pp. 1-11). UUM, Malaysia.

Nunan, A. (2014). Standardising assessment to meet student needs in foreign language modules in a university context: Is standardisation possible? *Language Learning in Higher Education*, *4*(1), 117-135. https://doi.org/10.1515/cercles-2014-0008

Raoofi, S., Binandeh, M., & Rahmani, S. (2017). An Investigation into writing strategies and writing proficiency of university students. *Journal of Language Teaching and Research*, *8*(1), 191-198. https://doi.org/10.17507/jltr.0801.24

Shah, P. M., Mahmud, W. H., Din, R., Yusof, A., & Pardi, K. M. (2011). Self-efficacy in the writing of Malaysian ESL learners. *World Applied Sciences Journal*, *15*(October), 8-11.

Singh, M. K. M. (2016). An emic perspective on academic writing difficulties among international graduate students in Malaysia. *GEMA Online® Journal of Language Studies*, *16*(3), 83-96. https://doi.org/10.17576/gema-2016-1603-06

Samuel, R., & Bakar, Z. (2008). The effectiveness of ' VELT ' in promoting English language communication skills : A case study in Malaysia. *International Journal of Education and Development Using Information and Communication Technology (IJEDICT)*, *4*(3), 109-126.

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed

of processing, and metacognitive knowledge. *Language Learning*, *53*(1), 165-202. https://doi.org/10.1111/1467-9922.00213

Shakir, R. (2009). Soft skills at the Malaysian institutes of higher learning. *Asia Pacific Education Review*, *10*(3), 309-315. https://doi.org/10.1007/s12564-009-9038-8

Tajuddin, A. J. A. (2015). *A Malaysian professional communication skills in English framework for English for occupational purposes courses* [Doctoral dissertation, University of Nottingham]. Nottingham eTheses. http://eprints.nottingham.ac.uk/id/eprint/28747

Tan, K. E., & Miller, J. (2007). Writing in English in Malaysian high schools: The discourse of examinations. *Language and Education*, *21*(2), 124-140. https://doi.org/10.2167/le663.0

Taylor, L., & Jones, N. (2006). Cambridge ESOL exams and the Common European Framework of Reference (CEFR). *Research Notes*, *24*, 2-24.

Weigle, S.C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wu, J. R., & Wu, R. Y. (2007). Using the CEFR in Taiwan: The perspective of a local examination board. *The Language Training and Testing Center Annual Report, 56*, 1-20.

Wudthayagorn, J. (2018). Mapping the CU-TEP to the Common European Framework of Reference (CEFR). *LEARN Journal: Language Education and Acquisition Research Network*, *11*(2), 163-180.

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education*, *23*(1), 31-48. https://doi.org/10.1016/j.linged.2011.09.004

Yoon, H. J. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing*, *34*, 42-57. https://doi.org/10.1016/j.jslw.2016.11.001

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, *66*, 130-141. https://doi.org/10.1016/j.system.2017.03.007

SOCIAL SCIENCES & HUMANITIES

Enhancing Learner Autonomy through Extensive Reading: The Case of Book Reports

Truong Thi Thanh Canh

English Faculty, Foreign Trade University Ho Chi Minh City, 72300, Vietnam

ABSTRACT

This paper investigates the autonomous learning strategies employed by students while participating in extensive reading (ER) and their perception of this method. Semi-structured interviews and autonomy strategies questionnaires are used as research instruments. The 25 participants are freshmen majoring in English. The findings illustrate some favored autonomous strategies that students used, such as summarizing, making inferences, taking notes, and using imagination. It is noticeable that using translation is an uncommon strategy chosen by these participants, which is somewhat unexpected in the Vietnamese context where students prefer it the most. For metacognitive and effective strategies, a sense of responsibility and monitoring received the most attention from the students. This result shows a positive attitude among the students in claiming that their autonomous learning can be enhanced via completing book reports. The findings from the current study provide insights into the implementation of book reports in extensive reading to enhance learner autonomy.

Keywords: Extensive reading, learner autonomy, reading strategies

INTRODUCTION

According to Mikulecky (2008), reading comprehension is one of the most crucial factors in English language learning because it functions as the backbone of instruction in all aspects of language learning. As Rodrigo et al. (2014) claim, good reading ability develops reading skills as daily habits. Nonetheless, recently, there has been a

concern that students lack love for habitual reading, even at the tertiary level.

In Vietnam, Mr. Nguyen Manh Hung, Minister of Information and Communications (Vietnamnet.vn), emphasized his worries about its reading rate because it has barely risen during the last 30 years. Specifically, in a survey conducted by Vietnam National University Ho Chi Minh city in 2016, only 30 percent read regularly, while over a quarter had no idea of reading. At the tertiary level, research conducted by Nguyen (2017) at People's Police University found that most non-major English students do not read much. This fact has created an alarming situation for educationists because the negative consequences can be foreseen. Some can be identified as poor academic performance, examination malpractice, mass failure, anti-social behaviors, poor understanding, fear and anxiety towards examinations and tests, poor execution of research projects and assignments, fall in the standards of education, among others (Issax & Kingley, 2020)

Teachers should motivate students to learn independently and take responsibility for learning to solve this problem. It means that effective variables, particularly motivation, are crucial to students' learning reading skill. For example, interactive reading activities may increase motivation to read more (Day & Bamford, 2002). In the EFL context, it is suggested that one way to improve reading skills in English is to read extensively (Nuttall, 1996) or what can be known as extensive reading (ER). To reinforce the belief in ER, Yamashita (2013) argues that ER has positive impacts on learners' attitudes or even fosters a love for reading. However, what is left uncertain is how students respond to ER implementation and what their attitudes are. Thus, this paper was conducted to gain an insight into the issue of how ER can be applied in an EFL context to motivate students' autonomous learning. The study was guided by the following two research questions to achieve that purpose:

Question 1: What are autonomous learning strategies used by students while participating in extensive reading activities? *Question 2*: What is students' attitude toward the extensive reading implementation?

LITERATURE REVIEW

Learner Autonomy and Autonomous Learning Strategies

Learner autonomy is a concept coined by Holec (1981) as a person's ability to take charge of their learning. The term was reported as reflecting critically, making decisions, and acting independently by Little (1991) and Sinclair (2000). It is known as a cognitive approach to motivation that focuses on "the individuals' thoughts and beliefs" (and recently also emotions) transferred into actions. Somehow, these processes are leveled up to another state where learners can experience active monitoring and consequent regulations of the cognitive process to active cognitive goals, which is considered metacognitive. Thus, when learners engage in autonomous learning, they can experience moving

from cognitive to metacognitive states. Literature has shown that those who can use cognitive learning strategies can succeed, differentiating them from less successful ones. Some sub taxonomies of cognitive strategies have been listed by Oxford (1990) as analysis, note-taking, summarizing, outlining, and reorganizing information. O'Malley and Chamot (1990) state that these strategies are specific measures or steps learners take to fulfill learning tasks.

For example, summarizing skills can help improve comprehension of texts and increase recall (Kinch & Van Dijk, 1978). Apart from summarizing, other cognitive skills can be named making inferences, making decisions, translating, applying grammar rules, taking notes, guessing meaning from texts, and using imagination (O'Malley & Chanmot, 1990). At a higher level, metacognitive strategies refer to the ability of learners to analyze items by themselves. Metacognition is cognition about cognition or thinking about thinking. It relates to active monitoring and consequent regulations and the orchestration of cognitive processes to achieve cognitive goals. O'Malley and Chamot (1990) consider metacognitive strategies as skills relating to planning, monitoring, or evaluating the success of a learning activity, which means evaluating the whole learning process.

The literature in reading comprehension reveals that readers with effective cognitive and metacognitive strategies have a good awareness of how to approach reading and monitor their learning, which boosts their learning autonomy. In this paper, the author considers the two strategies mentioned above in one particular reading activity: extensive reading to enhance students' learning autonomy. Furthermore, students imbuing these strategies also articulate their attitudes towards personal responsibility and learning capacity motivating their active participation in the learning process and encouraging them to learn responsively and independently.

Extensive Reading

Various researchers have attempted to provide a comprehensive definition for extensive reading (ER). Davis (1995) considers ER as a way to give learners time, encourage them, let them read as many materials as possible with pleasure, within their levels, and without washback effects. Grabe and Stoller (2002) believe that ER means learners read large quantities of material within their linguistic competence. Brown (2012) explains that extensive reading refers to the reading of large amounts of material, the level of which is convenient for the reader, and more importantly, which they choose themselves. According to Maley (2009), extensive reading is understood as a method that motivates learners to read for their pleasure and information regularly, in a vast number of materials and a wide range of topics with their own choices of books and at a fast speed. However, in this paper, the author will tailor the idea of fast speed into learners' suitable speed because she wants to encourage students' enjoyment of reading. Thus, extensive reading is defined as choosing their topic and genre

and reading at their pace in this study. All in all, based on the above researchers, some principles of ER given by Grabe and Stoller (2002) are applied in this study such as: (1) how students engage in reading activities; (2) what fluent reading skill is; (3) how reading is performed as a cognitive process; and (4) how the learners can draw meaning from their reading activities and (5) how their reading proficiency can be specified. Consequently, if learners participate in ER activity, they will gradually become more autonomous in their learning.

Extensive Reading on Learner Autonomy

Brown (2012) claims it is apparent that learner autonomy has strong links with extensive reading. Research has also shown a positive correlation between extensive reading and learner autonomy, which has been beneficial to students in language learning. Specifically, Dickinson (1995) argues that successful individualized reading experiences foster learner autonomy, learning success, and enhanced motivation. According to Bell (1998), the idea of giving students autonomy to choose the genre of material to read, as well as the pace at which to read, is in itself motivational because it addresses the needs and interests of individual learners. In addition to this, ER acknowledges and supports the fact that reading is an individual undertaking that allows individuals to learn at their pace, depending on their level of proficiency (Nation, 1997). From another perspective, this implies that ER offers flexibility to

learners and teachers in teaching and learning and would match the teacher-researchers teaching goal.

RESEARCH METHODOLOGY

Participants

The participants were twenty-five university freshmen students who studied reading two-course, approximate to a B1 in the Common European Framework of Reference for Languages (CEFR) level. The course for the spring semester of the 2018–2019 academic year lasted three months with 9 hours per week. Their English level is pre-intermediate.

Instruments

Semi-structured interviews for ten students and autonomy strategies questionnaire are used as research instruments for the study. An autonomy training strategies questionnaire was developed to measure the students' use of autonomous learning strategies. This questionnaire was adapted in part from Oxford (1990), O'Malley & Chamot (1990), and Channuan (2012). The questionnaire was divided into three main categories or parts with 29 statements in the form of a 5-point Likert scale as follows:

Part 1: *Cognitive Strategies*. Using background knowledge, summarizing, predicting, making inferences/ guessing, using resources, using imagination, taking notes, using mechanical means to store information, transferring, using keywords to find information, using translation and self-talk.

Part 2: *Metacognitive Strategies*. Using items such as planning, monitoring, and self-evaluating.

Part 3: *Students' Attitude toward Teacher's Roles*. It is emphasized that the questionnaire was translated into Vietnamese to avoid ambiguity before being administered to students. It was explained to them that they would remain anonymous and the collected results were for research purposes only. Then, a semi-structured interview was conducted at the end of the course to get a deeper understanding from students' feedback. The learner autonomy questions for the interview, which were adapted from Channuan (2012), include:

Question 1: Which strategies do you often/ rarely use while reading outside the class?
Question 2: When a teacher assigns you to write a book report:

- What do you normally do before you start reading extensively and writing a book report?
- While reading and writing a book report, have you encountered any problems, and how did you solve these problems?
- After you complete your book report, do you make other self-assessment?

Question 3: After finishing the course, can you take responsibility for your reading? How?

Regarding reliability, the conversations between the author and interviewees were conducted in Vietnamese, then transcribed into English, and that data were coded by numbering the students (e.g., Student 1 or Ss1).

Procedure

As for the reading skills, students had their textbook that followed the curriculum given by the university. However, apart from that textbook for the present study, they were also introduced to extensive reading, the goal and requirements they need to achieve and fulfill. Thus, the procedure was announced as the steps in Table 1.

Table 1

Timeline of ER implementation

Timeline	Content
Week 1	Teacher introduces extensive reading for students Teacher informs students with timeline and requirements that they need to fulfill
Week 2	Students hand in their plan for their reading, including the title of their chosen books

Table 1 *(Continued)*

Timeline	Content
Week 3	Students follow their plan in reading and ask the teacher for help if they need
Week 8	Students follow their plan in reading and ask the teacher for help if they need
Week 9	Students write and revise the book reports
Week 10	Students submit book reports
Week 11	Students answer the questions given by the teacher relating some points written their paper
Week 12	Students exchange their paper with one partner in the class and give thoughts on that paper orally on the last day of the course

This timeline was announced and explained in the first week of the course. Then it was sent to students individually to keep track of their process and follow the timeline. In addition to this class instruction, the teacher prepared a word file to denote other guidelines in detail, such as difficulty level for the reading materials, desired length of a book report, numbers of parts, or content. All of this preparation aimed to help students understand what they needed to do with the given task.

RESULTS

Use of Language Learning Strategies

Use of Cognitive Strategies. Table 2 shows the means obtained for each item in the questionnaire. The results were used to determine the strategies employed by the students and the frequency of the strategy used. From the results, it was found that students generally use all language learning strategies with high frequency.

Specifically, some cognitive strategies used most often were guessing the meaning of unknown words from the context (M=4.2), using a dictionary to find the meaning of the really important words (M=4.2), and using keywords to find information in the text (M=4.1). The other skills such as summarizing, making a prediction, using imagination, and background knowledge were preferred by students with mean scores ranging from 3.7 to 3.9. Surprisingly, that the translating strategy was used at a slightly lower level of frequency with M = 2.5.

Use of Metacognitive Strategies. In terms of metacognitive strategies, which involve planning, monitoring, and evaluating, the students regulated metacognition at a high level. For example, as revealed in Table 3, learners seem to be aware of their learning process when claiming that they know their weaknesses in reading and try to improve them (M = 3.7), or always ask themselves whether they understand what they are

reading (M = 4.2). In contrast, these students perceived that they had less responsibility towards making a reading plan (M = 2.8) and following their reading schedule (M = 2.5) or checking if they could finish the reading in time (M= 2.7).

Table 2

Use of cognitive strategies

Statements	Mean Score
Before I read, I thought about what I already knew about the topic, which helped me understand the story better.	3.9
I try to summarize (in my head or writing) important information that I read.	3.9
I usually predict a story while I am reading along.	3.6
I usually try to guess the meaning of unknown words from the context.	4.2
I use a dictionary to find the meaning of the really important words that I do not know.	4.2
If there are pictures in the text, I usually imagine what the text would be about.	3.7
I take notes while reading.	3.0
I write down or make lists of new words or phrases I see in the reading passages to be learned several times.	3.1
I try to understand the vocabulary from its prefix or suffix.	3.8
I periodically focus on specific information to achieve my reading objectives.	3.6
I usually ask myself questions about the texts and check if they make sense to ensure reading comprehension.	3.5
When I need to find some information in a text, I usually look for keywords.	4.1
I usually review the strategies I use while reading.	3.7
I translate from English into Vietnamese when I read the texts.	2.5
When I encounter a long, difficult text, I tell myself that I can read it, and I will try my best by using all the strategies that I have practiced.	3.6

Table 3

Use of metacognitive strategies

Statements	Mean Score
I start reading by browsing throughout the book and focusing on the content that interests me.	3.6
Before reading, I set my reading objectives in advance and read with those objectives in mind.	3.0
I have set a reading schedule, and I could follow it.	2.5
I finish reading faster after I make a reading plan.	2.8
While reading, I usually ask myself whether I understand what I am reading.	4.2
I know my weakness in reading and try to improve them by myself.	3.7
I always keep track of my reading progress.	3.4
I check my understanding by doing the follow-up exercise or summarizing the story.	3.5
After reading, I decided whether the reading strategies I used helped me understand the passages better, and I think of other strategies that I could have helped.	3.2
After reading, I check whether I accomplished my reading objectives, such as finishing the reading in time.	2.7

In the qualitative data, students provided more details to explain their choices in the survey questionnaire further. Most of the reading strategies used in extensive reading favored cognitive strategies such as guessing and predicting unknown words based on given clues or images. Some highlighted comments are as follows:

*When I read the book **Silence of the lambs**, there were many difficult words, so I had to use a dictionary to check up meanings, and sometimes I guessed with the context of the story.*

(Ss 5)

Or

*My chosen book for the book report is **The fault in our stars**. I read it because the content interests me. But a lot of words I do not understand, so I had to guess or look at the prefix or even checkup dictionary.*

(Ss 8)

In short, the quantitative and qualitative data show that while many participants prefer to employ cognitive strategies in their extensive reading, especially with books that have complicated content, and those who used metacognitive reading strategies were limited.

The Implementation of ER in Enhancing Students' Autonomous Learning and their Attitude

According to Thanasolous (2000) and Sinclair (2000), the learners' capacity to

control their learning positively impacts on their autonomy. In fact, from the semi-structured interview data, it can be assured that ER can help foster learner autonomy. In this book report project, ER allows learners to choose books they are interested in, no matter the genres. In addition, they could enjoy their own reading time, favorite place, or the manner or speed they preferred. To measure learner autonomy level, the author emphasizes two components*: (1) students' attitude towards their independence in learning; and (2) students' attitudes toward the teacher's role.* Firstly, in terms of independence in learning, most participants agree that ER helps them be more independent in their learning. Therefore, most participants claimed that a teacher should be a counselor instead of a controller who takes overall responsibility for students' learning (M = 4.0). A lot of language learning can be done without a teacher (M=3.3). Furthermore, they did not believe that the best way to learn a language is by teachers' explanation (M=2.7) or a teacher should choose materials for language classes (M=2.7). Thus, it is obvious that these participants are confident with their independence in managing the learning process.

In terms of students' attitudes towards ER, the transcribed information from the interview revealed that the majority of students stated that they felt motivated and excited with their book reports. The first reason is their freedom in choosing reading materials. As Ss 4 claimed:

I felt very excited and relaxed because I could choose my favorite type of book to read. I love to read non-fiction novels which are not allowed in the school textbook.

(Ss 4–Interview section)

Some others pointed out that they love the ER activity because, they had the chance to write their comments on what they had read in the critical part of their book reports. Sometimes, it is not easy to reflect on another person's writing. However, once they overcame that feeling, students felt more confident. From the researcher's perspective, the critical part is the most challenging in the book report because it requires a high level of metacognition.

In a nutshell, the participants show a positive attitude toward the implementation of ER with book reports. Nevertheless, they believed that this learning approach fosters their motivation to learn and helps them become autonomous learners.

DISCUSSION

The results reveal that students use most of the autonomous language learning strategies at a high-frequency level. Furthermore, cognitive strategies appear to be used most frequently by these students, which is similar to the results of the study carried out by Channuan (2012), Shin and Crandall (2014), and Nguyen (2018). These authors believe that cognitive strategies such as prediction and visualizing with given images effectively promote students' reading comprehension.

Besides, it is noteworthy that the finding of this present paper that the usage of the translating technique is the least used by the participants surprises the author because it opposes to the common belief that Vietnamese students seem to favor this reading strategy. For decades, Vietnamese learners have been used to learning in an exam-oriented environment. Consequently, they need to equip themselves with good grammar-based and translating methods in reading to be compatible with norm-referenced examinations (Le, 1999). However, up to now, the situation with the national examination system still witnesses no significant change. Therefore, the finding that fewer translating techniques are used in reading as claimed by the participants is a positive signal indicating the gradual reform in their perception of learning a foreign language.

Apart from the abovementioned issue, consciously or at least partially so, promoting independent learning is crucial for both learners and teachers in foreign language teaching. Therefore, with respect to reading skills, it is recommended that EFL students be exposed to extensive reading activity as an effective approach in enhancing their reading ability and their learning autonomy. Consequently, ER should be integrated into the training curriculum so that both teachers and students can navigate the potential benefits of implementing this activity. Nonetheless, some issues need addressing, such as the role of the teacher in manifesting the activity and interacting with students. Firstly, students should be responsible for

their reading processes, such as choosing materials and setting up a reading plan or reading pace. Secondly, teachers should be a counselor in helping them apply ER reading strategies to select suitable reading materials and gradually create their reading habits in the long term. Benson (2001) confirms that the teachers should not leave the learners to learn autonomy alone. Instead, they should actively encourage and provide the necessary support for the learners to enable them to take control of their learning. As a result, learners are more engaged in their learning process, and become more autonomous learners.

CONCLUSION

This study aimed to investigate autonomous learning strategies used by students during extensive reading and figure out their attitude towards the implementation of ER to enhance their reading autonomy. The findings support the belief that ER has positive effects on fostering learner autonomy. It allows learners to self-control their learning process and encourages them to be active learners. Therefore, it is understandable that students have positive comments on the implementation of ER. As a result, this study contributes to the literature on promoting the application of ER in reading training courses so that learners can be familiar with learner autonomy and the use of necessary learning strategies. Once students can control their learning process, they are ready to become autonomous learners as a part of their

lifelong learning. Therefore, it is strongly believed that extensive reading (ER) can enhance students' learning autonomy in English reading classes.

Due to the limited number of participating students, this study is small-scale and preliminary. Therefore, it cannot be expected to provide conclusive evidence regarding how students view ER implementation in general nor their favored autonomous learning strategies in English reading skills. Nonetheless, the results obtained may reflect the realities of the wider educational context beyond the local setting. The study's findings may also provide useful references to English teachers or educational reformers in other language teaching communities, where similar challenges exist in the implementation of ER in the English language.

ACKNOWLEDGEMENT

REFERENCES

Bell, T. (1998). Extensive reading: Why? And how? *The Internet TESL Journal, 4*(12), 1-6.

Benson, P. (2001). *Teaching and researching autonomy in language learning*. Pearson Education Ltd.

Brown, D. (2012). Online support systems for extensive reading: Managing the tension between autonomy and institutional education. *The Language Teacher, 36*(2), 11-16. https://doi.org/10.37546/jalttlt36.2-2

Channuan, P. (2012). *The effects of extensive reading using a learner autonomy training on reading ability and reader autonomy of Thai University students* [Doctoral dissertation, Chulalongkorn University]. Chulalongkorn University Intellectual Repository. http://cuir.car.chula.ac.th/handle/123456789/30617

Davis, C. (1995). Extensive reading: An expensive extravagance? *English Language Teaching Journal, 49*(4), 329-336. https://doi.org/10.1093/elt/49.4.329

Day, R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language, 14*(2), 136-141.

Dickinson, L. (1995). Autonomy and motivation a literature review. *System, 23*(2), 165-174. https://doi.org/10.1016/0346-251X(95)00005-5

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Pearson Education.

Holec, H. (1981). *Autonomy in foreign language learning*. Pergamon Press.

Isaac, E. A., & Kingsley, O. U. (2020). Effect of social media addiction on reading culture: A study of Nigerian students. *Library Philosophy and Practice (e-journal)*. 4170. https://digitalcommons.unl.edu/libphilprac/4170/

Kintch, W., & Van Dijk, T. A. (1978). Toward a model of discourse comprehension and production. *Psychological Review, 85*, 363-394. https://doi.org/10.1037/0033-295x.85.5.363

Le, V. C. (1999, October 13-15). Language and Vietnamese pedagogical contexts. In *Language and development: Partnership and interaction. Proceedings of the fourth international conference on language and development.* (pp. 13-15). Asian Institute of Technology.

Little, D. G. (1991). *Learner Autonomy 1: Definitions, issues and problems*. Authentik Language Learning Resources

Maley, A. (2009). Extensive reading: why it is good for our students and for us. *Teaching English*. https://www.teachingenglish.org.uk/article/extensive-reading-why-it-good-our-students-us

Mikulecky, B. S. (2008). *Teaching reading in a second language*. Pearson Education, Inc.

Nation, P. (1997). The language learning benefits of extensive reading. *The Language Teacher, 21*(5), 13-16.

Nguyen, N. A. (2017). *Using extensive reading to develop reading ability and motivation*. http://www.vnseameo.org/TESOLConference2015/Materials/Fullpaper/

Nguyen, T. (2018). Reading Strategies used by Vietnamese EFL and ESL University students. *VNU Journal of Foreign Studies*, *34*(2), 111-124. https://doi.org/10.25073/2525-2445/vnufs.4250

Nuttall, C. (1996). *Teaching reading skills in a foreign language* (2nd ed.). Heinemann.

O'Malley, M. J., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge University Press.

Oxford, R. (1990). *Language learning strategies: What every teacher should know*. Heinle & Heinle Publication

Rodrigo, V., Greenberg, D., & Segal D. (2014). Changes in reading habits by low literate adults through extensive reading. *Reading in Foreign Language, 26*(1), 73-91.

Shin, J. K., & Crandall, J. A. (2014). *Teaching young learners English: From theory to practice*. National Geographic Learning.

Sinclair, B. (2000). Learner autonomy: The next phase? In B. Sinclair, I. McGrath & T. Lamb (Eds.), *Learner autonomy, teacher autonomy: Future directions* (pp. 4-14). Longman. https://doi.org/10.1075/aals.1.19sin

Thanasolus, D. (2000). Autonomy and learning: An epistemological approach. *Applied Semiotics, 10*, 115-131.

Yamashita, J. (2013). Effects of extensive reading on reading attitudes in a foreign language. *Reading in a Foreign Language*, *25*(2), 248-263.

Exploring the Shortcomings of the Iranian MSRT English Proficiency Test

Mohammad Reza Ghorbani[1]*, Hadi Abbassi[2] and Abu Bakar Mohamed Razali[3]

[1]*Department of Foreign Languages, Faculty of Humanities, University of Bojnord, 9453155111, North Khorasan, Iran*
[2]*Department of Educational Sciences, Counselling, and Guidance, Faculty of Humanities, University of Bojnord, 9453155111, North Khorasan, Iran*
[3]*Department of Language & Humanities, Faculty of Educational Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*

ABSTRACT

The Iranian Ministry of Science, Research, and Technology (MSRT) English Proficiency Test (EPT) has been in use since 1992. While the MSRT-EPT is generally claimed to be reliable, valid, and practical, it does not assess speaking and writing skills. In this exploratory study, a qualitative approach was used to examine the MSRT-EPT test-takers experiences and language education experts' beliefs about the test as well as their congruence with each other through semi-structured telephone interviews. Convenience and purposive sampling procedures were used to select 15 participants. Inductive coding method was applied to determine invariant constituents. Then, the constituents were reduced to categories, and finally the categories were clustered into 11 themes. Dependability and validity of the study were established through triangulation, inter-coder agreement, and member checking technique. The problems associated with the MSRT-EPT and a lack of productive skills included a lack of correspondence between the test content and Ph.D. Candidates' needs, negative washback effect, non-theory-based content, inappropriate listening conditions, and a lack of test items originality. On the other hand, the candidates' and experts' perspectives were highly congruent. In light of these findings, the importance of designing a more comprehensive test including all facets

of the language proficiency construct was highlighted, and some suggestions were made for future research.

INTRODUCTION

While Standardized Tests (STs) are playing an increasingly prominent role in higher education decisions in recent years, there has always been a torrent of complaints about them. The criticism and grumbling associated with STs are not new phenomena. Proponents of STs argue that they are fair because they measure student ability objectively. In addition, due to their objectivity, STs can be used for comparison and accountability purposes (Churchill, 2015). However, opponents believe that STs are neither fair nor objective (Singer, 2019; Strauss, 2017) because they cannot measure students' actual progress through a one-time performance evaluation (Martinez & Miller 2018).

The International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) are two of the major and widely accepted English proficiency exams for non-native English language speakers intending to enroll in English-speaking universities worldwide. While these two tests differ in format, scoring, approach, and more, they determine students' English proficiency

level by assessing their reading, writing, speaking, and listening skills.

In the Iranian context, English is taught as a foreign language and a subject in high schools and universities. Therefore, Ph.D. students must pass one of the recognized English proficiency tests before graduation. Since taking the TOEFL or IELTS is expensive, the former Ministry of Culture and Higher Education (MCHE) developed a local standardized English Proficiency Test (EPT) known as the MCHE-EPT in 1992. In 2000, the name of the Ministry (MCHE) was changed to the Ministry of Science, Research, and Technology (MSRT). The MSRT was established in 2002. Consequently, the MSRT-EPT is required to be taken by all the Iranian Ph.D. candidates at the state-run universities and higher education institutes, and it is held almost every month. Therefore, this exam is of high importance and has serious consequences for stakeholders.

The MSRT-EPT is a standardized national test to assess the Iranian Ph.D. candidates' overall English as a Foreign Language (EFL) proficiency. This paper-and-pencil test consists listening comprehension, grammar (structure and written expression), and reading comprehension. All three parts of the MSR-EPT consist of multiple-choice questions. The listening comprehension section is comprised of 30 items. Candidates have 30-35 minutes to complete the items. The grammar section is also comprised of 30 items. Candidates have 20 minutes to complete the items. In the reading comprehension section

(four passages usually followed by ten questions), candidates are given 45–50 minutes to answer 40 questions. The total test duration is 1 hour 35–45 minutes. The duration is given in the range depending on the tasks (e.g., the length of the reading comprehension passages), and the allocated time may vary from one test administration to another.

The multiple-choice items in the MSRT-EPT are scored through a computerized scoring system. A test taker's MSRT-EPT score is only valid for two years from the date of taking the test. If the candidates fail to get the required minimum cutoff score (50%), they can register and retake the test without any restrictions. State scholarships are awarded only to candidates who perform above the MSRT-EPT cutoff score (at least 50 out of 100) to continue their studies abroad.

One of the drawbacks of the MSRT-EPT is the probability of guessing the correct answers by test-takers. It is because there are no negative points for wrong answers in the MSRT-EPT. Since test-takers have no marks deducted for giving incorrect answers, this lack of negative points for guessing can lead to chance achievement in test scores (Burton, 2001; Fulcher, 2010). However, Espinosa and Gardeazabal (2010) pointed out that if points were deducted for incorrect answers, test-takers may be cautious and not answer some questions even though they are more likely to choose correct answers.

Noori and Zadeh (2017) state that the MSRT-EPT is generally reliable, valid, and practical. It is well-designed, easily administered, and objectively scored. The benefits of the test include ease of accessibility, a computerized scoring system, and reasonable fees. However, the test does not assess the productive skills of speaking and writing. It is not based on real-world situations and students' needs. It is administered under different and inappropriate conditions. Since the test is not based on the latest testing trends, many students who pass the test cannot communicate in authentic contexts.

While developing and using tests based on the communicative approach was not possible in the past due to a lack of infrastructure facilities, the communicative assessment of all language skills is readily feasible using information and communication technologies in the 21st century (Yildiz, 2019). Thus, there is a strong need to study the MSRT-EPT shortcomings and help the decision-makers adjust the test to fulfill the requirements of the Iranian context by the emerging trends. Therefore, the following questions are formulated to identify the problems associated with the MSRT-EPT and compare the experts' beliefs with the test-takers experiences:

1. What are the problems associated with the MSRT-EPT based on the language education experts' perspectives and the Ph.D. candidates' experiences?

2. How congruent are the language education experts' perspectives and the Ph.D. candidates' experiences on the MSRT-EPT?

REVIEW OF RELATED LITERATURE

Schmidgall et al. (2019) pointed out that defining the assessment construct (e.g., overall English language proficiency), which is the basis for the meaning of test scores, is one of the key steps in the test development process. However, language proficiency unique to humans in its most complex form is an abstract, invisible ability in the brain, which has nothing to do with how a test is constructed. Language proficiency tests measure how well an individual has mastered a language. There are four domains to language proficiency: reading, writing, speaking and listening.

According to the latest theories, the development of these four integrated skills results from social interaction. Social interaction with the environment plays a key role in cognitive development (Vygotsky, 1978 as in Brown, 2000 & Kaufman, 2004). Larsen-Freeman and Long (1991) pointed out that the interactionist perspectives are better than other theories "because they invoke both innate and environmental factors to explain language learning" (p. 266).

Based on the sociocultural theory (interactionist approach), language emerges from social interaction. According to Bachman (2007), social context and abilities to interact in specific situations form the construct, implying that the construct definition in language assessment inevitably involves presenting ability-in-context. Although Norris (2016) acknowledged that task-based assessment conditions must

approximate real-life contexts to indicate the actual performance of test-takers, the MSRT-EPT lacks the speaking and writing assessment sections. Therefore, one question worth asking is whether the test measures the target construct relating to descriptions of the overall English language proficiency of the Iranian Ph.D. candidates who need to use it to take part in international conferences and publish articles in well-established academic journals. Based on the sociocultural theory, which underpins this study, this test seems to be deprived of the sociocultural features of real-world situations.

According to Purpura (2004), the overall language proficiency conceptualized as a multi-componential ability by many researchers consists of four modalities of listening, speaking, reading, and writing as well as linguistic elements such as vocabulary, grammar, phonology, socio-pragmatics. Powers (2013) asserted that testing English-language skills in all four domains drive teaching and learning and improves the overall communicative competence. Bruce (2018) argued that if an assessment does not adequately measure all facets of the intended phenomenon, construct underrepresentation occurs and detrimentally influences the test use, score interpretation, and evaluation.

Since the MSRT-EPT is a high-stakes test having a profound impact on many stakeholders at the national level, a comprehensive, four-skills assessment is in order. This test is similar to the Iranian National University Entrance Examination (UEE), in which listening,

speaking, and writing skills are not tested. Limiting language assessment to grammar, vocabulary, and reading skills in the UEE has led to a detrimental washback effect on students' English learning activities as well as English teachers' curricular planning and instruction (Ghorbani, 2008; Ghorbani & Neissari, 2015).

A study by Ghorbani et al. (2008) revealed that since test scores in the Iranian educational context provide the only benchmark to assess students' progress in schools, teachers usually rate their students based on their performance in the written exams. They argued that teachers might neglect the oral exams because they tend to teach to the test. The findings of another study by Ghorbani (2012) on the controversy over abolishing the UEE in Iran showed that most informants supported the incremental modification of the UEE. In contrast to the UEE, the MSRT-EPT included the listening section. However, the MSRT-EPT is similar to the UEE, which lacks the speaking and writing sections. Hence, identifying the MSRT-EPT problems is the first step for its modification.

Despite the significance of the MSRT-EPT, only a few studies have been conducted on it. Sahrai and Mamagani (2013) studied the validity and reliability of 10 MSRT-EPTs and found that it generally has acceptable reliability (p> 0.7) and validity. However, their study revealed that between the grammar and reading comprehension parts is higher than the correlation between the listening and grammar parts or the listening and reading comprehension parts. Although

the test correlates well with the previously validated and well-established TOEFL, it still requires more substantiation because the TOEFL excludes the speaking skill and measures vocabulary and grammar as separate rather than integrated skills. They believe that the test takers' poor performance in the listening comprehension section of the MSRT-EPT, compared to the reading comprehension and grammar parts, is generally attributed to the listening conditions of the test. They suggested an individual-based listening system to improve the quality of the listening comprehension section of the test. In this study, the informants' perspectives are sought to fill this gap.

Sichani and Tabatabaei (2015) studied the reading comprehension section of the MSRT-EPT using both qualitative and quantitative approaches. The quantitative phase used factor analysis to examine 65 English as a Foreign Language (EFL) students' and 25 experts' perspectives on the reading section. The explanatory factor analysis result did not confirm that the reading section assessed the reading skills. In the qualitative phase, most of the EFL experts and test-takers who were interviewed believed that different items on the reading section of the MSRT-EPT measured the reading ability of the test-takers. While Sichani and Tabatabaei (2015) focused on one section of the test, the present study addresses the test's shortcomings as a whole.

Noori and Zadeh (2017) investigated the strengths and weaknesses of different parts of the MSRT-EPT by analyzing

the test items. They also reviewed the MSRT-EPT related studies conducted to date. They concluded that while the test is generally reliable, valid, well-developed, easily accessible, and less expensive, it still needs more substantiated. To improve the quality of the test, they suggested including the speaking skill, using computerized assessment procedures, considering more integrative communicative items, providing better conditions for testing listening (e.g., using individual-based systems), and penalizing wrong answers (adding a guessing penalty). Each of the suggestions mentioned above is addressed in depth in this study.

Semiyari (2019) studied the MSRT-EPT scores' dependability using G-theory. They examined different sources of variations in isolation (persons, items, sections, gender, and fields of study) and their interactions. The analysis of 1600 pre-intermediate to intermediate participants' performance showed that the test scores were highly reliable. Furthermore, the researchers reported that gender and subject field was negligible, but the difference among persons' performance across items was considerable. This difference probably indicates that high reliability alone is not enough for such an important test.

Each of the studies mentioned above has focused on some specific features of the current MSRT-EPT. Narrowing down a topic and concentrating on its particular aspects can be the strength of a study. While these studies have contributed to a better understanding of the MSRT-EPT, the main weakness is that they have only addressed what is included in the test. The English language proficiency as a unitary construct, which covers all four language skills, is left under investigation. The current MSRT-EPT, therefore, needs to be examined for its shortcomings.

Based on a critical analysis of the current literature, studies have yet to explore the shortcomings of the MSRT-EPT from the perspectives of experts and test-takers. The current study collected the experts' and test-takers perspectives on the test through in-depth interviews using a qualitative approach and a phenomenological research design. By outlining the rationale for a comprehensive four-domain approach to the target construct assessment, the present study investigated the way the MSRT-EPT is viewed by Iranian language education experts who are aware of the theoretical issues associated with the test. It also investigated former Ph.D. candidates' perspectives as they have experienced the test and are aware of the practical issues.

In sum, this study explores the shortcomings of the present MSRT-EPT in measuring the Ph.D. students' overall communicative competence as a unitary construct. The theoretical perspectives of the experts and the practical perspectives of the test-takers can help testing authorities to improve the quality of the test. Although the studies reviewed show that the test is reliable, they are only limited to reading and listening skills. It is, therefore, necessary to identify the theoretical and practical shortcomings of the MSRT-EPT in terms of

all language skills. Since what is theoretical may be different from what is practical, this study was conducted to identify the problems associated with the MSRT-EPT.

RESEARCH METHOD

Research Design

One possible solution to assess the shortcomings of the MSRT-EPT is to examine the test-takers experiences and experts' beliefs. As emphasized by Edwards and Holland (2013) and Flick (2018), qualitative interviewing is generally used to investigate the experiences and perspectives of the interviewees to gain a better understanding of an issue.

This exploratory study used an interpretive phenomenological and qualitative epistemological approach to address the current MSRT-EPT shortcomings. It investigates the test-takers experiences and language education experts' views regarding the test and their congruence with each other. The phenomenological approach was used to describe the MSRT-EPT test-takers lived experiences, and the exploratory expert interview with an epistemological function (Bogner & Menz, 2009) was used to gain experts' knowledge. In this study, test-takers refer to the Ph.D. students in non-English fields, and experts refer to the English language education lecturers.

Ary et al. (2010) noted that since an experience has different implications for different people, researchers should use phenomenological methods like unstructured interviews to explore the perceptions and experiences of individuals. Leimeister (2010) believes that epistemology is the basis of appropriate research methods. Epistemology, the study or theory of knowledge, deals with all aspects of knowledge acquisition, including what constitutes knowledge, how knowledge is acquired or produced, and how its transferability can be assessed (Moon & Blackman, 2014). Epistemology was the most suitable approach in this study because it helped the researchers frame their study and discover knowledge.

A combination of the phenomenological approaches (focusing on the study of Ph.D. candidates' lived experiences) and the epistemological approaches (focusing on the discovery of the language education experts' knowledge) in this study helped the researchers address the problems associated with the MSRT-EPT more comprehensively. Furthermore, this research design enabled the researchers to compare and contrast the language education experts' beliefs with the test-takers experiences. Combining these two congruent approaches helped the researchers analyze and triangulate the data from two different sources, thus enhancing the credibility of the research findings and the study's strength.

Sampling and Participants

The convenience and purposive sampling method was used to recruit 15 participants for this study. In this method, since there is no equal opportunity for all qualified individuals in the target population to participate in the study, the study findings

are not necessarily generalizable to the population. The researchers used convenience sampling because the target subjects were nearer and more accessible to them. Purposive sampling was used to select the subjects suited for the study. The participants included eight test-takers (three male and five female) and seven experts (six males and one female). They were the only available subjects that could serve the purpose of the study. Therefore, the selection criteria and justifications for the number of different groups (e.g., male/female) were limited to the subjects' availability and suitability. The study's research objectives determined the choice of participants, and saturation determined the number of participants.

Due to the coronavirus pandemic, the researchers used personal cell phones to approach and recruit the participants. All participants are associated with the University of Bojnord (UB) and the Kosar University of Bojnord, Bojnord, North Khorasan province, Iran. They are academic staff (11 with Ph.D. degrees and four are Ph.D. students). The experts are proficient in English and native-like. They are familiar with the importance, structure, and function of the MSRT-EPT, and their work experience ranges from five to 32 years. The Ph.D. candidates had passed the MSRT-EPT, with at least an intermediate level of English proficiency. Since the researchers did not have permission to use the participants' names, the participants' initials were used throughout the paper.

Data Collection and Analysis

Due to the coronavirus crisis, the data were collected through in-depth telephone interviews during November 2020. The duration of each conversation was about half an hour. In the first phase, eight different semi-structured interviews (eight interviewees were asked the same questions) were held to elicit data from the test-takers about their experiences with the MSRT-EPT. First, the researchers prompted the participants to describe their experiences with the test carefully. Then, after describing the fundamental features of the test-takers common experiences, the researchers were better positioned to explore the experts' perspectives about the test.

In the second phase, seven different semi-structured interviews (seven interviewees were asked the same questions) were conducted to gather data about the test construct (language proficiency), including what it is, how it is acquired, how it is generated, how it is assessed, and when the results are judged to be adequate to claim that it is warranted or justified. There were only two phases of interviews in this study because the first phase addressed the lived experiences of the Ph.D. candidates, and the second phase addressed the theoretical knowledge of the language education experts regarding the MSRT-EPT. Finally, the test-takers and experts' perspectives were assessed to see how congruent they were.

After gathering the related data based on the research questions from the

participants, the researchers transcribed the test-related interviews verbatim. They searched for significant statements that had particular relevance to the MSRT-EPT. The researchers did not use any software for data analysis. The inductive coding method was manually applied to determine the invariant constituents in the data. Then, the constituents were reduced to categories, and finally, the categories were clustered into themes. Data collection and analysis were continued until saturation was reached. That is, further coding was no longer feasible.

The dependability and validity of the study were established through triangulation using two methods—phenomenology and epistemology—to understand the MSRT-EPT shortcomings. They were also enhanced by an inter-coder agreement in which two of the researchers coded the same transcript and compared the results. Furthermore, they were improved by member-checking. That is, the researchers, in the interpretation process, returned the results to the interviewees to review the interpretations and descriptions of the data and check for accuracy.

According to Johnson and Christensen (2017), classical phenomenologists suggest that researchers bracket or suspend their taken-for-granted orientation towards preconceptions about the phenomenon being studied to experience its essence vicariously. The researchers used the Epoche or bracketing technique in this study. In the phenomenological analysis, it is essential to mitigate the potentially detrimental impact of the researchers' preconceptions that could contaminate the research process. The researchers intentionally set their experiences aside and suspended their own beliefs. They assumed that each interviewee was y unique.

Instead of investigating just the variant part of the data, the researchers sought to understand the essence (commonality or invariant structure) of the experience. The researchers found that certain participants described the MSRT-EPT somewhat differently. While this information was useful in understanding and describing the interesting differences, the researchers were most interested in describing the essence of all the participants. The responses from each participant were considered in the discussion; however, the focus was on the general patterns and findings based on all the subjects' perspectives.

RESULTS

This study benefited from the test-takers experiences and the language education experts' perspectives with regard to the MSRT-EPT problems. The in-depth analysis of the telephone interview results led to some general themes. Based on the test-takers interviews, the following overarching themes were identified for the first research question.

Table 1

The test-takers perspectives on the problems associated with the MSRT-EPT

No	Theme	Frequency
1	Lack of productive skills (speaking and writing)	8
2	Lack of correspondence between the test content and Ph.D. candidates' needs	8
3	No positive washback effect	8
4	Inappropriate listening conditions	5
5	Lack of test items originality	1

Based on the experts' interviews, the following overarching themes were identified for the first research question.

Table 2

The language education experts' perspectives on the problems associated with the MSRT-EPT

No	Theme	Frequency
1	Lack of productive skills (speaking and writing)	7
2	Lack of correspondence between the test content and Ph.D. candidates' needs	7
3	Lack of positive washback effect	7
4	Not based on the latest theories	5
5	Inappropriate listening conditions	5
6	Lack of test items originality	1

The shortcomings of the MSRT-EPT based on the test-takers perceptions and experts' perspectives (Table 1 & Table 2) are combined and summarized in Figure 1 in order of theme frequency to answer the second research question (How congruent are the language education experts' perspectives and the Ph.D. candidates' experiences on the MSRT-EPT?).

As indicated in Table 1 and Table 2, the responses from both participants were in close alignment with each other. The main difference was related to the third theme in Figure 1. While five out of seven experts believed that the test content was not based on the latest theories, the test-takers did not mention this theme.

```
                    ┌──────────────────────────────────────────────┐
                    │ Lack of correspondence between the test       │
                    │ content and Ph.D. Candidates' needs           │
                    └──────────────────────────────────────────────┘
                    ┌──────────────────────────────────────────────┐
                    │ Lack of productive skills (speaking and       │
                    │ writing)                                       │
                    └──────────────────────────────────────────────┘
                    ┌──────────────────────────────────────────────┐
                    │ Not based on the latest theories              │
  ┌───────────────┐ └──────────────────────────────────────────────┘
  │ The           │ ┌──────────────────────────────────────────────┐
  │ shortcomings  │ │ No positive washback effect                   │
  │ of the        │ └──────────────────────────────────────────────┘
  │ MSRT- EPT     │ ┌──────────────────────────────────────────────┐
  └───────────────┘ │ Inappropriate listening conditions            │
                    └──────────────────────────────────────────────┘
                    ┌──────────────────────────────────────────────┐
                    │ Lack of test items originality                │
                    └──────────────────────────────────────────────┘
```
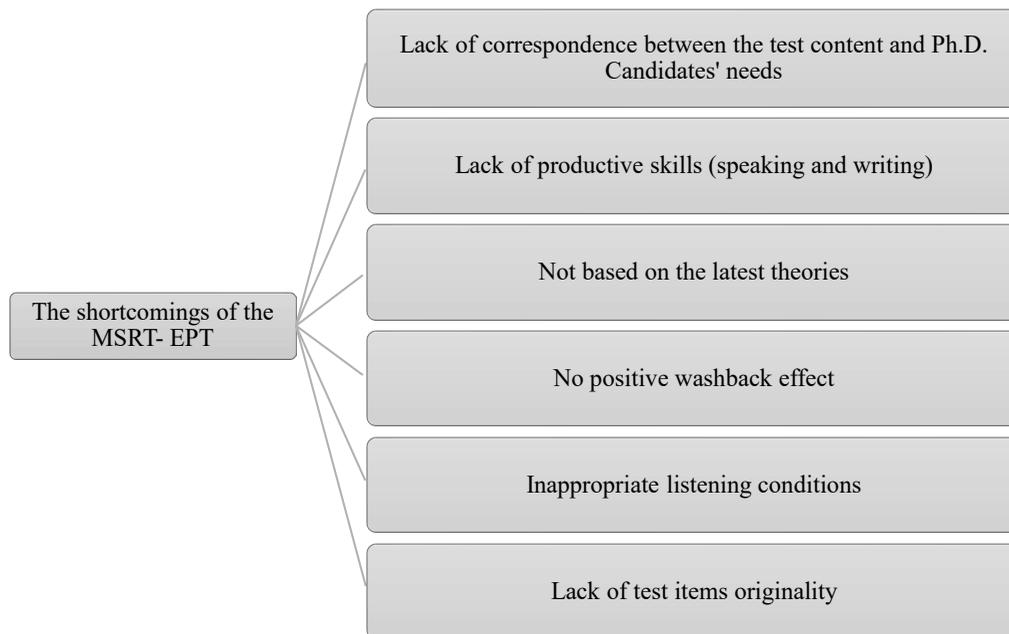
Figure 1. The shortcomings of the MSRT-EPT in order of theme frequency

DISCUSSION

As indicated in Table 1 and Table 2 as well as Figure 1, all the test-takers and language education experts unanimously referred to the problems associated with the MSRT-EPT, including lack of correspondence between the test content and Ph.D. candidates' needs, lack of productive skills, negative washback effect, inappropriate listening conditions, and lack of test items originality. In addition, seven experts referred to the non-theory-based content of the test as the main problem.

A lack of speaking and writing skills suggests that the MSRT-EPT does not adequately measure all aspects of the intended construct (overall language proficiency). Language proficiency has multiple facets. When one of the dimensions is not used in the measurement, construct underrepresentation occurs and negatively affects the test use, score interpretation, and evaluation (Bruce, 2018). A lack of speaking and writing assessment is the first and most frequent theme. This claim is corroborated by an informant (G. H. K.) as follows: *"Language includes four main skills. Grammar, vocabulary, and pronunciation are subskills. Listening and reading alone cannot assess language proficiency. Test-takers are not required to produce anything in the MSRT-EPT. There is no speaking and writing section"*.

This finding is inconsistent with the results of a previous study by Sahrai and Mamagani (2013), who studied the validity and reliability of 10 MSRT-EPTs and claimed that the test generally

possesses acceptable reliability and validity. It also contradicts Noori and Zadeh's (2017) conclusion that the MSRT-EPT is generally reliable, valid, and well-developed. However, it is consistent with their suggestion that including the speaking and writing components can improve the quality of the test as a whole. Therefore, although the test has acceptable reliability and validity, including the productive skills can improve these two key characteristics.

The second most frequent theme is that a lack of correspondence between the test content and Ph.D. candidates' needs. It is confirmed by a respondent (G. H. K.) who mentioned that: *"Ph.D. students are highly expected to write articles in English and take part in international conferences. The test should be comprehensive and based on the candidates' needs. These reading and listening multiple-choice items are not based on the real-world context. Writing and speaking are very important for Ph.D. students. I think, at least, writing should be added to the content of the test"*. It implies that if the test were designed based on the candidates' needs, it would probably reinforce the test-takers motivation and encouragement.

This finding aligns with Powers' (2013) argument that communicative competence is key in English-language proficiency and involves all main language skills (reading, listening, speaking, and writing). In addition, Ph.D. candidates need to use English for article writing and take part in international conferences. Therefore, the four language domains should be included

in the test in an integrative way to satisfy all the needs of the intended test-takers.

A lack of positive washback effect is the third most frequent theme emphasized by all participants. One of the experts (M. E. S.) elaborated on this problem: *"The test-takers just focus on mastering decontextualized grammar and vocabulary. They try to attend private language institutes and classes where they just study grammar and vocabulary books like 504 and 1100 Words so that they can boost their test-taking strategies based on what appears on the test. They cram for the exam. They just want to pass the exam and meet the Ph.D. requirements."*

It is in line with previous research findings regarding the negative washback effect of high-stakes tests on both teachers and students in the Iranian context (Ghorbani, 2008; Ghorbani et al., 2008; Ghorbani & Neissari, 2015). However, the MSRT-EPT washback effect is different from that of other nationwide Iranian tests like the UEE, where teachers adjust their teaching methods, and learners adjust their learning activities based on what appears in the test. University professors do not teach any specific materials related to the MSRT-EPT to be affected by the test format. However, future Ph.D. candidates will focus on the features that appear in the test and ignore what is excluded. Hence, if the test is designed to exert a positive effect, it can be used as a driving force to promote useful learning activities.

The fourth important problem identified from the responses was inappropriate

listening conditions during the test. Ten out of 15 participants referred to this challenge. One of the test-takers (S. K.) stated that: *"Loudspeakers did not function very well. It took the test-administrators a few minutes to adjust and operate them. The voice was not clear. It was vague. It was interrupted. Test-takers were stressful. Loudspeakers were too close to test-takers. A few candidates were next to the loudspeakers. They were not able to understand anything. It was not a good environment for listening at all."*

A language expert (M. E. S.) who had taken the MSRT-EPT a few years ago also reported that: *"The test-takers were not satisfied with the listening conditions at all. I took this test in 2009. I really could not understand what I heard. English Language students have to get 80 out of 100 to meet the Ph.D. requirements. It is too difficult to get the pass mark even for the English language candidates because the quality of the listening part is not acceptable."* When the testing conditions are different for different candidates, the reliability and validity of the test are questionable.

This finding is in line with Sahrai and Mamagani's (2013) recommendation as well as Noori and Zadeh's (2017) suggestion that providing better conditions for testing listening skills through using individual-based systems is essential to improve the quality of the test. Educational technology advancements increasingly make it more feasible to use computer-based testing with individual headsets to provide better listening conditions. As the IELTS is held under the same condition for all candidates in Iran, the MSRT-EPT can also be held under the same condition.

Five out of seven language education experts believed that the test was not based on the latest theories. None of the test-takers referred to this theme because they were probably unaware of the assessment theories and just reflected on their experiences. One of the experts (J. Z.) stated that: *"Items in a test should not be isolated. Language skills are integrated in the real world. In the MSRT-EPT, it is quite clear that skills are not tested in an integrative manner. For example, grammar cannot be separated from reading and writing. It is better to test grammar in students' writing and reading."*

This finding is consistent with the sociocultural theory that language emerges from social interaction. Social context and abilities to interact in specific situations form the construct. That is, construct definition in language assessment inevitably involves presenting ability-in-context (Bachman, 2007). The test in the present format is not based on real-world situations in which there is a natural interaction and meaningful communication.

A lack of originality in the test items is the last theme. Only one test-taker and one expert referred to this problem. One of the experts (M. A. R.) mentioned that: *"All test items have already been used elsewhere. There is no board of exam to develop original items. The test is not standardized. The items are probably taken from the TOEFL and IELTS samples."* It implies

that the test-takers might pass the test by studying sample tests in which some items are sometimes included in the real test.

Furthermore, a test-taker (H. D.) stated, *"I think you can find all the listening section in the market. The items are taken from the TOEFL. Recently it seems that they have tried to change the test and improve it. However, I think test-takers will find the reference of the test items. I think developing a language proficiency test at the national level is too difficult. The test developers have to resort to international standardized test samples to design their local tests."* These findings indicate that since the test items are not original, some test-takers may just review previous sample tests and manage to get a pass. It will, in turn, affect the validity of the test. Furthermore, the listening section is difficult to be developed by non-native English speakers. Due to this, they use the available listening material developed by native speakers.

To answer the second research question (How congruent are the language education experts' perspectives and the Ph.D. candidates' experiences on the MSRT-EPT?), the test-takers experiences and experts' perspectives were compared and contrasted. As indicated in Tables 1 and 2, the responses from both participants were in close alignment with each other. The main difference was related to a theme that five experts had emphasized. While five out of seven experts believed that the test content was not based on the latest theories, none of the eight test-takers referred to any theories.

Since the test-takers field of study was not related to second language education, they might be unaware of the assessment theories. However, they shared test-related problems without mentioning any theories, which indicate a lack of congruence between the test content and the latest theories. For example, the following quotation from one of the test-takers (M. S. H.) confirmed this claim: *"This test lacks speaking. It does not help us improve our communication abilities. It is of no use for enhancing letter or article writing. It is for improving translation only. The test must include a writing section with open-ended items."* This statement refers to the fact that the test's content does not represent the construct of interest accurately.

While the experts' beliefs focused on the theoretical aspects of the MSRT-EPT, the test-takers experiences focused on the practical problems of the test in this study. The MSRT-EPT is mandatory for all Ph.D. students regardless of their field of study; however, the passing mark is at least a score of 80 out of 100 for English major students and 50 out of 100 for the rest of Ph.D. candidates. Therefore, it may justify why the experts' perspectives are somehow similar to the Ph.D. candidates' perceptions.

In sum, the findings of this study suggest that the MSRT-EPT in the present form underrepresents the construct of language proficiency. Furthermore, a lack of alignment between the test content, objectives, the latest related theories, and Ph.D. candidates' needs has led to a detrimental washback effect. Therefore, it

is high time for language education policy-makers and testing authorities to revise and improve the quality of the test based on the latest research findings.

CONCLUSION

This exploratory study examined the MSRT-EPT test-takers experiences with the test and the language education experts' beliefs about the test's shortcomings through semi-structured telephone interviews. The findings revealed that the test is not designed in alignment with the latest theories or the findings of recent studies on language education and assessment. Moreover, since it does not assess productive speaking and writing skills, it underrepresents the construct it claims to measure. A lack of alignment between the test content and its objectives has a pernicious washback impact on the Ph.D. candidates and the related instructors. Inappropriate listening conditions in some testing centers are another serious problem, which has made the test more biased.

These findings highlight significant implications for foreign language policy-makers, testing authorities, test developers, and test-takers. The evaluation based on the test outcome without considering the internal and external factors that affect the reliability and validity of the test may result in incorrect interpretation and decisions. Consequently, as pointed out by most participants, identifying the problems and inadequacies associated with the current test and accordingly rectifying them can improve the validity of score interpretations.

When the quality of the MSRT-EPT is improved, the decision-making will be enhanced accordingly.

In sum, this study highlighted the importance of designing a more comprehensive test, including all facets of the language proficiency construct. The findings of the study contributed to the betterment of the MSRT-EPT in the future. Despite the identified shortcomings, the test can be redesigned and improved to include the productive skills of speaking and writing. Although these skills were difficult to assess in an integrative way at the national level in the past, the development of educational technologies has made it possible to easily include such skills in the test to cater to the needs of the intended candidates.

Testing authorities can improve the quality of the listening conditions, which are different from one context to the other, by using computer-based testing. Each candidate is provided with their headsets. In addition, acoustic standards can be used to create a good listening environment. Poor acoustics in some test settings makes it difficult for test-takers to make the best use of their listening time. Ongoing refurbishment is a chance to modify and improve the acoustic conditions of testing centers.

The current study was limited to the views of eight test-takers and seven language education experts on the problems associated with the MSRT-EPT. The study served as a preliminary investigation of the test. Further quantitative and qualitative studies involving a larger and wider group

of stakeholders are suggested to support or reject the perspectives demonstrated in this study and uncover other test dimensions. It is hoped that future studies provide more insights for the improvement of this high-stakes test. In addition, some of the research topics not covered in this study or ignored in previous studies need further investigation. They are as follows:

To date, no study has addressed the extent to which the test-takers success in the MSRT-EPT can enable them to write scientific papers in English or take part in international conferences. Few studies have dealt with the reasons for not testing in all four domains. The extent to which the testing conditions can affect the MSRT-EPT candidates' performance needs to be investigated. As English language testing authorities have relied on the MSRT-EPT results for decision-making for years, the extent to which such decisions are sound must be investigated.

Since a lack of productive skills in the MSRT-EPT exerts a negative washback effect, future studies can address the degree of the MSRT-EPT washback impact. In addition, future studies can focus on how a comprehensive approach to testing English language proficiency can be implemented. Finally, researchers can study whether it is justifiable to use the MSRT-EPT instead of the well-established standardized tests like the TOEFL and IELTS in the Iranian context.

REFERENCES

Ary, D., Jacobs, L. C., & Sorensen, C. (2010). *Introduction to research in education*, (8th ed.). Cengage Learning.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41-72). University of Ottawa Press. https://doi.org/10.2307/j.ctt1ckpccf.9

Bogner, A., & Menz, W. (2009). The theory-generating expert interview: Epistemological interest, forms of knowledge, interaction. In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing experts* (pp. 43-80). Palgrave Macmillan. https://doi.org/10.1057/9780230244276_3

Brown, H. D. (2000). *Principles of language learning and teaching.* Longman.

Bruce, B. F. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. SAGE Publications. http://dx.doi.org/10.4135/9781506326139.n144

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment and Evaluation in Higher Education, 26*(1), 41-50. https://doi.org/10.1080/02602930020022273

Churchill, A. (2015). *Bless the tests: Three reasons for standardized testing*. The Thomas B. Fordham Institute. https://fordhaminstitute.org/

national/commentary/bless-tests-three-reasons-standardized-testing

Edwards, R., & Holland, J. (2013). *What is qualitative interviewing?* Bloomsbury Publishing. https://doi.org/10.5040/9781472545244

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54*(5), 415-425. https://doi.org/10.1016/j.jmp.2010.06.001

Flick, U. (2018). Doing qualitative data collection - Charting the routes. In U. Flick (Ed.), *The SAGE handbook of qualitative data collection* (pp. 3-16). SAGE Publications. https://doi.org/10.4135/9781526416070.n1

Fulcher, G. (2010). *Practical language testing.* Hodder Education/Routledge.

Ghorbani, M. R. (2008). The washback effect of the university entrance examination on Iranian English teachers' curricular planning and instruction. *The Iranian EFL Journal, 2*, 60-87.

Ghorbani, M. R. (2012). Controversy over abolishing Iranian university entrance examination. *Asian Education and Development Studies, 1*(2), 139-152. https://doi.org/10.1108/20463161211240115

Ghorbani, M. R., & Neissari, M. (2015). Washback effect of the Iranian Concours on senior high school students' EFL learning activities. *Iranian Journal of Language Testing, 5*(1), 1-28.

Ghorbani, M. R., Arshad, A. S., Sahandri, M., & Nooreen, N. (2008). The washback impact of the Iranian university entrance examination on pre-university English teachers. *The Journal of Asia TEFL, 5*(3), 1-29.

Johnson, R. B., & Christensen, L. B. (2017). *Educational research: Quantitative, qualitative, and mixed approaches* (6th ed.). SAGE Publications.

Kaufman, D. (2004). Constructivist issues in language learning and teaching. *Annual Review of Applied Linguistics, 24*, 303-319. https://doi.org/10.1017/S0267190504000121

Larsen-Freeman, D., & Long, M. (1991). *An introduction to second language acquisition research.* Longman.

Leimeister, S. (2010). Epistemological foundation and research methodology. In *IT Outsourcing Governance* (pp. 9-13). Gabler Verlag. https://doi.org/10.1007/978-3-8349-6303-1_2.

Martinez, S., & Miller, R. (2018, November 26). Measuring growth in test scores is key to understanding student progress. *EdSource*. https://edsource.org/2018/measuring-growth-in-test-scores-is-key-to-understanding-student-progress/604959

Moon, K., & Blackman, D. (2014). A guide to understanding social science research for natural scientists. *Conservation Biology, 28*, 1167-1177. https://doi.org/10.1111/cobi.12326

Noori, M., & Zadeh, S. S. H. (2017). The English proficiency test of the Iranian ministry of science, research, and technology: A review. *International Journal of English Language & Translation Studies, 5*(3), 21-26.

Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230-244. https://doi.org/10.1017/S0267190516000027

Powers, D. E. (2013). Assessing English language proficiency in all four language domains: Is it really necessary? In D. E. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies* (pp. 1.1-1.7). Educational Testing Service.

Purpura, J. (2004). *Assessing grammar.* Cambridge University Press. https://doi.org/10.1017/CBO9780511733086

Semiyari, S. R. (2019). Investigating MSRT (MCHE) scores' dependability by means of G-theory: Variations due to persons, items, sections, gender, and subject fields. *International Journal of Language Testing, 9*(2), 56-81.

Sahrai, R., & Mamagani, H. (2013). The assessment of the reliability and validity of the MSRT proficiency test. *The Educational Assessment Journal, 10*(3), 1-19.

Schmidgall, J., Oliveri, M. E., Duke, T., & Grissom, E. C. (2019). *Justifying the construct definition for a new language proficiency assessment: The redesigned TOEIC Bridge® tests - Framework paper* (Research Report No. RR-19-30). Educational Testing Service. https://doi.org/10.1002/ets2.12267

Sichani, E. F., & Tabatabaei, O. (2015). Construct validity of MSRT reading comprehension module in Iranian context. *English Language Teaching, 8*(9), 173-186. https://doi.org/10.5539/elt.v8n9p173

Singer, S. (2019, July 5). *Gadfly on the Wall Blog: Standardized tests are not objective measures of anything*. National Education Policy Center (NEPC). https://nepc.colorado.edu/blog/standardized

Strauss, V. (2017, April 19). 34 problems with standardized tests. *The Washington Post*. https://www.washingtonpost.com/news/answer-sheet/wp/2017/04/19/34-problems-with-standardized-tests/

Yildiz, T. E. T. (2019). Emergent trends and research topics in language testing and assessment. *Akdeniz Eğitim Araştırmaları Dergisi, 13*(28), 66-81. https://doi.org/10.29329/mjer.2019.202.4

Aligning the Language Criteria of a Group Oral Test to the CEFR: The Case of a Formal Meeting Assessment in an English for Occupational Purposes Classroom

Priscilla Shak¹* and John Read²

¹*Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia*
²*School of Cultures, Languages and Linguistics, The University of Auckland, 18 Symonds Street, Auckland Central, Auckland 1010, New Zealand*

ABSTRACT

The Malaysian Education Blueprint (MEB) 2015-2025 has set in motion efforts from all stages of education to align programs, courses, and syllabuses to the Common European Framework of Reference (CEFR) benchmark. This exercise has brought on major revamps in all aspects of English language education in the nation. This study will present such an undertaking in a public university in Malaysia and detail how the language criteria for an oral group test of an English for Occupational Purposes course have been aligned to the stipulated CEFR level. The actual assessment task involved groups of four or five students conducting a meeting of their established company. Data for the study came from an analysis of the audio recordings of nine group meetings, along with post-assessment interviews and focus group discussions involving three EOP instructors. Based on the data analysis, this study recommends a revised set of language criteria for the assessment. Furthermore, it demonstrates how an alignment of the scoring criteria with the descriptors of the targeted CEFR scale can be achieved through a systematic comparison of the language functions (LFs) produced in the meeting task to the targeted CEFR descriptor scales. The revised language component for the meeting assessment could help ease instructors' assessment of students interactional skills and allow them to gauge

better their students' attainment of the skills required in a formal meeting context.

INTRODUCTION

The English Language Education Reform prompted recent prominent transformations of Malaysia's English language education landscape due to the implementation the Malaysian Education Blueprint (MEB) 2015-2025. The MEB, launched in 2015, is a reform plan spanning all stages of education from preschool to tertiary levels, which has resulted in the unified alignment of the English curricula of these institutions to the Common European Framework of Reference (CEFR) (Council of Europe, 2001). The CEFR includes specifications of six levels of proficiency, each of which has been adopted in the MEB as the aspirational target for one level of education in Malaysia: A1for preschool, A2 for primary, B1 for secondary, B2 for post-secondary, and B2 to C1 for university (Ministry of Education Malaysia, 2016).

The CEFR originated as a project sponsored by the Council of Europe in the late 20th Century to promote language learning among adults who had completed their compulsory education. However, it has subsequently become influential at all levels of education in Europe and many other countries worldwide (Byram & Parmenter, 2012; Read, 2019). It is often seen primarily as an assessment scale, and it does serve as a point of reference for many standardized international tests, including IELTS, TOEFL, and TOEIC (Don & Abdullah, 2019; Abidin & Jamil, 2015). However, it has a much broader scope than that: there are multiple scales in the framework that "are accompanied by a detailed analysis of communicative contexts, themes, tasks and purposes" and the "CEFR is used in teacher education, the reform of foreign language curricula, the development of teaching materials and for the comparability of qualifications" (Council of Europe, 2020b).

There have been numerous critics of the CEFR, both in general terms (Fulcher, 2004; Hulstijn, 2007) and more specifically about problems in defining the B2 level for university admission in Europe and Australia (Deygers et al., 2018a; Deygers et al., 2018b). In addition, closer to home Foley (2019) has raised concerns about how the use of the CEFR as a benchmark has been implemented in various ASEAN countries, including Malaysia. Nevertheless, applied linguists have recognized the appeal of the framework to policymakers as a means of articulating language education goals according to internationally defined levels of proficiency and as a tool for accountability in education. As McNamara (2014) has pointed out, "the functionality of a universal letter/number system to code the six levels is a key feature of the CEFR, which makes it attractive to administrators and policymakers" (p. 227).

In Malaysia's case, policymakers insist that a form of standardization is required,

especially to align English graduates' language proficiency across universities and as a form of quality control. As such, it is the public higher learning institutions' role to help the Ministry achieve this target. Accordingly, this article aims to investigate how the assessment of a specific course at a Malaysian university can be aligned to the CEFR B2 benchmark.

The EOP Meeting Assessment as a Test Task

The context of the present study is a course in English for Occupational Purposes (EOP) at a Malaysian university. The students undertake a group project to establish a company, and they are assessed based on their language performance in the task of a simulated company meeting. The main objective of the EOP course is to improve the students' employability by enhancing their language skills to secure future employment and communicate effectively in future workplaces. These include interviewing, presentation, and meeting skills. Specifically, this study focuses on the formal meeting assessment of the EOP course, which is detailed in the next section.

A review of the literature reveals that the meeting test task is somewhat unconventional. For example, Shehadeh (2017) pointed out that there are relatively few studies that investigated the use of task-based language testing (TLBT) in the English for Specific Purposes (ESP) realm despite both sharing similar underlying principles, which are "goal-oriented," "has a real outcome" and "reflects real-life language use and language need" (Shehadeh, 2018, p. 1).

When learners are engaged in a task, they actively focus on meaning-making through interaction in the target language (Nunan, 1989). At the same time, tasks naturally encourage collaboration between learners (Bruton, 2002). In attempting their tasks, learners interact with one another and engage in collaborative efforts to complete the task assigned as there is a real need to do so for mutual benefits (Nakatsuhara, 2013; Shak, 2014; Shak, 2016; Taylor 1983). Therefore, tasks enable language learners to function in "extended, realistic discourse" and help them learn how to use language appropriately for real communicative purposes (Taylor, 1983, p. 70). According to Skehan (1998), managing tasks engages the "naturalistic acquisitional mechanism" that helps learners to develop language skills (p. 95).

For an assessment task to be authentic, it should "parallel those in the real world" (Messick, 1996, p. 3). It means that a task should simulate the target context as closely as possible. Ellis (2003) also highlighted the need for task-based assessment to represent "real-world" behavior and activities (p. 285). In an earlier study undertaken by the first author to investigate the learners' perception of a task-based group project work related to the current study, it was found that the participants viewed the tasks assigned as comparable to a real-world task (Shak, 2014). In addition, for a test task to be useful, it should be informed by the

real-world language use domain (Bachman & Palmer, 1996). Finally, these authors discussed the notion of 'interactiveness,' which refers to the match between the abilities engaged by the test task and those that learners require in the target language use (TLU) context. Following Bachman and Palmer's visual representation, the TLU domains and tasks for this study are presented in Figure 1.
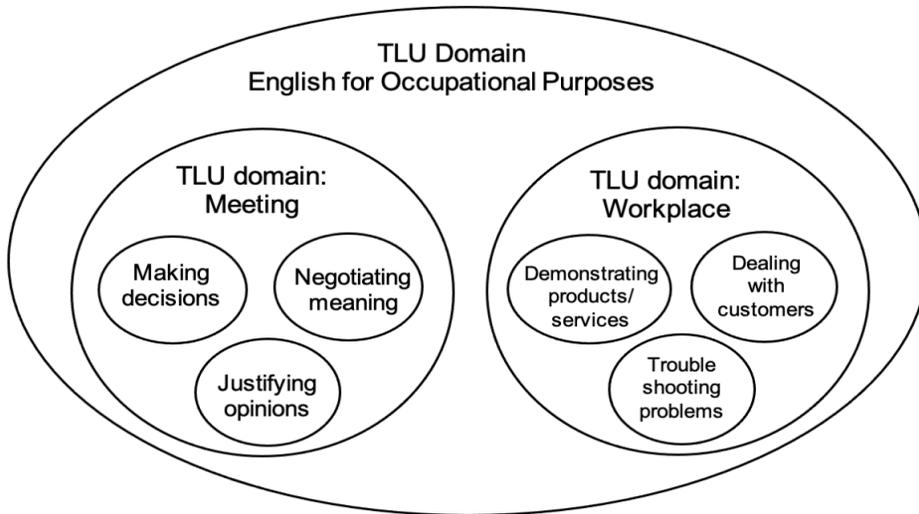
Figure 1. English for Occupational Purposes TLU domain and TLU tasks

As illustrated in Figure 1, the tasks in the TLU domain that apply to the EOP meeting require the test takers to make decisions, negotiate meaning and justify opinions. These functions are among those that are necessary for the successful completion of the meeting assessment task.

Previous studies have highlighted the central role of discourse analysis in offering insights into the nature of interactions in various testing contexts (McNamara et al., 2002; Nakatsuhara, 2013; van Batenburg et al., 2018; Woodward-Kron & Elder 2015). In addition, researchers studying institutional talk have identified formal meeting talk as a genre distinct from other institutional discourse and ordinary conversation (Angouri & Marra, 2010; Asmuß, 2013; Asmuß & Svennevig, 2009; Drew & Heritage, 1992; Svennevig, 2012a; Svennevig, 2012b). Therefore, assessments focusing on this genre should concentrate on its distinctive characteristics and the acquisition of relevant skills to perform the meeting tasks. The appropriate tool for this purpose is Language Function Analysis, which is discussed further in the Data Analysis section below.

The Present Study

The main objective of this study, which is part of a larger-scale project, is to recommend a revised marking scheme for the meeting assessment of the EOP course offered by a Language Centre in a public university in Malaysia. The paper focuses on the alignment of the assessment criteria to the stipulated CEFR B2 level. As such, the paper addresses the following two research questions:

1. What problems did the EOP instructors face when using the existing marking scheme to assess their students' interactional competence?

2. How can the existing marking scheme be revised to align with the CEFR B2 level?

Two sets of qualitative data were obtained from the EOP instructors to address the first research question: individual interviews after the assessment and a Focus Group Discussion (FGD). The synthesized data provided specific details regarding the problems faced by the instructors when assigning marks to their students and their thoughts on the alignment to the CEFR level. For the second research question, results from a Language Function Analysis (LFA) performed on audio recordings of the meeting assessment task were compared to the benchmarked CEFR B2 level descriptor scales for formal discussion (meetings), and recommendations were made based on the findings. The result is a recommended revised version for the language component of the meeting assessment marking scheme.

The EOP Meeting Assessment

The main purpose of the EOP meeting assessment was to evaluate whether the students had acquired the language skills needed to communicate successfully in a meeting setting. In addition, students were tested on their abilities to use language in a formal context and handle such workplace demands in the future. Based on their group project and the roles or positions, each of the students participated in a meeting assessment following a pre-agreed agenda for their group's meeting. The students' main task was to resolve their agenda items to their meeting objective(s). While performing the different roles assigned to them for the meeting test task, students were expected to utilize various language functions such as agreeing, clarifying, suggesting, justifying, negotiating, reciprocating, and interrupting to resolve their agenda items.

The assessment of the meeting task was guided by a marking scheme that contained a list of 16 Likert-type scale items. In accordance with the task-based nature of the EOP group project, the marking criteria focused on the abilities of the students to undertake the meeting task. The evaluation form covered three main components: content and organization (30 marks), presence (20 marks), and delivery, language, and grammar (30 marks). Table 1 lists the items for each of the components. Each item was graded according to a scale of one (very poor) to five (excellent), and each student was assigned individual marks.

While the study was being conducted, the center reviewed all of its English courses

Table 1

EOP meeting assessment's marking criteria

Content and organisation (30%)	Quality of ideas or contents presented in the meeting Sufficient support for ideas Active contribution in the discussion Organized and clear presentation of ideas Perform role assigned effectively Adhere to correct meeting procedures
Presence (20%)	Physical appearance, neatness, and grooming Posture, gestures, mannerism, and movement Eye contact and rapport with group members Listens attentively and shows respect when others are speaking
Delivery, language and grammar (30%)	Enthusiasm and vocal variation (freedom from monotone) Preparation and knowledge of materials (confident and organized) Vocabulary and use of appropriate words (meeting terminologies) Freedom from distracting "uh"s and "like"s Pronunciation, enunciation, audibility, and clarity Grammar

to align them to the Common European Framework of Reference (CEFR) to implement the nationwide English Language Education Roadmap standardization process under the Malaysian Education Blueprint (MEB). As mentioned in the Introduction, part of the MEB requirements is for all English courses in public universities across Malaysia to be aligned to the CEFR's B2 or C1 levels. Given this, the English Language Unit of the Centre determined that the EOP course would be aligned to the CEFR B2 level. This alignment meant that the EOP course would need to produce language learners capable of demonstrating a B2 level of proficiency. As such, it is important that the course assessments could determine whether the learners can perform at this level. Due to this, the assessment criteria of the course would need to be revised according to this benchmark so that an accurate assessment of the learners' proficiency can be correctly mapped to the targeted level.

MATERIALS AND METHODS

The formal meeting assessment involved groups of four or five students. Based on a meeting agenda prepared by the students in advance, each group member was assigned an agenda item based on their role in the project. It provided an information gap as each student had information not available to the others. Following formal meeting conventions, a chairperson was appointed for each group to lead the meeting. Each group was given between 20 to 25 minutes to complete the task. In total, nine meeting groups were audio-recorded.

Each test-taker was awarded individual marks based on the three main rating criteria: a). content and organizations, b). presence, and c). delivery, language, and grammar (Table 1). This paper will focus on the third criterion, the delivery, language, and grammar component.

Participants

In total, 42 second-year undergraduates taking the EOP course and three full-time EOP instructors participated in the study. The student participants had scored Band 1 or 2 in the Malaysian University English Test (MUET), which is a prerequisite for university entrants. The instructor participants recruited the student participants (30 females and 12 males) from their respective classes. Each instructor recruited three groups from their classes. All the instructors were experienced in teaching the EOP course.

Procedures

Each meeting assessment session was attended by the instructor (as evaluator), one group of students (as test-takers), and the first author (as non-participant observer). All the assessment sessions were audio-recorded, as it is less intrusive than video recording for data collection during an assessment event. All the audio files were downloaded into the NVivo 12 software and transcribed orthographically using the transcribe feature of the software. In total, nine transcripts were obtained and analyzed.

All the instructors' post-assessment interview sessions were conducted the week after the meeting assessments. For the post-assessment interviews, a set of semi-structured questions was utilized (Appendix A). Questions relevant to this part of the study included the instructors' feedback regarding their students' performance and their difficulties assigning marks. In total, 136 minutes of recorded data were obtained. In addition, all instructor participants attended a focus group discussion (FGD) as a follow-up to their post-assessment interviews. The FGD was conducted to obtain collective input from the instructors to identify similar issues faced in assigning marks and discuss possible solutions to the problems faced. The FGD lasted for approximately 1 hr 48 min. Appendix B shows the FGD questions.

Data Analysis

The Language Function Analysis (LFA) procedures reported here are situated within a larger project focusing on using group oral assessments in the EOP classroom. For the LFA, both the audio recording and verbatim transcriptions were used concurrently. Therefore, it was necessary to identify the language functions (LFs) that required extensive re-listening and re-reading, and contextual information was essential. The O'Sullivan et al. (2002) Observation Checklist was utilized as an initial operational coding guide (Table 2) to ensure systematic coding of the LFs. Although developed for "real time" use in the Cambridge Main Suite examination paired

speaking test, the successful application of O'Sullivan et al. (2002) checklist was also reported in other studies of oral group tests (Brooks, 2003; Nakatsuhara, 2013).

To ensure that the LFs were coded reliably, the first author and a second coder specializing in English language testing coded all nine transcripts. In instances where there was coding disagreement, specifically those associated with codes where the kappa values were below 0.4, indicating less to a fair agreement (Fleiss et al., 2003; Landis & Koch, 1977; Sim & Wright, 2005; Vierra & Garrett, 2005), the items were further examined and discussed. Upon reaching a final consensus, the kappa values for these items were recalculated. The overall Cohen's kappa value for all of the codes for all the sources is 0.94. Thus, it indicates a high level of inter-coder reliability. In addition, for all codes, average kappa values between 0.71 to 1.0 were obtained.

For the instructors' post-assessment interviews and the focus group discussion (FGD), the audio files were transcribed verbatim orthographically in Word document file format (.docx). The transcripts were then uploaded to NVivo and prepared for coding. Several rounds of close and repeated reading were done before the data were segmented and subjected to thematic analysis coding, allowing researchers to focus on the content highlighted by the participants (Zacharias, 2012). Fereday and Muir-Cochrane (2006) refer to this as "a form of pattern recognition within the data" (p. 82), thus enabling the authors to focus on the specific theme of interest. After the initial coding, the codes

and categories were further refined for final data coding before the data was reported.

For the instructors' post-assessment interviews, the themes were coded under two main categories. The first category coded was the challenges in group discussion assessment, which was further sub-coded into i) the scripted discussion; ii) quantity versus quality; iii) role assignment; iv) personality and v) proficiency. The second category coded focused on the challenges posed by the marking criteria. Similarly, for the FGD, the two main categories identified in the post-assessment interviews were used in the NVivo coding. The sub-themes coded under the theme of the challenges in group discussion assessment were i) the scripted discussion, ii) role assignment, iii) monopoly of talk, and iv) proficiency.

Meanwhile, the sub-themes coded under the theme of the challenges in group discussion assessment were i) generic language component, ii) group collaboration, and iii) interpretation of the assessment items. For this study, codes related to the language component of the marking criteria were highlighted in the results section. Data obtained from the post-assessment interviews and the FGD were instrumental in providing the writers with the directions in which the revised assessment criteria should take; most importantly, they need to move towards a more CEFR-aligned format.

RESULTS

Table 2 presents the range of language functions and corresponding percentage of

test-takers use. Additional LFs not found in the original checklist (O'Sullivan et al., 2002) are shown in bold italic typeface. For example, eight additional LFs under *Interactional* functions were identified, while four additional functions under the *Managing interaction* functions were found.

Table 2

The percentage of test-takers for each of the language functions used

Informational functions	%	Interactional functions	%	Managing interaction	%
Expressing opinions	90.5	Asking for opinions	61.9	Reciprocating	42.91
Providing information	83.3	***Asking for confirmation***	59.5	***Nominating***	33.3
Elaborating	76.2	***Confirming***	59.5	***Concluding***	26.2
Justifying opinions	71.4	***Commenting***	54.8	Changing	23.8
Suggesting	66.7	Agreeing	54.8	***Interrupting***	21.4
Describing	31.0	Negotiating meaning	52.4	Deciding	19.0
Staging	14.3	Asking for information	50.0	***Prompting***	4.8
Speculating	14.3	***Acknowledging***	47.6	Initiating	4.3
Summarizing	14.3	***Instructing***	33.3		
Comparing	7.1	***Assisting***	33.3		
Expressing preferences	4.8	***Assuming responsibility***	26.2		
		Modifying	16.6		
		Disagreeing	9.5		
		Granting permission	9.5		

*Additional LFs in **bold italics** typeface

As can be seen in Table 2, the meeting assessment elicited the highest number of *Interactional* functions (14 LFs), followed by *Informational* functions (11 LFs) and *Managing Interaction* Functions (8 LFs). It demonstrated the propensity of the meeting test task to elicit the desired functions, which in turn indicated the overall effectiveness of the group oral in prompting interaction among the meeting participants. Thus, it

can be regarded as validating the use of the task to assess the test-takers interactional competence.

Apart from that, the additional LFs identified under the *Interactional* and *Managing Interaction* functions were also unique to the test task, which exemplifies how a specific-purpose assessment task could elicit LFs distinct from other types of group interaction. As presented in this section, identifying the LFs elicited from the test task is crucial in recommending a revised language component for the meeting assessment. It will be addressed further in the Discussion section.

The Instructors' Perspectives

This section presents the data collected from the three EOP instructors' post-assessment interview and focus group discussion (FGD) sessions. It primarily discusses the instructors' concerns regarding their difficulties in evaluating their students' interactional skills and assigning student marks. The instructors' post-assessment interviews were necessary to gain their feedback based on their assessed groups and their personal opinions regarding the assessment task. Meanwhile, the FGD was utilized to obtain collective input regarding what the instructors recognized were the main assessment issues regarding the use of the meeting test task. It was especially useful to gauge their views on what needed to be done to improve the meeting assessment further. The results in this section are based on the synthesized findings.

As the meeting discussion was individually assessed, Instructor 2 expressed that some students did not "care about other people" but focused only on speaking during their turns. As such, interaction and input to each other's topics were minimal, and the desired scaffolding did not occur. These test-takers, it seemed, focused only on presenting their ideas, and, as soon as they had voiced their opinions, they ceased to contribute. "When they're not speaking, you know that they're not in the meeting already… Only doing their part, and that's it", said Instructor 2. Although she observed such behavior, Instructor 2 could not penalize her students as such criteria were not stipulated in the marking scheme. Nevertheless, it was an issue for Instructor 2 as she could not adequately assess her students' interactional skills.

Since the meeting assessment was meant to gauge the test-takers abilities to engage in group interaction, they needed to be involved in the co-construction of the interaction rather than merely presenting their ideas. Therefore, the existing marking criteria that focus on language and grammar components are not particularly relevant for assessing the test-takers interactional abilities. For example, one component focused on vocabulary use, specifically meeting terminologies and useful meeting expressions, but that did not cover the test-takers abilities to use such expressions to co-construct the discussion by continuing, elaborating, negotiating and sustaining the topics being considered.

Both Instructor 1 and Instructor 2 agreed that aligning the existing marking scheme to the CEFR would help improve the validity of the marking scheme in assessing the test-takers interactional skills more effectively and fairly. Instructor 1 believed that the test-takers language abilities could be better gauged if they were assessed based on more specific criteria and "not just by performing [the meeting task]." It implies that the test-takers performance should not be judged solely based on their language abilities to complete their own assigned role but also the means through which they collaborated with the others to accomplish the joint task.

Instructor 2 stressed the need to assess both language and meeting management skills as "they are inter-related. Because if you are able to conduct the meeting, definitely, you have a certain degree of language ability in order to carry out all the procedures, convey ideas clearly and understand others." Hence, in her opinion, the assessment criteria should take these aspects into account. As East (2016) has argued, although to a certain extent, task completion is dependent on linguistic abilities, it may not be a sufficient criterion to assess proficiency in this specific context, where proficiency also involves the ability to engage and interact with each other's thoughts and opinions in order to reach a consensus.

For Instructor 3, the existing marking scheme did not pose any problems for her. She typically adhered to it fairly strictly and would award marks based on the criteria stipulated. Hence, she did not assess components absent from the marking scheme. Interestingly, this was an aspect that she did not realize and only became aware of when attending the FGD. It illustrates how relevant interactional skills might have been neglected in these oral assessments as the focus was just on the linguistic aspects of the test-takers abilities. Nevertheless, Instructor 3 agreed that alignment to the CEFR would entail some revisions to the existing language criteria and believed this move would be more positive.

Overall, although all the instructors agreed that the existing marking scheme allowed them to gauge the competencies required to perform the meeting task and could provide information regarding the test-takers abilities to participate in the discussions, the criteria lacked focus on the use of specific language functions, especially those associated with the group interaction in a meeting. This aspect could be improved with alignment to the relevant CEFR scale.

As the study was being undertaken when the alignment of the EOP course to the CEFR had been proposed in line with the Ministry's standardization exercise, there was increased awareness on the instructors of the need to comply with this requirement. As a result, both Instructor 1 and Instructor 2 could pinpoint the specific table for the Formal discussion (Meetings) scale in the CEFR. Table 3 shows the illustrative descriptors for spoken interaction in that context.

Table 3

CEFR's formal discussion (meetings) illustrative descriptors scale (Council of Europe, 2020a, p.78)

	Formal discussion (Meetings)
C2	Can hold their own in a formal discussion of complex issues, putting an articulate and persuasive argument at no disadvantage to other participants. Can advise on/handle complex, delicate, or contentious issues, provided they have the necessary specialized knowledge. Can deal with hostile questioning confidently, hold on to the turn and diplomatically rebut counter-arguments.
C1	Can easily keep up with the debate, even on abstract, complex, unfamiliar topics. Can argue a formal position convincingly, responding to questions and comments and answering complex lines of counter-argument fluently, spontaneously, and appropriately. Can restate, evaluate and challenge contributions from other participants about matters within their academic or professional competence. Can make critical remarks or express disagreement diplomatically. Can follow up questions by probing for more detail and can reformulate questions if these are misunderstood.
B2	Can keep up with an animated discussion, accurately identifying arguments supporting and opposing points of view. Can use appropriate technical terminology when discussing their area of specialization with other specialists. Can express their ideas and opinions with precision and present and respond to complex lines of argument convincingly. Can participate actively in routine and non-routine formal discussion. Can follow the discussion on matters related to their field, understand in detail the points given prominence. Can contribute, account for, and sustain their opinion, evaluate alternative proposals and make and respond to hypotheses.
B1	Can follow much of what is said related to their field, provided interlocutors avoid very idiomatic usage and articulate clearly. Can put over a point of view clearly, but has difficulty engaging in debate. Can take part in a routine formal discussion of familiar subjects clearly articulated in the standard form of the language, or a familiar variety that involves exchanging factual information, receiving instructions, or discussing solutions to practical problems. Can follow argumentation and discussion on a familiar or predictable topic, provided the points are made in relatively simple language and/or repeated, and opportunity is given for clarification.

Table 3 *(Continued)*

	Formal discussion (Meetings)
A2	Can generally follow changes of a topic in formal discussion related to their field, which is conducted slowly and clearly.
	Can exchange relevant information and give their opinion on practical problems when asked directly, provided they receive some help with formulation and can ask for repetition of key points if necessary.
	Can express what they think when addressed directly in a formal meeting, provided they can ask for repetition of key points if necessary.
A1	No descriptors available
Pre-A1	No descriptors available

DISCUSSION

As the authors were made aware of the need for the EOP course to align to the CEFR B2 benchmark, careful consideration was given to meeting this requirement. Hence, in making recommendations for improvement, the authors decided to incorporate the relevant CEFR scale for formal discussion and meetings into the assessment scheme to illustrate what the test-takers should do at the B2 level. However, it has to be pointed at this juncture that a higher number of the LFs produced by the test-takers corresponded more closely to the descriptors below the dividing line after the second statement in the B2 level descriptors. It indicated that the test-takers were likely to be at the lower range of B2 performance, which was to be expected as it represented a more realistic target for Malaysian students with MUET Band 1 and 2 scores. Nevertheless, there were also instances where the more proficient test-takers could produce LFs that reflected higher-level descriptors. Therefore, it indicated that the meeting assessment task was able to elicit LFs beyond B2 level performance. However, as the EOP course has been benchmarked at the B2 level, the revisions were made based on comparison to this level of descriptors.

In order to incorporate elements of the CEFR descriptors into revised language criteria for the meeting test, the authors examined the LFs generated from the meeting assessment, specifically those that yielded higher percentages of test-taker use (ranging from 50% to 90.5%) and compared these to the CEFR descriptors. Table 4 illustrates this comparison.

After examining the corresponding LFs to the CEFR descriptors, the recommended revisions for the language and delivery components were put forth and presented in Table 5 to replace the existing delivery, language, and grammar components of the meeting assessment (Table 1).

Table 4

CEFR B2 descriptors scale for formal discussion and meeting and the corresponding language functions

Level	Descriptors scale for formal discussion and meetings	Corresponding Language Functions
B2	Can keep up with animated discussion, accurately identifying arguments supporting and opposing points of view. Can express his/her ideas and opinion with precision, present and respond to complex lines of arguments convincingly. Can participate actively in routine and non-routine formal discussion. Can follow the discussion on matters related to his/her field, understand in detain the points given prominence by the speaker. Can contribute, account for, and sustain his/her opinion, evaluate alternative proposals and make and respond to a hypothesis.	(Dis)agreeing Supporting Negotiating meaning Expressing/Asking for opinions Justifying opinions Suggesting Asking for confirmation/ Confirming Elaborating Commenting Asking for/Providing information

Table 5

Recommended revisions for the language and delivery components

Language and Delivery
Can present with confidence and enthusiasm (vocal variation, e.g., freedom from monotone).
Can use accurate vocabulary and grammar (appropriate meeting terminologies and sentence structure).
Can speak with correct pronunciation (enunciation, audibility, and clarity).
Can speak fluently (free from lengthy/frequent pauses and distracting fillers, independent of notes).
Can contribute ideas and suggest alternatives.
Can respond to ideas by (dis)agreeing, commenting, confirming, and negotiating meaning.
Can sustain discussion by elaborating, supporting, and justifying opinions and/or arguments.

As presented in Table 5, the recommended version incorporates 'can do' statements, characteristic of the CEFR. These statements correspond to the B2 level of the CEFR's formal discussions and meetings scale. In this revised version, four of the descriptors from the original CEFR list are integrated. Where broader behavioral features are indicated in the CEFR, they are represented more explicitly in the revised version of the marking scheme. For example, at the CEFR B2 level, students 'can keep up with animated discussion, accurately identifying arguments supporting and opposing points of view' (Table 4). These skills are represented in the revised version's abilities to 'present with confidence and enthusiasm' and sustain the discussion by 'elaborating, supporting, and justifying opinions and/or arguments.' It is also worth pointing out that the recommended version does not emphasize accuracy in grammar and pronunciation. Not because these are not important but mainly because these features could be better tested through the other types of assessment that the test-takers have to perform in the EOP course, such as the test, presentation, proposal, and portfolio tasks. As such, the assessment of the meeting task should concentrate more on the abilities of the test-takers to perform interactional functions in such a setting. As Galaczi and Taylor (2018) have recommended, CEFR descriptors should be further refined to meet stakeholder needs. In the case of this study, one of the considerations for the revision of the assessment criteria is the concept of test localization, which "stipulates that for a test to be valid, its design and development must take into consideration the population, context, and the domain in which the test is used" (Abidin & Jamil, 2015, p. 1).

This study has utilized the qualitative bottom-up approach to gain insights into the language produced by the test takers to substantiate the recommendations for a revised marking scheme. At the same time, the post-assessment interviews and FGD with the instructors revealed concerns about the marking scheme and the need to align it with the benchmarked CEFR level, which has illuminated aspects that required improvement.

One of the main aims of language proficiency testing in ESP is to assess test-takers performance based on a simulated setting to predict their capacities to tackle such real-world demands in the future (Basturkmen & Elder, 2004; Douglas, 2000; Woodward-Kron & Elder, 2015). The results of the LFA indicated that, in addition to the LFs found in the assessment of dyads, the group format could generate a wider range of LFs, which lends support to its use for assessing the interactional competence of language learners. Most importantly, the group meeting task could generate language functions that reflect those in natural workplace settings. It is an important aspect of the EOP course as students are exposed to realistic and meaningful interaction. When "the language learners are functioning in the target language in situations similar to the ones they experience every day, they may start internalising English and their motivation may increase" (İlin, 2014, p. 2).

As illustrated in this study, identifying LFs in a meeting setting is instrumental in informing the design of revised marking criteria for the language component of the meeting evaluation form. The recommended language descriptors make it easier for the instructors to evaluate a student's performance. However, as the stakeholders require, they align with the CEFR's formal discussion and meeting descriptors. Despite skeptics' claims, the CEFR can serve as a rich resource for rating scale development and adapted to various testing conditions (Deygers & Van Gorp, 2015; North, 2014; Weir, 2005a; Weir, 2005b; Abidin & Jamil, 2015).

CONCLUSION

This study has illustrated how the language criteria of an EOP meeting assessment can be aligned to the CEFR by demonstrating in detail the steps involved in the alignment process. Qualitative data obtained from the EOP instructors' post-assessment interviews and FGD were utilized to identify the specific issues they faced while assigning students marks to help determine areas requiring revision. In addition, the LFA provided empirical evidence of the LFs elicited by the task. It enabled them to be compared to the CEFR descriptors, which led to the recommended revised criteria.

The methodological implication of the study is that data from the corpus of students' meeting assessment events are a rich and viable resource for the alignment of assessment criteria with the objective and learning outcome of a course. By examining in-depth what was produced by the test-takers in an actual assessment event and comparing this to the targeted performance descriptors, CEFR-compliant assessment criteria could be devised to ensure that the assessment method correlates with the desired level of performance. In this case, the LFA was useful to help gauge the effectiveness of the meeting test task to elicit the desired language output and served as an effective method to map the elicited output to the CEFR's B2 level descriptors for formal meetings and discussions. The result was the recommended CEFR-aligned marking criteria for the language component as presented earlier.

The limitation of this study is that data were collected from just a small number of instructors. Despite this, feedback from these experienced instructors indicated that they were aware of the shortcomings of the assessment scheme utilized then. Another shortcoming is that the trial of the revised assessment has yet to be undertaken. Nevertheless, the proposed revised criteria presentation to the three instructor participants and preliminary discussions indicated that the recommended version would likely ease the challenges of grading the students. In addition, the resulting assessment marks would better reflect the students' interactional abilities. Another limitation concerns the focus of the recommended revisions based on the B2 level descriptors. It has to be acknowledged that it is possible for other lower (B1 below) or higher levels (C1 and C2) LFs can manifest during the formal

meeting assessment. Nevertheless, as highlighted earlier, since the Centre has determined the EOP course to be aligned to the B2 level, the main focus of the revisions in this study was placed on this level's descriptors. Nonetheless, similar processes may be adopted for the other CEFR level descriptors in other contexts based on the steps undertaken in aligning the marking criteria detailed in this study.

An area worth exploring in the future is the trialing and implementing this revised marking scheme to gauge its effectiveness and a further detailed examination of other assessment criteria to enhance further the overall assessment of the students' interactional abilities.

ACKNOWLEDGEMENTS

REFERENCES

Abidin, S. A. Z., & Jamil, A. (2015). Toward an English proficiency test for postgraduates in Malaysia. *SAGE Open, 5*(3), 1-10. https://doi.org/10.1177/2158244015597725

Angouri, J., & Marra, M. (2010). Corporate meetings as genre: A study of the role of chair in corporate meeting talk. *Text and Talk, 30*(6), 615-363. http://doi.org/10.1515/TEXT.2010.030

Asmuß, B. (2013). Conversation analysis and meetings. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 2006-2008). http://doi.org/10.1002/978140598431.wbeal0210

Asmuß, B., & Svennevig, J. (2009). Meeting talk. *Journal of Business Communication, 46*(1), 3-22. http://doi.org/10.1177/0021943608326761

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Basturkmen, H., & Elder, C. (2004). The practice of LSP. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 672-694*)*. Blackwell. https://doi.org/10.1002/9780470757000.ch27

Brooks, L. (2003). Converting an observation checklist for use with the IELTS speaking test. *Cambridge ESOL Research Notes, 11*, 20-21.

Bruton, A. (2002). From tasking purposes to purposing task. *ELT Journal, 56*(3), 280-288. https://doi.org/10.1093/elt/56.3.280

Byram, M., & Parmenter, L. (Eds.) (2012). *The Common European Framework of reference: The globalisation of language education policy*. Multilingual Matters. https://doi.org/10.21832/9781847697318

Council of Europe (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020a) *Common European Framework of reference for languages: Learning, teaching, assessment - Companion volume*. Council of Europe Publishing.

Council of Europe (2020b). *The Common European Framework of reference for languages*. Council of Europe Publishing.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, 32*(4), 521-540. https://doi.org/10.1177/0265532215575626

Deygers, B., Carlsen, C. H., Saville, N., & Van Gorp, K. (2018a). The use of the CEFR in higher education: A brief introduction to this special

issue. *Language Assessment Quarterly, 15*(1), 1-2. https://doi.org/10.1080/15434303.2017.1421957

Deygers, B., Carlsen, C. H., Saville, N., & Van Gorp, K. (Eds.) (2018b). Special issue: Language tests for academic enrolment and the CEFR. *Language Assessment Quarterly, 15*(1), 1-108. https://doi.org/10.1080/15434303.2017.1421957

Don, Z. M., & Abdullah, M. H. (2019, May 17). What the CEFR is and isn't. *Free Malaysia Today*. https://www.freemalaysiatoday.com/category/opinion/2019/05/27/what-the-cefr-is-and-isnt/

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.

Drew, P., & Heritage, J. (1992). Analyzing talk at work: An introduction. In P. Drew, & J. Heritage (Eds.), *Talk at work: Interaction in institutional settings* (pp. 3-65). Cambridge University Press. https://doi.org/10.1075/fol.1.2.08ade

East, M. (2016). *Assessing foreign language students' spoken proficiency: Stakeholder perspectives on assessment innovation*. Springer.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80-92. https://doi.org/10.1177/160940690600500107

Fleiss, J. L., Levin, B. A., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley-Interscience.

Foley, J. (2019). Issues on assessment using CEFR in the Region. *LEARN Journal: Language Education and Acquisition Research Network, 12*(2), 28-28.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1,* 253-266. https://doi.org/10.1207/s15434311laq0104_4

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 218-236. https://doi.org/10.1080/15434303.2018.1453816

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal, 91*, 663-667. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

İlin, G. (2014). Student-teacher judgements on Common European Framework: Efficacy, feasibility and reality. *Journal of Language and Literature Education, 9*, 8-19. https://doi.org/10.12973/jlle.11.221

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

McNamara, T. (2014). 30 years on–evolution or revolution? *Language Assessment Quarterly, 11*, 226-232. https://doi.org/10.1080/15434303.2014.895830

McNamara, T., Hill, K., & May L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221-242. https://doi.org/10.1017/s0267190502000120

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report series, 1996*(1), i-18. https://doi.org/10.1002/j.2333-8504.1996.tb01695.x

Ministry of Education Malaysia. (2016, April 27). Executive summary: *Malaysia education blueprint 2015-2025 (Higher Education)*. Kementerian Pendidikan Malaysia.

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral test. Language Testing and Evaluation, 30*. Peter Lang.

North, B. (2014). *The CEFR in practice. English Profile Studies, 4*. Cambridge University Press.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge University Press.

O'Sullivan, B., Weir, C. J. & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, *19*(1), 33-56. http://doi.org/10.1191/0265532202lt219oa

Read, J. (2019). The influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific region. *LEARN Journal, 12*(1), 12-18.

Shak, P. (2014). Incorporating task-based group project work in English for Occupational Purposes Course: The instructors' perspectives. *MANU Journal, 21*, 77-97.

Shak, P. (2016). Taken for a ride: Students' coping strategies for free-riding in group work. *Pertanika Journal of Social Science & Humanities, 24*(1), 401-414.

Shak, P. (2019). *Towards a framework for effective group oral assessment in the ESP classroom* [Unpublished Doctoral dissertation]. University of Auckland.

Shehadeh, A. (2017). Foreword: New frontiers in task-based language teaching. In M. Ahmadian & M. Mayo (Eds.). *Recent perspectives on task-based language learning and teaching* (pp. vii-xxi). De Gruyter Mouton. http://doi.org/10.1515/9781501503399-015

Shehadeh, A. (2018). Task-based language assessment. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching*. (pp. 1-6). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0379

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268. https://doi.org/10.1093/ptj/85.3.257

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Svennevig, J. (2012a). Interaction in workplace meetings. *Discourse Studies*, *14*(1), 3-10. http://doi.org/10.1177/1461445611427203

Svennevig, J. (2012b). The agenda as a resource for topic introduction in workplace meetings. *Discourse Studies*, *14*(1), 53-66. http://doi.org/10.1177/1461445611427204

Taylor, B. P. (1983). Teaching ESL: Incorporating a communicative, student-centered component. *TESOL Quarterly, 17*(1), 69-88. https://doi.org/10.2307/3586425

van Batenburg, E. S. L., Oostdam, R. J., van Gelderen, A. J. S., & de Jong, N. H. (2018). Measuring L2 speakers' interactional ability using interactive speech tasks. *Language Testing, 35*(1), 75-100. https://doi.org/10.1177/0265532216679452

Vierra, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistics. *Family Medicine, 37*(5), 360-363.

Weir, C. (2005a). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3) 281-300. https://doi.org/10.1191/0265532205lt309oa

Weir, C. J. (2005b). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Woodward-Kron, R., & Elder, C. (2015). A comparative discourse study of simulated clinical roleplays in two assessment contexts: Validating a specific-purpose language test. *Language Testing, 33*(2), 251-270. https://doi.org/10.1177/0265532215607399

Zacharias, N. T. (2012). *Qualitative research methods for second language education: A coursebook*. Cambridge Scholars Publishing.

APPENDICES

Appendix A

Post-assessment interview questions (adapted from Shak, 2019)

1. What do you think of your students' overall performance for the meeting assessment?
   Potential prompts:

   a) Are you happy with the performance of the groups?
   b) Are you happy with the students' performance?

2. For the formal meeting assessment, were there any successful group discussions that stood out?
   Potential prompts:

   a) Why was/were the discussion(s) successful?
   b) What did the students do to make the discussion successful?

3. Did any of the students perform well beyond your expectation of him/her?
   a) Why was the student's/students' performance successful?
   b) How did this affect your marking?

4. During the meeting assessment, were there any students who performed badly?
   a) Why were the students' performance less successful?
   b) What did the students do/fail to do?

5. Do you think the group discussion assessment format is suitable for assessing your students' language skills?

   Follow-ups if YES:
   a) Why?
   b) How?

   Follow-ups if NO:
   a) Why?
   b) What method(s)/format(s) would you suggest instead?

6. In your opinion, is the use of the group discussion assessment fair for the students?

   Follow-ups if YES:
   a) Why?
   b) Please elaborate on why you feel that it is a fair assessment.
   c) What do you do to ensure that the students are assessed fairly in the group assessment?

   Follow-ups if NO:
   a) Why?
   b) Please elaborate on why you feel that it is not a fair assessment.
   c) What do you think can be done to improve the fairness of the group discussion assessment?

7. During their group assessment, the students were assigned different roles. Do you think this will favor some students (i.e., the chairperson of the meeting) while placing the others at a disadvantage?

Follow-ups if YES:
a) Why?
b) How do you think this can be prevented?

Follow-ups if NO:
a) Why?

8. For the group assessment, is there a specific marking scheme that you adhere to? (Refer to marking scheme)
   a) Did you follow the marking scheme strictly when assessing your students? Why? If not, how did you do it?
   b) How did you use the marks sheets? Do you go according to the list of items in the score sheet?
   c) Do you think the marking scheme reflects the objectives of the meeting discussion assessment? How so? If not, how do you think this can be done?
   d) Do you think the marking criteria allow for effective assessment of the specific language skills required to perform the group discussion task?
   e) Do you think that the marking criteria are suitable for assessing the individual language abilities of the students?
   f) Do you think that the marking criteria are fair for all students?
   g) Do you agree with all the items in the marking scheme? Explain.
   h) Did you face any problems while using the marking scheme? Please explain.

i) Did you have any difficulty assessing all the students within the duration of their group assessment?
j) How did you ensure that the assessment was done within the timeframe for each of the students?
k) In your opinion, how can the marking scheme be improved?

9. The course outline specified groups of four students for the group project. In groups where there were more/extra member(s),
   a) How had the extra student affected the assessment process?
   b) How did you manage the assignment of marks in bigger groups?

10. What did you pay attention to when assigning marks to your students? (eg. Language/performance/cooperation)

11. How did/would you assess students who were quiet during the assessment?
    a) Those who are naturally quiet
    b) Those who are weak in the language
    c) Those who cannot get a word in because of other members who manipulate discussion
    d) Those who chose not to contribute when given a chance (the free-rider?)

12. How did/would you assess students who manipulated most of the talk time during the assessment to get a higher score?

13. How did you use your knowledge of your students to help you in assigning their marks?

14. How did you ensure that everyone gets the marks they deserved and that you have marked them fairly?

15. Were your marks set by the end of the assessment? Did you review your marks? How did you do this?

16. What are your suggestions to make the group assessment process more effective?

17. Do you have anything to add?

Appendix B

*Focus group discussion questions
(adapted from Shak, 2019)*

1. What do you think about the topic that has brought us here today (meeting assessment)?

2. I understand that in this Centre, the course chairperson makes most of the decisions about the course design. What are the roles of the other instructors of the course in the decision-making process?
    Items covered:
    *   Course design

    *   Course assessment

    *   Course content

3. In your opinion, what are the major problems in implementing the group discussion assessment format?
    Items covered:
    *   Time constraints

    *   Numbers of students in a group

    *   Students who free-ride (or do not contribute much to the discussion).

    *   Students who monopolize the discussion

    *   The different personalities

    *   The marking scheme

    *   The allocation of marks

    (individual versus group marks)

    *   Whether the marks reflect the individual student's language abilities

    *   Whether the marks given are generalizable to other settings. (i.e., whether being able to perform well in the group discussion assessment means being able to perform in other oral tasks competitively as well)

4. What do you think can be done to overcome the problems you (the instructors) face?

5. Could you provide any suggestions on how the group discussion assessment process can be improved?
    Items covered:
    *   Planning

    *   Strategies to ensure fair evaluation of the students

    *   Marking scheme/criteria

        o   Task versus construct considerations

        o   How to ensure that the student's skills can be captured and are reflected in their scores

      o   How to ensure that the
marking sheet is practical
for use for the group
discussion assessment

6.   Do you have anything to add?

SOCIAL SCIENCES & HUMANITIES

PERTANIKA JOURNALS

Aligning a University English Language Proficiency Measurement Tool with the CEFR: A Case in Malaysia

Nurul Najwa Baharum*, Lilliati Ismail, Nooreen Nordin and Abu Bakar Razali

Department of Language & Humanities Education, Faculty of Educational Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

The Common European Framework of Reference (CEFR) describes the capability of learners' language skills at six reference levels. It is internationally recognised as the standard language proficiency framework for describing language learning, teaching and assessment. Many countries, including Malaysia, have attempted and invested tremendous efforts to adopt the CEFR as a reference for language ability at all levels of education. However, there are many ways of adopting CEFR, and it is a continuous process of alignment between curriculum and assessment. In this regard, this study is carried out to examine how a Malaysian university attempts to demonstrate this alignment by correlating the scores obtained from English language proficiency courses in the university, called the English Language Competence Score Average (ELCSA), to a CEFR-aligned English language proficiency test (Linguaskill). The results showed an overall significant positive correlation that varied in strength. The overall correlation was 0.371, a positive but weak correlation whereby the strongest correlation was seen between ELCSA and CEFR Writing score with a correlation of 0.417, which is positive and moderate in strength. Therefore, it could be identified that a score of 3.25 and 3.5 on the ELCSA can be considered equivalent to a Linguaskill score of 160 (CEFR Band B2). It could be considered that the B2 CEFR level could be subdivided into lower and higher B2. However, there is a need to correlate ELCSA with other CEFR-aligned tests and perform further revisions to the English language proficiency programme at the university to successfully benchmark the programme and its assessment tool, ELCSA, with the CEFR.

Keywords: Assessment, benchmarking, CEFR, English language programme, language testing

INTRODUCTION

There are increasing concerns in establishing standards for the English Language in terms of international benchmarking worldwide (Read, 2019); such language benchmarks standard can be an expressive scale of language ability (Inguva, 2018). Establishing these standards can be quite important in securing places in international higher education institutions and for employment in international companies. The Common European Framework of Reference, CEFR, has become an international benchmark for language competency in many countries, even beyond Europe, such as Mexico, Canada, Japan, and Vietnam. Additionally, many international high-stakes tests such as the IELTS, TOEFL and TOEIC have now been aligned to the CEFR, further underscoring how the framework has gained acceptance and credibility worldwide (Don & Abdullah, 2019). However, despite CEFR being adopted worldwide, research has also claimed that the CEFR still lacks links with stakeholders, socio-educational contexts and empirical validation (Ali et al., 2018).

The Malaysian government acknowledges and stresses the mastery of the English language to gain economic and social leverage in the globalised world. Therefore, it is essential to establish standards and benchmarks that are accepted worldwide to measure proficiency levels among Malaysians. The English Language Standards and Quality Council (ELSQC) and the English Language Teaching Centre (ELTC) of the Malaysian Ministry of Education were given the task to align the

Malaysian English education curricula and assessment with the CEFR, as well as to develop a roadmap for systematic reform of Malaysia's English language education (Prakash, 2019). Following this educational shift, the primary and secondary levels of education have replaced their English language textbooks with CEFR-aligned textbooks. These actions were also followed by the alignment of SPM and MUET examinations in which the results of the test takers English language proficiency were banded against the CEFR descriptors (Sufi & Stapa, 2020).

One of the key issues that surfaced during the adoption of the CEFR in Malaysia was the fear that the Malaysian National Education Philosophy would be side-lined and European cultural values and elements would instead dominate local and national content (New Strait Times, 2019). However, the ministry has organised programmes for teacher training, curriculum familiarisation and adaptation, as well as continued efforts in providing more resources. The use of of-the-shelf CEFR-aligned textbooks (as textbooks for National primary and secondary schools) that were carefully selected, vetted and revised to suit the Malaysian context, by working closely with the publishers, has proven to be more cost-effective and offers a wider acceptance of other cultures along with providing a variety of ways of using the English language in different contexts (Sani, 2018). Implementing the English Language Education Reform in Malaysia was foreseen to be complex, costly and requires persistent efforts and tremendous patience. However,

all these endeavours are for the national advancements that will benefit Malaysia's current and future generation (Ministry of Education, 2015).

Research Problem

While there are some positive indications in referencing language performance against the CEFR at the pre-tertiary level, as stated in the Cambridge Evaluation study in 2017, the impact of such an initiative is not yet seen at the university level (Zulkefli, 2017). The English Language Education Reform in Malaysia stresses the importance of implementing the CEFR in universities to address problems related to poor English communication skills among graduates, which would inadvertently negatively affect their learning experience, employability potential and realise the national agenda (Ministry of Education, 2015). Based on The Roadmap, it was stated that university students are to possess a CEFR B1 level upon university entrance. The Department of Higher Education, Malaysia, stipulated that 'international students' must also sit for exams that reference the CEFR to fulfil the English requirements for university admission purposes (Jaafar, 2019). Additionally, students are required to reach a proficiency of CEFR B2/C1 upon graduation. In accomplishing the required CEFR condition among university graduates, The Roadmap implies that students' English language proficiency may need to be reassessed by the institution prior to their completion of studies (Sufi & Stapa, 2020).

Furthermore, the Malaysian Qualifications Agency (MQA), responsible for quality assurance and accreditation of Malaysian universities, specified that English proficiency courses that are not equated with the international benchmark, CEFR, can no longer be used to fulfil university requirements (Malaysian Qualification Agency, 2020a). Furthermore, the MQA also stated that a CEFR minimum proficiency of C1, or its equivalent in the relevant language, is required to pursue certain job positions in tertiary level institutions (Malaysian Qualification Agency, 2020b). Therefore, there are increasing attempts, demands, and a heightened level of importance for universities to be CEFR-aligned not only of their courses but also their entry and exit grade requirements of universities.

Noticeably, the alignment of CEFR within the tertiary level of education is underexplored and is an area of concern in which further research is required as it affects the efficiency of the English Language proficiency among university graduates. Given that the CEFR is required by the Ministry of Education Malaysia to be aligned with the curricula in the tertiary level education and considering that the Malaysian government has invested a substantial amount of money and effort in aligning the curriculum, it is important to investigate the alignment of English language proficiency assessments at different levels of education to the CEFR standards. Furthermore, investigating the alignment between university English language proficiency evaluation measures

and the CEFR could provide invaluable information to policymakers and test developers about the predictability and comparative values of the university English proficiency assessment with a well-recognised international standard for language education and assessment, namely the CEFR.

Hence, this paper seeks to fill the gap in the literature by attempting to align the accumulative scores obtained by undergraduate students who took an English proficiency programme in a Malaysian public university with the CEFR scores based on the Linguaskill test students sat for. The English proficiency score selected for this study is the ELCSA accumulative score. ELCSA stands for English Language Competence Score Average, an accumulated score derived from a package of English language proficiency courses in University Putra Malaysia. More specifically, the paper will attempt to firstly examine the relationship between the scores obtained in the ELCSA and the overall as well as individual language skill scores on the Linguaskill test, and secondly, identify the ELCSA score that is equivalent to a CEFR B2 level which has been targeted as the minimum CEFR level for Malaysian university graduates. By doing so, the paper can contribute to a greater understanding and contextualisation of the CEFR. Furthermore, aligning ECLSA scores to the CEFR will help provide comparative scores in ELCSA with Linguaskill. It could then provide indications of test-takers CEFR levels based on ELCSA accumulative proficiency scores and could assist and contribute to the university's benchmarking efforts of an internally developed English proficiency measurement tool with international standards.

LITERATURE REVIEW

Benchmarking

There are various definitions given on the concept of benchmarking in the literature. For example, Bogan and English (1994) stated that benchmarking is the continuous pursuit of best practices. By establishing measurement points, comparisons can be made for reasons of learning, adapting and ultimately resulting in better performance, which is the main purpose of benchmarking (Fisher, 1996). The essence of benchmarking is also inspiring ongoing learning and boosting organisations to be at their best (Zairi, 1996). The intent of benchmarking is to aid organisations in establishing a baseline performance criterion that should be complied with (Nwabuko et al., 2020). Similarly, Keegan and O'Kelly (2012) consider benchmarking as a method of comparison between organisations to obtain insights from each other. Benchmarking is operative in identifying best practices, and these practices are applied for the benefit of the organisation (Alosani et al., 2016).

In education, especially with second language learning, benchmarking is required when measurable standards are set for learning (Inguva, 2018). Benchmarking in assessments ordinarily attends to the purposes of evaluating and monitoring program efficiency, planning curriculum and

instruction, communicating expectations for learning and predicting future performance whereby it would operate finest when it is specifically designed to deliver the data required for enhancements to be made (Herman & Osmundson, 2010). Benchmarking could also provide information about the position of a specific student, class, or institution in terms of ranking (Canadian Language Benchmarks, 2012). There are many views to the term benchmarking as it is used in various contexts. Nevertheless, it could be said that benchmarking is an ongoing process of seeking the best practices by making comparisons and creating points of reference so that the effectiveness of a particular programme could be identified and further improved.

Regarding benchmarking language learning and assessment, the CEFR framework has proven to be an influential baseline for the development of language curricula and assessment around the world (Read, 2019). However, benchmarking curricula to the CEFR has brought a great deal of discussion whereby some countries found it problematic to strike a balance between the appeal of establishing mutual international standards and the importance of representing the unique educational and social contexts of distinct countries in language learning (Read, 2019). In Taiwan, for example, attempts were made to adopt the CEFR, which meant that their recognised tests needed to be calibrated against the CEFR (Wu, 2012). However, Wu (2012) pointed out that there were several problems

with the process of calibrating tests to the framework, such as the conceptual difficulty in comparing the results of tests that have been designed differently and the lack of technical expertise to confirm the alignment of CEFR upon their tests. Furthermore, Wu (2012) mentioned an unclear relationship between the assessment of English language proficiency according to the CEFR and the grading criteria used by universities. Additionally, it was reported that Taiwan students did not have the exposure to the language to use it communicatively as described on the CEFR scales (Cheung, 2012).

On the other hand, there were also instances where some researchers suggested developing a new framework of reference altogether. For example, in China, rather than adapting the CEFR, the development of a Common Chinese Framework of Reference for Languages (CCFR) or currently known as China Standards of English (CSE) which has been established without much reference to the other frameworks and with their separate tests as measures of student achievement was proposed (Jin et al., 2017). Meanwhile, there are instances where these efforts to align the CEFR with curricula succeeded. For example, in Japan, a team of language researchers at the Tokyo University of Foreign Studies undertook a project to adapt the CEFR to the Japanese context, which successfully resulted in a version of the framework labelled CEFR-J whereby they added sublevels (A1.1, A1.2 and A1.3) to reflect better the degree of English ability (Markel, 2018).

English Proficiency Courses in a Malaysian University

One of the measures taken by universities to improve English language proficiency among students is to offer a range of English language courses required for students to pass as part of their graduation requirements (Rethinasamy & Chuah, 2011). It is also a measure taken by one of the research universities in Malaysia, Universiti Putra Malaysia (UPM). They revamped their English language proficiency level courses and developed an innovative package referred to as the English Language Experience (ELEx). The purpose of the ELEx package is to engage students with the language in a variety of formal and informal situations as well as involve more student-centred courses and task-based language activities. ELEx consists of three components, namely conventional courses (LPE), non-alphabet grade preparation courses (CEL), and language activities (LAX). The number of English skills courses, CEL courses, and LAX activities that students need to take is determined by the MUET results obtained before students start their studies at UPM. Therefore, students who get low results in MUET need to take more English courses and activities than students who achieve high results.

UPM enhanced the ELEx package by implementing a cumulative, and summative assessment of language performance referred to as the English Language Score Competency Average (ELCSA). The ELCSA is obtained calculating average achievement points for the two components of ELEx, namely conventional courses (LPE) and preparatory courses (CEL). LAX activities are not included in this calculation because they serve as support (scaffold) to forming English language skills by allowing students to use the language and build confidence in its use. The assessment of this English language achievement, named English Language Competence Score Average (ELCSA), will be calculated at the final stage of the study program and will be stated in the student transcript. It is also important to add that the ELSCA is isolated from the existing CGPA. Therefore, it does not interfere with nor affects the student's CGPA. Therefore, ELCSA serves as a cumulative summary of the student's achievement in their English language skills.

As mentioned, the targeted level for university graduates is the B2 level of the CEFR whereby at this level, it is expected that graduates can understand complex texts, tackle other abstract topics, engage in discussions as well as be able to communicate with native speakers with ease (Ministry of Education, 2015). A major motivation in introducing the ELCSA is to provide a measure that can indicate the student's English language performance according to the CEFR bands. The ELSCA scores could act as a comparison point compared to other CEFR achievement tests such as IELTS, TOEFL, Linguaskill and MUET. In addition, it could evaluate the effectiveness of the ELEx package.

The ELEx Program at UPM and the Development of ELCSA

In 2013, the Centre for the Advancement of Language Competence (CALC) in UPM executed its undergraduates' English Language Experience (ELEx). The students must follow a carefully developed set of courses and activities during the whole duration of their educational programme.
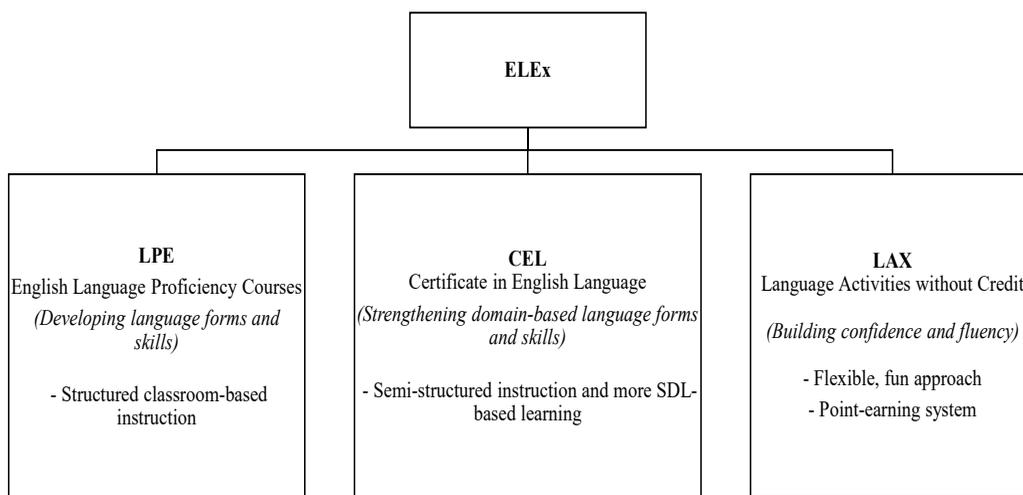

Figure 1. Overview of the ELEx Structure

The components that construct the ELEx are portrayed in Figure 1 (Centre for the Advancement of Language Competence, 2013). As displayed, the ELEx package comprises three significant parts. The LPE component focuses on building the basis of language whereby the knowledge of vocabulary and grammar would be solidified to achieve language accuracy and fluency. Similarly, the CEL component is also constructed to assist in the mastery of vocabulary and grammar. In addition, it also emphasises domain-based learning, whereby it serves to accommodate learning English for general, academic and professional purposes. On the other hand, the LAX component focuses on incidental learning via task-based activities, aiming improve students' confidence and familiarity in using the language.

Recognising that students vary in levels of proficiency from being very limited to very proficient users of English, the ELEx package is designed to cater to students' specific language needs, which is identified based on the levels that they have achieved in their MUET results (Band 1 to Band 6). MUET is a compulsory test that students have to take in order to be admitted into a university. Thus, ELEx provides students of MUET Band 1 or 2 with an intensive programme that aims to supply essential assistance to help foster their confidence in the language while assisting them to meet

their immediate needs required in academic tasks. For those with MUET bands of 3 to 6, the package would provide programmes and courses that aim to enhance further and polish their language competency as well as amplify their confidence and fluency in order to be more linguistically marketable (Abdullah et al., 2015).

The implementation of the ELEx package has proven to show positive outcomes in improving language proficiency, especially among the less proficient students, whereby the students portrayed higher willingness to use the language, which resulted in higher participation and interaction in various contexts (Mustafa, 2018). Although the ELEx package is compulsory for all students, it is seen as more of assistance for students to cope better with their studies rather than an obligation or a test (Sani, 2020). It could be said that this package delivers and is in line with the aspiration of the Ministry of higher education in developing graduates that possess adequate English language abilities. In fact, in the 14th parliament meeting on July 22, 2019, the Ministry of Education mentioned and

acknowledged the ELEx package from UPM as one of the government's efforts in assisting youths in mastering the English language (Parlimen Malaysia, 2019)

Since its implementation, the assessment for the ELEx package for each student was evaluated via an alphabetical grade for the LPE component, a 1 to 4 level for the CEL component and a Satisfactory / Unsatisfactory grade for the LAX activities. However, a comprehensive evaluation of the student's English language proficiency was not provided at the point of graduation. Therefore, the English Language Competence Score Average (ELCSA) was established, and the students will obtain scores ranging from 0.0 to 5.0. It is obtained by calculating average achievement points for two components of ELEx, namely the conventional courses (LPE) and preparatory courses (CEL). LAX activities are not included in the calculation as the activities in LAX were for scaffolding purposes that provided opportunities to use the language and build confidence. The division of courses and calculation of the student's ELCSA is shown in Table 1.

Table 1

Courses are taken into account for the determination of ELCSA according to MUET results

MUET band level	CEL courses	Number of CEL courses	LPE courses	Number of LPE courses	Total courses
1 – 2	CEL2102, CEL2103 and one of the courses CEL2105 / 2106/2107	3	LPE2301, LPE2501	2	5

Table 1 *(Continued)*

MUET band level	CEL courses	Number of CEL courses	LPE courses	Number of LPE courses	Total courses
3 – 4	CEL2103 and one of the courses CEL2102 / 2105/2106/2107	2	LPE2301, LPE2501	2	4
5 – 6	CEL2103	1	LPE2402 and / or LPE2502	1 or 2	2 or 3

The scores that will be given for the achievement of each CEL and LPE course are in Table 2 and Table 3.

Table 2

Scores for Certificate in English Language (CEL) Course

Level	Score
1	0
2	3.0
3	3.5
4	4.0

Table 3

Score for courses of Language Proficiency in English (LPE)

Alphabetical Grade	LPE2301*	LPE2501*	LPE2402**	LPE2502**
A	4.0	4.0	4.5	4.5
A-	3.7	3.7	4.0	4.0
B+	3.3	3.3	3.7	3.7
B	3.0	3.0	3.3	3.3
B-	2.7	2.7	3.0	3.0
C+	2.3	2.3	2.7	2.7
C	2.0	2.0	2.3	2.3
C-	1.7	1.7	2.0	2.0
D+	1.3	1.3	1.7	1.7
D	1.0	1.0	1.3	1.3
F	0	0	0	0

Taken only by MUET students 1 - 4

** One or both courses are taken only by MUET Students 5-6

The calculation of ELCSA is based on the total score obtained divided by the number of selected LPE and CEL courses taken (i.e., on average). The average score obtained will determine the level of ELCSA as described in Table 4. As mentioned before, LAX activities are excluded from the calculation because their main purpose is to build confidence in using English. The LPE 2401 course is also excluded because it is in special preparation for students with MUET results 1 and 2. Higher scores are given for LPE2402 and LPE2502 courses as these two courses are high-level courses taken only by MUET students 5 and 6.

Table 4

Scale for English Language Competence Score Average (ELCSA)

Score	Competency	Grade	Estimated CEFR
3.90 above	Excellent	A+	C2
3.725 - 3.89	Very High	A	C1
3.5 - 3.724	High	A-	B2
3.0 - 3.49	Competent	B+	B2
2.5 - 2.99	Average	B	B1
2.5 and below	Low	B-	B1

Linguaskill English Language Proficiency Test

Linguaskill is one of the tests provided by Cambridge Assessment English and has just recently been introduced in Malaysia in 2020. In implementing the CEFR, Cambridge Assessment English played a contributing role and possesses increasing, ongoing and various evidence that supports it to be the embodiment and reflection of the CEFR in multiple aspects (Cambridge Assessment English, 2021b). Linguaskill is a CEFR-aligned, computer-based, multi-level test that assesses one's English language proficiency in writing, reading, listening and speaking (Cambridge Assessment English, 2019). Table 5 illustrates the Linguaskill scores and corresponding CEFR levels.

Table 5

CEFR scores and levels

Cambridge English Scale Score	CEFR Level
180+	C1 or above
160–179	B2
140–159	B1
120–139	A2
100–119	A1
82–99	Below A1

The Linguaskill test provides two test options, namely Business and General, whereby Linguaskill Business assesses the familiarity of the test-taker towards the language of business. At the same time, Linguaskill General would focus on assessing English used in daily life (Cambridge Assessment English, 2019). Linguaskill Business has replaced BULATS that was officially discontinued on December 6, 2019 (Cambridge Assessment English, 2021a). While the Ministry of Higher Education Malaysia (MOHE) had specified several English competency tests (e.g., MUET, IELTS, and TOEFL iBT) that can be recognised by universities to meet English language requirements for student admission, the Linguaskill, Cambridge English Qualifications and OET was also added to the list in 2020 (Cambridge Assessment English, 2021d) Linguaskill reports up to a maximum score of 180+ which is equivalent to C1 or above on the CEFR scale (Cambridge Assessment English, 2021c). Linguaskill was developed by a team of experts and is supported by artificial intelligence. The trial report in April 2016 shows that the Linguaskill test scores are reliable and precise (Cambridge Assessment English, 2016). An analogous measure, the Rasch reliability, was used, and each test obtained a reliability coefficient over .90, which is considered adequate. Whereas the target level of precision was roughly 90% in which most of the tests that failed to reach the target precision were at the extremes of the CEFR: Level A1 or below and C1 or above (Cambridge Assessment English, 2016).

For this research, the Linguaskill General test was used. As mentioned earlier, the Linguaskill General test assesses language used in day-to-day life. The test would include topics involved with studying and working, making plans, travel and technology. Thus, it makes the test suitable for a broad spectrum of organisations, university admissions or exits. The test could also be used for recruitment roles that do not require specialist business terminology; for instance, it would be suitable for employees who are required to showcase their strong command in English to perform their roles effectively.

The Linguaskill General test has three modules which are reading and listening, speaking and writing. The reading and listening tests are adaptive according to the candidate's proficiency level, meaning that each candidate would face a different set of items on their test based on how well they answered the previous question (Cambridge Assessment English, 2018). Although there are not a fixed number of questions, each question the candidates' answer would help the computer understand their level better. The test finishes when the candidate has answered enough questions for Linguaskill to identify their level accurately. The writing test uses innovative auto-marker technology whereby the computer automatically marks it. Meanwhile, a hybrid approach was taken to mark the Linguaskill Speaking test, which uses auto-marking technology and human examiners to ensure efficiency (Xu et al., 2020).

RESEARCH METHODOLOGY

Generally, there are two approaches in which alignment to the CEFR can be adopted: the direct alignment and the indirect alignment (Bruce & Hamp-Lyons, 2015). The direct alignment would require much expertise, resources, and funding which is made possible by large organisations such as Cambridge English Assessment (Ali et al., 2018). Due to time and financial constraints, the direct approach would not be feasible. Alternatively, the indirect approach to the CEFR is adopted by mapping test scores to the CEFR-aligned scores. However,

certain factors regarding the language test such as its purpose, format, test-takers, and the scoring system should be considered before the indirect alignment can be made (Ali et al., 2018). This indirect linkage via 'equation' to an existing test already linked to the CEFR is one of the recommended approaches in the Council of Europe's Manual (Cambridge English Language Assessment, 2011).

According to the Standards for Educational and Psychological Testing (American Educational Research Association, 1999), scores can be considered 'comparable' or 'equivalent' when the test's features are closely similar to each other (Lim, 2017). In this case, this study attempts to uncover the relationship between ELSCA scores and the Linguaskill test scores. Both are designed to measure English proficiency directed towards the goal of real-world applications. Therefore, in order to fulfil the purpose of this research, a quantitative,

correlational design was utilised in this study involving the collection of quantitative data followed by a correlational analysis as the study intended to examine the extent to which two or more variables relate to one another (Rockinson-Szapkiw, 2012).

Data Collection and Analysis

The researchers were given access to the participant's ELSCA scores and their full Linguaskill test report, including their overall CEFR score and language skill scores. The researchers used Excel to compile the participant's scores accordingly and then proceeded to use the IBM SPSS Statistics software to calculate the Spearman Rho correlation to uncover the relationship between the two variables. A Spearman Rho correlation was used in this study as it can describe two variables in a monotonic relationship. It should be mentioned that the Spearman Rho correlation seems most befitting as it is suitable for data that is, either ordinal, interval and ratio variables, continuous and non-normally distributed (Schober et al., 2018). In ensuring the standard of quality when assessing the correlational analysis, outliers were addressed and removed. The presence of outliers is common in data collection due to various reasons. It, therefore, is crucial to be dealt with prior to the analysis to ensure the overall reliability of the results (Kwak & Kim, 2017). Additionally, a scatter plot was constructed to observe the relationship between the two variables further, and a trend line was identified.

Sampling

The participants were 197 final year undergraduates from six Science, Technology, Engineering and Mathematics (STEM) based faculties. Table 6 shows the participant's profile.

Table 6

Participants' profile

Participants	Category	Number of test takers	Percentage (%)
Gender	Male	41	20.8
	Female	156	79.1
Age range	21-23	153	77.7
	24 and above	44	22.3
Faculty of	Biotechnology and Biomolecular Sciences	34	17.3
	Computer Science and Information Technology	32	16.2
	Engineering	34	17.3
	Food Science and Technology	34	17.3
	Medicine and Health Sciences	30	15.2
	Science	33	16.7

As shown in Table 6, the participants were 197 final year students (M= 41, F= 156) from six STEM-based faculties in UPM. The purposive sampling method, specifically the Homogenous Sampling, was applied as this sampling form focuses on a particular characteristic of a population where they share similar traits (Etiken et al., 2016). In this case, the participants were chosen according to the following criteria; 1) Participants have completed their undergraduate programmes and therefore also obtained their ELSCA scores. 2) Participants have taken the Linguaskill test and obtained their CEFR band level. 3) Participants were among the STEM-related faculties. The number of participants from each faculty ranged from 30 to 34. The highest number of participants were from the Faculty of Engineering (n= 34), the Faculty of Food Science and Technology (n= 34), and the Faculty of Biotechnology and Biomolecular Sciences (n= 34). It is followed by the Faculty of Computer Science and Information Technology (n= 32) and the Faculty of Medicine and Health Sciences (n= 30). This study specifically chose Science Technology Engineering and Mathematics (STEM) undergraduates because based on their MUET scores, the STEM undergraduate students have varied levels of English language proficiency, which may provide better insights into the correlation between the CEFR and ELCSA

scores. For courses related to English, they are required to meet a MUET band 4 to be admitted in the course (UPM, n.d.). Furthermore, past research has shown that STEM graduates have low employment rates, possibly due to a lack of multiple skills and English proficiency (Murtaza & Saleh, 2018; Thomas, 2019). Additionally, the participants obtained both an ELSCA and a Linguaskill General score, thus allowing the comparison and correlation between ELCSA and Linguaskill.

RESULTS

The performance of the students on the ELCSA and CEFR is presented in Table 7.

Table 7

Performance of Respondents on ELCSA and CEFR

	N	Minimum	Maximum	Mean	Std. Deviation
ELCSA	197	3.000	4.000	3.667	0.211
CEFR	197	122	180	168.43	10.012

The mean ELCSA and CEFR Linguaskill scores were 3.667 and 168.430, respectively. Thus, the Linguaskill score indicates that, on average, the UPM STEM undergraduates had successfully achieved the B2 level as targeted by Malaysia's Ministry of Education.

A correlational analysis between the CEFR and ELCSA scores is presented in Table 8.

Table 8

Correlation between ELCSA and CEFR scores

			ELCSA	Overall (CEFR)
Spearman's Rho	ELCSA	Correlation Coefficient	1.000	0.371**
		Sig. (2-Tailed)	0.000	0.000
		N	197	197

**. Correlation is significant at the 0.01 level (2-tailed)

In fulfilling research objective 2, results show a positive, weak relationship according to the Guilford Rule of Thumb between ELCSA and CEFR scores. In addition, results of Spearman Rho correlation indicated that there was a significant positive association between the overall ELCSA scores and CEFR scores, (r_s (195) = 0.371, p < .05).
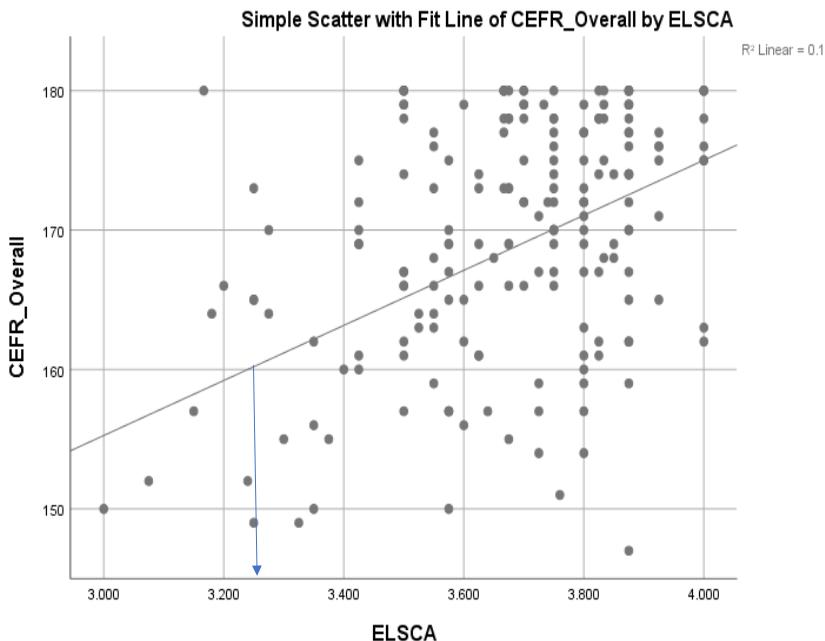
Figure 2. Simple Scatter plot of ELCSA and CEFR scores

Scores on the CEFR and ELCSA were also placed on a simple scatter plot, and based on the trend line in the scatter plot (Figure 2), a score of approximately 3.25 on the ELCSA can be considered equivalent to a Linguaskill score of 160 (CEFR Band B2). However, the trend line does not allow for predicting the C1 Band based on the ELCSA.

The correlations between the language components in the Linguaskill and the ELCSA are presented in Table 9.

Table 9

Correlation between ELCSA and CEFR scores

			Writing (CEFR)	Reading	Speaking	Listening
Spearman's Rho	ELCSA	Correlation Coefficient	0.417**	0.360**	0.249**	0.179*
		Sig. (2-Tailed)	0.000	0.000	0.001	0.012
		N	196	195	188	197
** Correlation is significant at the 0.01 Level (2-Tailed)						

Regarding the individual skills, there are correlations of varied strengths between the ELCSA and each of the four skills. There is a positive and moderate relationship between ELCSA and CEFR Writing scores with a correlation coefficient of 0.417. Also, there was a positive, low relationship between ELCSA and CEFR Reading scores with a correlation coefficient of 0.360 and CEFR Speaking scores with a correlation coefficient of 0.249. However, although positive, the relationship between ELCSA and Listening scores was negligible, with a correlation coefficient of 0.179. The relationship that was considered best and strongest was that of ELCSA and writing skills. For that reason, as well as writing being especially important in academic contexts, this relationship is further explored as in Figure 3 in order to determine the ELCSA score that would best reflect a B2 CEFR level.

The scatter plot of scores on the ELCSA, and the Linguaskill Writing skill is presented in Figure 3.
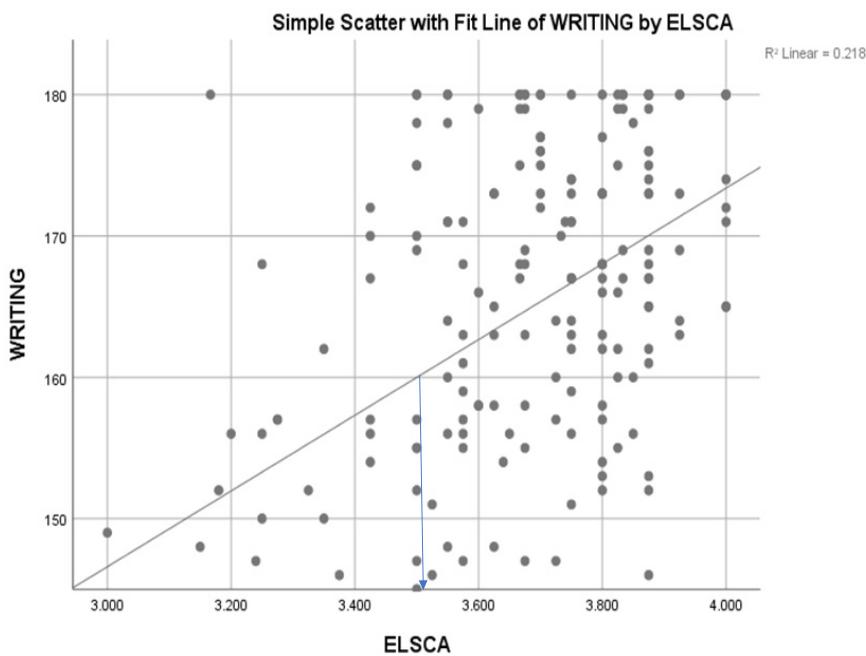

Figure 3. Simple Scatter plot of ELCSA and CEFR Writing scores

As the writing component of the Linguaskill test was the language skill that yielded the strongest correlation with the ELCSA, the scatter plot was used to identify the ELCSA score comparable to a B2 CEFR level. Based on the trend line, the ELCSA score of approximately 3.5 could be identified as equivalent to the CEFR Writing score of 160, which the Linguaskill test specifies as representing the B2 level.

DISCUSSION

The study demonstrates that there is a positive relationship between the ELSCA and CEFR scores. It means that the students that managed to get a high score in ELSCA also managed to get a high CEFR score in the Linguaskill test, and it is likewise for those who received low scores in ELSCA also possessed a low CEFR score. The correlations did, however, differ to some extent in terms of strength. The possible explanation for why the correlations varied in terms of strength is that, though the two scores both measure English proficiency for real-world applications, the two measurements' nature and grading scale differ. While the ELCSA score is cumulative based on language courses taken over time, the Linguaskill test is an English proficiency test. In terms of the grading scales, ELCSA is a score that ranges from 0.00 to 5.00, while the highest possible score obtained in the Linguaskill has a maximum score of 180, which is considered, as equivalent to a C1 and above grade on the CEFR. However, despite the varied strengths of the correlation, the data shows that a positive correlation exists, indicating that the variables move in the same direction.

Furthermore, this paper has shown that it is possible to use an established test that is CEFR aligned as a reference to determine the required scores that match a B2 level in a university English proficiency programme. For example, this study shows that a 3.25 score in the Writing component of the ELCSA corresponds to the B2 CEFR level. In comparison, a 3.5 overall score for ELCSA corresponds to the B2 CEFR level for overall English language proficiency. Thus, it could be assumed that a student who achieves a score of 3.25 in the ELCSA Writing component is at the B2 level of proficiency in terms of writing skills. Also, achieving a score of 3.5 in the ELCSA overall score would mean that a student is at B2 level for overall English language proficiency. This benchmarking is useful, as it can indicate a student's CEFR level using an internally developed university English language programme. Determining the score corresponding to the B2 CEFR level is also important as university students are expected to have a minimum B2 level of proficiency upon graduation. Notably, in so far as the students' performance is concerned, 72.6% of the STEM participants in the study managed to achieve the target that the Malaysian Ministry of Education had set by obtaining the minimum CEFR level of B2 for Malaysian university graduates. The other ten per cent of the participants had exceeded the target and managed to achieve C1, while only 17.3% achieved B1 and fell below the Ministry target.

Previous studies had mentioned that caution should be taken when aligning assessments using CEFR as it was implied that although the different tests use related criteria and are based on descriptors of the same however the perceived equivalence is only assumed (Foley, 2019). Additionally, it should also be considered that even though tests such as IELTS has been aligned to the CEFR, the alignment does not refer to the scores of specific language skill; instead,

it refers to the overall scores (Ali et al., 2018). Nonetheless, a study conducted by Universiti Malaysia Pahang attempted to contextualise the CEFR with their English Writing Language Proficiency Test. Their preliminary analysis has shown that the CEFR-A1 is sufficient in describing their lowest band (Band 1) and that the CEFR C2 and C1 would describe their highest bands, namely band 8 and 9. It was also mentioned that it was necessary to further describe the subcategories of the level of proficiency in order to address all of their bands as their English proficiency test had nine bands altogether (Ali et al., 2018). Therefore, it could be said that, despite being cautious of comparability aspects and over emphasis on standardisation, attempts for an alignment can be made possible. However, it is important to note that fundamentally, the CEFR was originally devised to assist the planning of curricula and that the common reference levels are for further facilitation (Foley, 2019).

CONCLUSION

In seeking to align the accumulative ELSCA scores with the Linguaskill CEFR scores, the authors conclude that there is a positive correlation between the ELSCA scores and the CEFR scores—which shows that there is a possibility in using performance in an English language proficiency programme to predict CEFR levels. Furthermore, this study has also shown that the ELCSA can be used with either the Linguaskill overall score or the writing score to predict and determine CEFR levels, especially to indicate whether or not the student has attained B2 in the CEFR as required by the Ministry of Education for university students upon graduation. Due to this alignment, it can be said that UPM is on the right track in benchmarking its language proficiency programmes with the CEFR. However, it is important to ensure the efficiency of their language programmes and make improvements where necessary.

It is suggested that for future research, attempts should be made to benchmark language programmes in different higher learning institutions to the CEFR. Given that the Linguaskill test is now accepted and adopted in the admission and exit requirements of universities in Malaysia as an alternative to MUET, IELTS, TOEFL and other tests, language centres should consider providing training for students to prepare for such tests or even become centres to carry out the tests. It could further enhance the curriculum of language centres and the practices of language instructors to be more CEFR-aligned. Consequently, this would increase the student's familiarisation with the CEFR and help them develop their language proficiency in line with the CEFR. In sum, this paper contributes knowledge that an alignment between a language proficiency programme of a Malaysian university and the CEFR does exist and that it is pertinent for other institutes to work in unanimity to benchmark their language proficiency programs towards the CEFR so that the level of standards of the English Language in Malaysian universities are acceptable and further credible.

REFERENCES

Abdullah, M. H., Wah, H. S., & Eng, W. B. (2015). Delivering the ELEx package: An English language experience approach for developing undergraduates' language proficiency. *Journal of Language and Communication, 2*(2), 283-296.

Ali, Z. M., Ali, F., Radzuan, N. M., Alwi, N. N. M., Abu, N. L., & Kassim, Z. (2018). Contextualising the CEFR: the Universiti Malaysia Pahang English language proficiency writing test. In *11th Annual International Conference of Education, Research and Innovation (ICERI 2018)* (pp. 4892-4902). IATED Publisher.

Alosani, M. S., Al-Dhaafri, H. S., & Yusoff, R. Z. (2016). Mechanism of benchmarking and its impact on organizational performance. *International Journal of Business and Management*, *11*(10), 172-183. https://doi.org/10.5539/ijbm.v11n10p172

American Educational Research Association. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Bogan, C. E., & English, M. J. (1994). *Benchmarking for best practices: Winning through innovative adaptation.* Mcgraw-Hill.

Bruce, E., & Hamp-Lyons, L. (2015). Opposing tensions of local and international standards for EAP writing programmes: Who are we assessing for? *Journal of English for Academic Purposes, 18*, 65-77. https://doi.org/10.1016/j.jeap.2015.03.003

Cambridge Assessment English. (2016). *Linguaskill.* Cambridge Assessment English. https://www.cambridgeenglish.org/Images/429191-linguaskill-listening-and-reading-trial-report.pdf

Cambridge Assessment English. (2018). *Meet the research team behind Linguaskill.* Cambridge Assessment English. https://www.cambridgeenglish.org/Images/521541-linguaskill-research-team.pdf

Cambridge Assessment English. (2019). *Linguaskill simplify your English language testing.* Cambridge Assessment English. https://www.cambridgeenglish.org/Images/466596-linguaskill-simplify-your-english-language-testing.pdf

Cambridge Assessment English. (2021a). *BULATS*. Cambridge Assessment English. https://www.cambridgeenglish.org/exams-and-tests/bulats/

Cambridge Assessment English. (2021b). *Fitness for purpose*. Cambridge Assessment English. https://www.cambridgeenglish.org/research-and-validation/fitness-for-purpose/#cefr-relationship

Cambridge Assessment English. (2021c). *How results are presented.* Cambridge Assessment English. https://www.cambridgeenglish.org/exams-and-tests/linguaskill/information-about-the-test/how-results-are-presented/

Cambridge Assessment English. (2021d). *Malaysia: Ministry of Higher Education recognises English language exams from Cambridge*. Cambridge Assessment English. https://www.cambridgeenglish.org/news/view/malaysia-ministry-of-higher-education-recognises-cambridge-english-exams/

Cambridge English Language Assessment. (2011). *Using the CEFR: Principles of Good Practice.* University of Cambridge. https://www.cambridgeenglish.org/images/126011-using-cefr-principles-of-good-practice.pdf

Canadian Language Benchmarks. (2012). *English as a second language for adults.* Centre for Canadian Language Benchmarks. https://www.canada.ca/

content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf

Centre for the Advancement of Language Competence. (2013). *ELEx Package (English language experience) for undergraduates* Centre for the Advancement of Language Competence. https://calc.upm.edu.my/dokumen/11203_ELEX_Briefing_BI.pdf

Cheung, H. (2012). Academic perspectives from Taiwan. In M. Byram, & L. Parmenter (Eds.), *The Common European Framework of Reference: The globalisation of language education policy* (pp. 224-230). Multilingual Matters.

Don, Z. M., & Abdullah, M. H. (2019, May 27). What the CEFR is and isn't. *Free Malaysia Today*. https://www.moe.gov.my/en/menumedia/printed-media/newspaper-clippings/what-the-cefr-is-and-isn-t-free-malaysia-today-27-mei-2019

Etiken, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1-4. https://doi.org/10.11648/j.ajtas.20160501.11

Fisher, J. G. (1996). *A managers guide to staff incentives and performance improvements*. Kogan Page Publication.

Foley, J. A. (2019). Issues on assessment using CEFR in the region. *Language Education and Acquisition Research Network Journal, 12*(2), 28-48.

Herman, J. L., & Osmundson, D. R. (2010). *Benchmark assessment for improved learning (AACC Policy Brief)*. University of California. https://doi.org/10.1037/e685212011-001

Inguva, M. (2018). Effective use of benchmarking: The context of the centre for preparatory studies in Oman. *Arab World English Journal*, 93-104. https://doi.org/10.24093/awej/MEC1.7

Jaafar, M. S. (2019). *Syarat kompetensi Bahasa Inggeris kepada pelajar antarabangsa* [English language competency requirements for international students]. Malaysian Qualifications Agency. https://www.mqa.gov.my/pv4/document/pekeliling/2019/JPT-Penambahbaikan%20Syarat%20Kompetensi%20BI%20Pelajar%20Antarabangsa%20IPTS%2009Dis.pdf

Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the common Chinese Framework of Reference for Languages: Challenges at macro and micropolitical levels. *Language Testing in Asia*, *7*(1), 1-19. https://doi.org/10.1186/s40468-017-0032-5

Keegan, R., & O'Kelly, E. (2012). *Applied benchmarking for competitiveness: A guide for SME owner/ managers*. Oaktree Press.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407-411. https://doi.org/10.4097/kjae.2017.70.4.407

Lim, G. S. (2017). *The Occupational English Test and IELTS: A Benchmarking Report*. Cambridge Assessment English.

Malaysian Qualifications Agency. (2020). *Kompilasi dasar jaminan kuasa pendidikan tinggi (2009-2020): Edisi* ketiga [Compilation of higher education power guarantee policies (2009-2020): 3rd ed.]. https://www.mqa.gov.my/pv4/document/compilationPolicy/KOMPILASI%20DASAR%20JAMINAN%20KUALITI%20PENDIDIKAN%20TINGGI%2020092020.pdf

Malaysian Qualifications Agency. (2020). *Malaysian qualification agency*. Programme Standards: Language. https://www2.mqa.gov.my/qad/v2/garispanduan/2020/PS%20Language%20-%20updates%20Oct%202020.pdf

Markel, S. (2018). The CEFR and English education in Japan. *Journal of Policy Studies*, *56*, 33-38.

Ministry of Education. (2015). *English language education reform in Malaysia: The Roadmap 2015-2025.* Ministry of Education Malaysia.

Murtaza, S. F., & Saleh, N. S. (2018). English language use in Malaysian government and private civil engineering workplaces. *International Journal of Education & Literacy Studies*, *6*(3), 84-91. https://doi.org/10.7575/aiac.ijels.v.6n.3p.84

Mustafa, Z. (2018, July 11). Proficient way to better English. *New Strait Times*. https://www.nst.com.my/education/2018/07/389562/proficient-way-better-english

New Strait Times. (2019, June 27). *We can make CEFR work for us*. *New Straits Times*. https://www.nst.com.my/opinion/letters/2019/06/499490/we-can-make-cefr-work-us

Nwabuko, L. O., Igwe, N. J., Okengwu, M. C., Nwabuko, M. A., & Ekere, O. (2020). Benchmark for partnership in human and material resources provision for adult education programmes in the South-East zone of Nigeria. *Global Journal of Health Science*, *12*(5), 46-56. https://doi.org/10.5539/gjhs.v12n5p46

Parlimen Malaysia. (2019). *Mesyuarat Kedua, Penggal Kedua Parlimen Keempat Belas 2019* [Second Meeting in the Second Term of the 14th Parliamentary]. Parlimen Malaysia. https://www.parlimen.gov.my/files/jindex/pdf/JDN22072019.pdf

Prakash, G. (2019, May 23). Education Ministry says using European benchmark for English-language option teachers. *Malay Mail*. https://www.malaymail.com/news/malaysia/2019/05/23/education-ministry-says-using-european-benchmark-for-english-language-optio/1755690

Read, J. (2019). The influence of the Common European framework of Reference (CEFR) in the Asia Pacific region. *LEARN Journal: Language Education and Acquisition Research Network Journal*, *12*(1), 12-18.

Rethinasamy, S., & Chuah, K. M. (2011). The Malaysia university English test and its use for placement purposes: A predictive validity study. *Electric Journal of Foreign Language Teaching*, *8*(2), 234-245. https://doi.org/10.2139/ssrn.2146007

Rockinson-Szapkiw, A. J. (2012). *Step 4: Selecting and justifying your research design.* http://amandaszapkiw.com/artifacts/resources/tutorials/research-process/Step-4-Selecting-and-Justifying-Your-Research-Design.pdf

Sani, R. (2018, February 7). *Building capacity to increase English Proficiency*. *New Strait Times*. https://www.nst.com.my/education/2018/02/332968/building-capacity-increase-english-

Sani, R. (2020, March 18). *Meeting university requirements*. *New Strait Times*. https://www.nst.com.my/education/2020/03/575837/meeting-university-requirements

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Sufi, M. K. A., & Stapa, M. (2020). Should the CEFR illustrative scales be localised to Malaysian higher education standards? – A conceptual paper. *International Journal of Academic Research in Business and Social Sciences, 10*(9), 885-897. https://doi.org/10.6007/IJARBSS/v10-i9/7882

Thomas, J. (2019, November 28). *Malaysia: Between education and skills*. *The Asean Post*. https://theaseanpost.com/article/malaysia-between-education-and-skills

UPM. (n.d.). *Specific Program Requirements for Bachelor Program*. Universiti Putra Malaysia.

https://fsktm.upm.edu.my/upload/dokumen/BGAKA1_Specific_Program_Requirements_(Bachelor).pdf

Wu, J. (2012). Policy perspectives from Taiwan. In M. Byram, & L. Parmenter (Eds.), *The Common European Framework of Reference: The globalisation of language education policy* (pp. 213-223). Multilingual Matters.

Xu, J., Brenchley, M., Jones, E., Pinnington, A., Benjamin, T., Knill, K., Seal-Coon, G., Robinson, M., & Geranpayeh, A. (2020). *Linguaskill Building a validity argument for the Speaking test.* Cambridge Assessment English. https://www.cambridgeenglish.org/Images/589637-linguaskill-building-a-validity-argument-for-the-speaking-test.pdf

Zairi, M. (1996). *Benchmarking for best practice: Continuous learning through sustainable.* Heinemann.

Zulkefli, A. S. (2017, November 4). M'sian students, teachers show vast improvement in CEFR as per Cambridge Evaluation Study 2017. *New Straits Times.* https://www.nst.com.my/news/nation/2017/11/299126/msian-students-teachers-show-vast-improvement-cefr-cambridge-evaluation

SOCIAL SCIENCES & HUMANITIES

PERTANIKA
JOURNALS

In-service Teachers' Familiarisation of the CEFR-aligned School-based Assessment in the Malaysian Secondary ESL Classroom

Charanjit Kaur Swaran Singh¹*, Harsharan Kaur Jaswan Singh¹, Dodi Mulyadi², Eng Tek Ong³, Tarsame Singh Masa Singh⁴, Nor Azmi Mostafa¹ and Melor Md Yunus⁵

¹*English Language and Literature Department, Faculty of Languages and Communication, Universiti Pendidikan Sultan Idris, 35900 Tanjung Malim, Perak, Malaysia*
²*English Education Department, Faculty of Foreign Language and Culture, Universitas Muhammadiyah Semarang, 50273 Semarang, Indonesia*
³*Department of Educational Studies, Faculty of Human Development, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia*
⁴*English Language Unit, Language Department, Institute of Teacher Education, Tuanku Bainun Campus, 14000 Bukit Mertajam, Pulau Pinang, Malaysia*
⁵*Faculty of Education, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*

ABSTRACT

The main purpose of this study is to investigate in-service teachers' familiarization of the CEFR-aligned school-based assessment (SBA) in the Malaysian secondary ESL classroom. It also intends to explore teachers' knowledge, understanding, and perceptions of the CEFR-aligned SBA. The study also examined the implementation of the SBA and the challenges that TESL teachers faced embracing the CEFR-aligned SBA in their ESL classroom. An exploratory mixed-method research design was employed. Data were collected by administering a survey to 108 in-service teachers, and 12 in-service teachers participated in the interview. The results show that the in-service teachers have rather a good level of familiarization with CEFR-aligned SBA and a moderate level of awareness and comprehension of the CEFR-aligned SBA. However, the in-service teachers are aware of the importance of CEFR-aligned SBA to assist students to improve their proficiency. In-service

teachers exhibit a good understanding of selecting the appropriate assessment tools and methods to assess students' learning. In-service teachers expressed their struggles and concerns regarding implementing CEFR-aligned SBA effectively, including lack of training, sourcing for good materials to teach, students' negative attitude towards the teaching and learning process, students' attendance, time constraint and their workload. In conclusion, the implementation of the CEFR-aligned SBA is crucial as it is a national agenda and teachers' involvement in executing the assessment is obligatory.

INTRODUCTION

Assessment is an inseparable part of teaching and learning, as it assists teachers in monitoring students' progress and the achievement of educational goals. Assessment has always been part of the education curriculum. Teachers can assess students' learning through a formative or summative manner (Box et al., 2015). Teachers can use formative assessment to focus ongoing development of the student's language. Formative assessment allows teachers to evaluate students in 'forming' their competencies and skills to assist them in monitoring performance (Singh et al., 2017). So, when a student shares a suggestion or makes mistakes, teachers must offer feedback to improve the student's language ability (Liu & Li, 2014). Summative assessment assists

teachers to summarize and measuring student attainment generally at the end of a course or unit of instruction. Both forms of assessment are important and necessary as they serve different purposes. Assessment helps teachers make decisions about curriculum, attainment of learning outcomes, grades, achievement, placement, instructional needs, and formation of skills and competencies of students. Teachers must incorporate assessment in the teaching and learning process as it can enhance or promote learning. Therefore, assessment must be formative and embedded with teaching. Brown and Abeywickrama (2010, p. 3) refer to assessment as an ongoing process encompassing many methodological techniques. These techniques include teachers' effort to appraise the students' response to a question and written work. Assessment is also defined as 'appraising or estimating the level or magnitude of some attribute of a person (Mousavi, 2009, p. 36). Hancock and Brooks-Brown (1994) opine assessment as an active process that enables the teacher and student to monitor the student's performance. Assessment has always been a concern in all educational institutions where one form of assessment is used. The question about the effectiveness of assessing student ability is of great concern.

School-based Assessment

In Malaysia, the entry and introduction of school-based assessment (SBA) is in line with the National Philosophy of Education, an ongoing effort toward developing the potentials of individuals

in a holistic and integrated manner to produce individuals who are intellectually, spiritually, emotionally, and physically balanced and harmonious. In line with current trends in assessment, SBA or PKBS (*Penilaian Kendalian Berasaskan Sekolah*) has been introduced into Malaysian schools under the New Integrated Curriculum for Secondary Schools. Now 'coursework' has been recommended for a few secondary school subjects. The Ministry of Education introduced the school-based oral assessment for both *Bahasa Malaysia* and English Language in 2003. It is a compulsory component for Secondary Five candidates taking the *Sijil Pelajaran Malaysia* (SPM) Examination. It gives all educational stakeholders the power to improve teaching and learning practices.

Inception of CEFR-Aligned SBA

The Malaysian Ministry of Education implemented *Kurikulum Standard Sekolah Rendah* (KSSR) or the Standard Curriculum for Primary Schools (SCPS) in 2011. The main purpose for introducing the curriculum was to set national standards and performance for all primary school level subjects, including ESL (Sidhu et al., 2018). A modular structure approach was introduced for the four language skills under the Standard Curriculum for Primary Schools (SCPS). In addition, phonics approaches for basic literacy, language arts and penmanship were introduced. Furthermore, importance was placed on critical and creative thinking skills (CCTS) specifically for incorporating and fostering

higher-order thinking skills (Ministry of Education, 2017). The Standard Curriculum for Primary Schools (SCPS) emphasised a learner-centred approach and focused on the 4Cs (communication, critical thinking, creativity, and collaboration) of traversal skills required for 21st-century learning. The Standard Curriculum for Primary Schools also focused on the e-assessment through the Information Communication Technology (ICT) tools. Teachers should not just focus on assessing students' skills and competencies, but students must be taught to exhibit cognitive operations at higher levels. The Malaysia Education Blueprint (2013-2025) recognises the importance of developing and applying 21st-century curriculum and assessment (Ministry of Education Malaysia, 2013). It aligns with the government's policy to enhance English Language mastery among teachers and students, exceeding the English Language curriculum benchmark internationally.

Consequently, this study investigates the in-service teachers' familiarisation of CEFR-aligned school-based assessment (SBA) in the Malaysian secondary ESL classroom. More specifically, it explored teachers' knowledge, understanding, and perceptions of the CEFR-aligned SBA. Therefore, this study will answer the following research questions: What is the in-service teachers' familiarisation and knowledge of CEFR-aligned SBA? What are in-service teachers' mastery of formative assessment? How is SBA implemented in the secondary ESL classroom? What are the challenges faced by the teachers in implementing SBA?

Past Studies on In-service teachers' Familiarisation of CEFR-aligned School-based Assessment. Uri and Aziz (2018) carried out a study on CEFR implementation in Malaysia based on the teachers' awareness and the challenges. Their study reported that the introduction and implementation of CEFR in Malaysia began with forming the English Language Standards and Quality Council (ELSQC) in 2013. The Council extended help to the English Language Teaching Center (ELTC) to support the Ministry of Education to uplift the English language proficiency of Malaysian students. The Council introduced the CEFR framework into the education system and developed a roadmap for systematic English language education reforms. The need to align CEFR to the education system was crucial in the Malaysia Education Blueprint as it aims at enhancing the standards to meet international benchmarks (Azman, 2016). However, a study conducted by Malakolunthu and Hoon (2010) on teachers' perspectives of school-based assessment in Kuala Lumpur revealed that they need a proper grading guideline and the implementation procedures; in other words, they still lacked the information on implementing formative assessment skills. In addition, teachers shared that they lack basic knowledge of school-based Oral English Assessment (OEA).

The roadmap, implemented in 2013, was anticipated for completion in 2025 with the hope to deliver the best language education beginning from pre-school up to tertiary education (Uri & Aziz, 2018). The findings of this study showed that teachers are familiar with CEFR and believe that implementing CEFR onto the Form 5 English syllabus and assessment can assist in upgrading students' English proficiency, thus enabling them to compete at a global level. Their findings also revealed that adopting the CEFR framework would solve the graduate employability issues in Malaysia. On the other hand, some teachers agreed that they have limited knowledge and exposure to the CEFR. Therefore, it may also slow down the CEFR implementation process in our educational context. Other related problems that surfaced with CEFR include teachers' English proficiency, teachers' cooperation, and willingness to learn and shortage of experts who can write and produce CEFR aligned textbooks, inadequate training and the mindset of teachers who believe that it is challenging and complicated to integrate CEFR in their instruction were reported in the study.

The CEFR-aligned SBA puts emphasis on both peer and self-assessment as one of the important components for developing autonomous language learners (Little, 2013). It is a holistic approach in which cognitive, affective, and psychomotor domains are equally assessed. Thus, it can be concluded that many teachers view CEFR-aligned SBA as a transformative approach to assessment practices (Sidhu et al., 2018). The CEFR-aligned ESL secondary school curriculum restructure has proposed an innovative assessment system in the education system. The formative SBA complements the summative assessment putting forward the

significance of learner autonomy to ensure enhanced language learning.

Past studies in second language assessment abound; these have provided data empirically to support research on formative SBA. Formative assessments are deemed effective in facilitating student learning provided they are implemented in problem-based learning and inquiry-based (Darling-Hammond, 2012; Grob et al., 2017; Weiss & Belland, 2016). Teachers and students must collaborate in the formative assessment process. It would then allow the teachers to understand and monitor students' level of achievement and knowledge. Only then can teachers use the information obtained from the students' mastery of knowledge to get information about their strengths and weaknesses to adjust teaching and learning, thereby enhancing the instructional value of assessment. Details regarding students' strengths and weaknesses can reveal weaknesses in teaching and provide useful information to improve teaching. It may also suggest that students have not mastered a particular unit or syllabus content that is being assessed. It could be due to the weaknesses in instruction and thus necessitates implementing more effective teaching strategies (Cizek, 2010). The combination of formative assessment and summative assessment are well-practiced in some schools and educational institutions. However, teachers still lack the confidence to implement formative assessment and summative assessments successfully due to their inability to carry out the assessment process successfully, complexities involved

or fear that this approach may disrupt the teaching and learning process.

SBA's main focus and initiative under the CEFR-aligned ESL curriculum restructure on implementing formative assessment in secondary schools. Teachers were given a variety of strategies for incorporation during the teaching process to collect evidence related to student learning and help learners improve mastery of learning. As a result, teachers were exposed to some training guiding them on implementing the formative assessment. However, SBA has been implemented in the Malaysian school context. Therefore, not much empirical evidence can be gathered or shared on implementing the CEFR-aligned SBA in Malaysian secondary ESL classrooms.

Therefore, this study aimed at investigating the in-service teachers' familiarisation of CEFR-aligned school-based assessment (SBA) in the Malaysian secondary ESL classroom. More specifically, it explored teachers' knowledge, understanding, and perceptions of the CEFR-aligned SBA. The study also examined the SBA implementation and the challenges TESL teachers faced in embracing the CEFR-aligned SBA in their ESL classroom.

METHOD

According to Creswell (2012), a research design is a blueprint known as the initial step in planning and organising the research process (Toledo-Pereyra, 2012) that regulate factors that might affect the validity of the finding. Therefore, an exploratory mixed-

method research design entailing two phases was employed (Creswell, 2012).

In-service teachers from twelve different schools participated in the study. The schools were selected randomly and located in Perak, Kuala Lumpur, Negeri Sembilan, Selangor, Kedah, Johor and Sarawak. The twelve schools were labelled as School 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12. Google survey approach was used in the study. Respondent confidentiality and anonymity are some of the advantages of a Google survey. In addition, such a survey can reach a larger number of respondents in a different location (Bourque & Fielder, 2003). Furthermore, the Google survey gives respondents flexibility as they can take their time to answer all the questions given. According to Punch (1994), respondents will give more honest responses, and the process avoids interviewer bias. Other advantages of the google survey are permitting quick and inexpensive data collection, as it only involves mailing expenses (Creswell, 2012), and this is the most economical form of data collection.

A group of individuals who have the same characteristics constitute a population (Creswell, 2012). The study population is selected from lower and upper secondary school's in-service ESL teachers in Malaysia. A total of 108 in-service teachers responded and were assigned numbers ranging from 1 to 108. The study is divided into two phases. In the first phase, which took a quantitative approach, the researcher administered a survey to 108 in-service teachers. In the study's second phase, which employed a qualitative approach, the researcher elicited in-service teachers' knowledge of CEFR-aligned school-based assessment. Therefore, samples that were selected need to be those who are experts in concern (Kruger & Stones, 1981) and who "understand the central phenomena" (Creswell, 2012, p. 206). The sample size selected was based on the study's judgement and purpose, as opined by Groenewald (2004). In this study, twelve in-service ESL teachers volunteered to be interviewed.

The survey employed in the study had two sections. The first section included respondent demographic background. Section B explored in-service teachers' familiarisation of CEFR based on a 4-point Likert Scale where a score of 1 reflected strong disagreement while a score of 4 indicated a firm agreement. The survey validity was checked by a panel of four experts— three TESL lecturers and one teacher who has been the master trainer for CEFR. The reliability of the survey was performed through a pilot study with 28 teachers from another district in Perak. The reliability of the survey was 0.954 based on the Cronbach alpha.

The researchers approached twelve in-service teachers from each school from the lower and upper levels. All the twelve teachers were interviewed. The interview was conducted to triangulate data gained from the survey instrument. Data obtained from the survey were analysed using descriptive statistics using the SPSS (version 20), and the interview data were analysed thematically.

RESULTS AND DISCUSSION

This section discusses findings from the survey. The survey data revealed in-service teachers' familiarisation of CEFR, in-service teachers' knowledge of CEFR-aligned SBA, goals of formative assessment and formative assessment strategies. In addition, the data obtained from interviews showed how in-service teachers implement SBA and the challenges that TESL teachers faced in embracing the CEFR-aligned SBA in their ESL classroom. The following Table 1 explains the in-service teachers' familiarisation of CEFR.

Table 1

Means and standard deviations of in-service teachers' familiarisation of CEFR, CEFR-aligned SBA and mastery of formative assessment

	Mean	Std. Deviation
In-service teachers' familiarisation of CEFR	48.9630	5.21666
In-service teachers' knowledge of CEFR-aligned SBA	23.4352	3.72733
Goals of Formative Assessment	14.1852	1.96297
Planning of formative assessment: Initial stage	17.4167	2.20503
Planning of formative assessment: Developmental stage	17.2130	2.22581
Planning of formative assessment: Closure	17.6667	2.18320
Formative assessment strategies	17.6389	2.38554

These are seven main constructs based on a 4-point Likert scale ranging from strongly disagree to strongly agree. The constructs were adopted from a manual on school-based assessment (SBA) prepared by the Malaysian Ministry of Education. The formative assessment has been divided into four subheadings: goals of formative assessment, planning of formative assessment: initial stage, planning of formative assessment: developmental stage and planning of formative assessment: closure. The findings shown in Table 1 reveal that the in-service teachers strongly agree and had rather a good familiarisation of CEFR ($M = 48.96$, $SD = 5.21$). However, in-service teachers' knowledge of CEFR-aligned SBA is moderate ($M = 23.43$, $SD = 3.72$), indicating that they lack awareness and comprehension of CEFR-aligned SBA. In terms of understanding the goals of formative assessment ($M = 14.18$, $SD = 1.96$), in-service teachers' mastery and understanding of the formative assessment goal is still at the infancy level. As for the planning of formative assessment, the initial stage ($M = 17.41$, $SD = 2.205$) indicates that teachers can plan activities to incorporate formative assessment at the beginning of the instruction. Planning of formative assessment: developmental stage ($M = 17.21$, $SD = 2.225$) shows teachers can plan the activities to be assessed at the developmental stage fairly. Planning of

formative assessment: closure (M = 17.66, SD = 2.183) showed teachers could plan and assess students throughout instruction. Teachers' ability to construct formative assessment strategies (M = 17.63 SD = 2.385) revealed that assessing student performance during teaching permits them to monitor student learning.

Implementation of SBA in the ESL classroom

A Range of Assessment Tools Employed. This section reports on the SBA implementation by the in-service teachers in the secondary ESL classroom. For the third research objective on in-service teachers' implementation of SBA in the ESL classroom, data were elicited from the interviews conducted with the teachers. It is essential to determine the types of assessment tools in-service used to evaluate the ESL students' performance. Therefore,

further analysis was carried out to investigate the types of assessment tools in-service teachers employ to carry out the formative assessment in the classroom. Based on the interviews conducted with the in-service teachers, various assessment tools were employed, including portfolio assessment, peer-assessment, presentation, exercises, worksheet, pair-work, role-play, authentic assessment, and exercises from the textbook (Table 2). The findings revealed that in-service teachers emphasise both formative assessment and summative assessment. Teachers employed the assessment tools to allow students to show their mastery of learning based on the topics taught. Teachers can activate formative assessment to monitor students' progress during the teaching and learning process. Students can only develop and build sound knowledge and fluency in English, which they can apply to survive in life outside the classroom.

Table 2

In-service teachers' use of assessment tools

Teacher	Types of assessment tools	Types of assessment methods	SBA related activities	Type of feedback
Teacher 1	Exercises, worksheets, listening module, role-play, dialogues, pair-work	-	Sourcing for materials based on topics given, two times a week	Marks, grading, written feedback
Teacher 2	exercises from textbooks, additional worksheets, pair work, presentation, writing exercises	Peer assessment, authentic assessment	Hardly any homework is given	Written feedback, verbal feedback,

Table 2 *(Continued)*

Teacher	Types of assessment tools	Types of assessment methods	SBA related activities	Type of feedback
Teacher 3	Workbook, other authentic materials including newspaper, presentation	-	Work is given in the class, no homework	Written feedback, verbal feedback, grades
Teacher 4	projects, mind-map, group discussion	Portfolio assessment	Work is given in the class, no homework	End unit test, written feedback and oral feedback, rubric (band 1–6)
Teacher 5	Class task, exercises	Peer assessment, self-assessment, video, brochure, diorama, essay writing, creating advertisement, writing song	Work is given in the class, no homework	Constructive feedback, rubric, written feedback
Teacher 6	Debate, activity books, worksheets, role-play, presentation	-	Homework	Verbal feedback, written feedback,
Teacher 7	Reflections, role-play,	Portfolio assessment, peer evaluation	giving students take home homework or extra worksheets, homework is given after every lesson.	Grading, marks, written feedback, grade them using an offline system
Teacher 8	Mind map	-	-	Oral feedback
Teacher 9	Exercise, worksheets	-	-	Oral feedback
Teacher 10	Discussion	-	rarely give homework	Oral feedback

Table 2 *(Continued)*

Teacher	Types of assessment tools	Types of assessment methods	SBA related activities	Type of feedback
Teacher 11	mind-map, exercises, worksheets, videos and PBL, exercises, worksheets a	-	-	Oral feedback, written feedback, grading
Teacher 12	Worksheets, mind-maps/ I-think maps, individual/ pair/ group presentations, exercises, reflection	Peer evaluation	-	Oral feedback, grading, star rating

Peer-assessment. Teachers have also provided written and oral feedback on students' performance in the class. Both formative assessment and summative assessment implementation in SBA is apparent based on the interviews with the teachers. Evidence on summative assessment implementation is apparent using worksheets and grading. Two teachers (Teacher 2 & 5) employed peer-assessment that is highly recommended by the CEFR-aligned ESL curriculum to encourage learner autonomy in the ESL classroom. Peer-assessment includes students providing judgments based on the work submitted by their peers. Peer assessment has been effective to assist the teachers to modify teacher assessment (Brown, 2004; Li, 2017; Liu & Li, 2014; Pope, 2001), on the other hand some scholars reject the notion of integrating peer assessment into formal assessment (Anderson, 1998; Cheng & Warren, 1999). The obtained findings concur with Li (2017) who carried out a study on 77 students involved in a peer assessment activity and reported that peer assessment can improve students' learning provided the students are given sufficient training. Matsuno (2017) also supports it, researching if peer assessment can be implemented employing FACET analysis. Findings showed that peer assessment is a practical approach and can be used as a supplementary assessment in class. One of the problems scholars faced using peer assessment is when the learners have to assess more than thirty peers, resulting in not assessing them thoroughly (Domingo et al., 2014). Teachers and scholars are doubtful in terms of the effectiveness of assessing students through peer assessment, but much research has proved that peer assessment is still beneficial in most of the educational contexts as it helps to promote student learning (Liu & Li, 2014; Pope, 2001) autonomy, motivation and responsibility (Brown, 2004; Pope, 2001).

The Adaption of CEFR-aligned SBA. Teacher 8 expressed that she has limited knowledge and exposure to CEFR-aligned SBA. Despite training and exposure given, Teacher 8 is still unclear what exactly CEFR is. She admitted that she is unfamiliar with the CEFR-aligned SBA. She further mentioned that the adaption of CEFR-aligned SBA for English Language Education is still not taken seriously among the language teacher in Malaysia. Teachers still lack understanding and have lots of confusion about the method and the framework of the CEFR- aligned SBA. However, she knows SBA and finds SBA as one of the effective efforts towards developing the proficiency level of the English language among the students. SBA ensures the integration of all four language skills, and her role in encouraging students to participate in the language activities can help strengthen their understanding. She also mentioned that it is very important to teach students to connect ideas and concepts when they learn to increase their confidence. Findings obtained from Teacher A are in line with findings reported by Uri and Aziz (2018), as most teachers have restricted knowledge and little information on CEFR. However, Uri and Aziz (2018) also reported that the teachers know the significance and the importance of the CEFR framework to help learners enhance English proficiency levels. Policy developers were optimistic about the implementation of the CEFR despite the obstacles and challenges faced.

Other factors that could impede CEFR implementation include teachers' attitude and resistance towards CEFR, negative perception and lack of training (Uri & Aziz, 2018).

Portfolio Assessment. Teacher 4 explained in his interview how he implemented portfolio assessment for his students. He shared that he used portfolio assessment as one of the assessment methods to assess them. He also mentioned that to implement and assess students using the portfolio assessment, teachers must adopt it. He continued sharing those portfolios will allow students to exhibit their work, progress, and achievement. The teacher shared two reasons for using portfolio assessment: the core element of the SBA-aligned curriculum emphasised both formative and summative evaluation. The teacher added that he usually instructs his students to create portfolios for a particular unit, not throughout the whole year. He limits monthly exams and replaces them with portfolio assessments. Each task and activity given to the students will be compiled in the portfolio, and students were asked to record the scores obtained. The teacher mentioned that portfolios show cumulative efforts and learning of a student over time. He also shared that portfolio assessment is valuable as it offers data about student improvement and skill mastery. Teacher 4 explained:

> Teacher 4: I will have my students create portfolios of their work for a particular unit...
>
> I will try not to do monthly exams and will replace them using portfolio assessment...
>
> Portfolios show the cumulative efforts and learning of a particular student….

According to Singh and Samad (2013), portfolio assessment is becoming significant as an assessment strategy that gives a holistic view of student performance. It is also viewed as an alternative to the shortcomings of the traditional form of examination. Paulson, Paulson, and Meyer (1991) stated that "portfolios offer a way of assessing learner learning that is different from the traditional methods. Portfolios allow the teachers to observe the students in a wider context which include students taking responsibility towards their own learning, taking risks, and developing creative alternatives to make judgments of their own performances."

Based on the interview conducted with Teacher 5, she shared her experience implementing SBA in her class. She explained that each student must complete at least one assessment for each unit taught throughout the year. All the units are from the English textbook. There are five unit plans that students must complete: People and Culture, Health, Social Issues, Environment and Science and Technology) to be taught in a year. Hence, upon teaching

the unit, the students must complete an assessment consisting of two tasks in that unit. The assessment can be in a video form, brochure, diorama, essay writing, creating an advertisement, writing song, and so on. This assessment is completed apart from worksheets and exercises given during the lesson. The teacher will give written feedback and grade their work according to the rubrics. The best assessment will also be displayed in their classroom or language room. Teachers are the leading players to ensure the assessment process is carried out appropriately in class. Teachers' skills, knowledge, commitment, and competency are the main elements to ensure success in any assessment planned for the students (Malakolunthu & Hoon, 2010; Pantiwati et al., 2017).

According to Torrance (1995), past studies have shown that teachers plan and execute assessment practices well to differentiate the assessment tools and methods deemed important for their students. Chapman and Snyder Jr (2000) and Stillman (2001) divulged that SBA is valuable and powerful for teaching, learning and assessment; teachers must be equipped with the appropriate skills, knowledge, competencies, and commitment to implement it successfully. Findings from Malakolunthu and Hoon (2010) revealed that teachers have limited knowledge, including content, learning outcomes, assessing students and some ideas to carry out the Oral English assessment activities. However, they reported teachers' inability

to assess the students accordingly because of improper guidelines prepared by the Ministry of Education. Therefore, teachers find it challenging to implement SBA (Malakolunthu & Hoon, 2010).

Findings from the teachers' interviews showed that teachers put in their efforts to implement peer assessment, portfolio assessment and self-assessment under SBA as directed by the CEFR-aligned ESL curriculum that would help to enhance learner autonomy. The portfolios assist the teacher to observe students' learning over a period based on the units assigned. These portfolios contained a variety of unit plans based on the textbook that students must complete. The use of portfolio assessment would also benefit teachers in improving their teaching practice, allowing them to see new directions and developments in instruction that would benefit their students (Knight, 2002; Mohtar, 2010).

CEFR-aligned SBA Activities for All the Language Skills. To further confirm on teachers' understanding of the CEFR-aligned SBA, the teachers were also asked during the interview session to share how they implemented activities for all the language skills. The findings obtained from the interview are reflected in Table 3. All the teachers agreed that they carried out the activities for all the four skills in an integrated manner. The teachers shared they usually plan reading and listening activity together. The teacher instructed

the students to listen while their friends are reading. Students must ask questions after each paragraph and at the same time they have to come up with higher order thinking skill question (Teacher 3). All the students must bring their textbook so that they can complete the listening tasks. As for Teacher 6, she used the audio from British Council websites to conduct the listening activity because it covers many topics. She also shared that she ensures the topics selected from British Council reflect the unit plans of the textbook. Teacher 6 preferred selecting materials from British Council because they are authentic, and she also get the reading materials from websites like National Geography. As for the writing skill, teacher 6 instructs students to write essay based on books or book review. Teacher 7 shared that she prefers to use lot of worksheets and grade the students using the offline system. She also gives feedback on the work submitted to her so that the students can improve. All the teachers expressed that they aware of the need to assess the students formatively so that they can acquire the competency levels, and this is supported by Ashraf and Zolfaghari (2018). The assessment stipulated in the English language syllabus is in line with the competence level based on CEFR descriptor. So, teachers must grade students' competency levels based on CEFR descriptors.

Table 3

SBA language activities conducted by teachers

Teacher/ Language skills	Listening skills	Speaking skills	Reading skills	Writing Skills
Teacher 1	listen to dialogues and answer the questions & listen to songs on YouTube that relates to the topic learnt	group discussion to express an opinion about general issue & pairing dialogues about one's routines	identify true false statements & match the words or phrases with the correct meaning.	Response to an email from a friend & write a guided composition with the note's expansion.
Teacher 2	Listen to conversations, advertisements, announcements	group discussion and talk about real life events	True/ False, identify title/ sub-title/main ideas	Rearrange paragraphs, guided essay (WH-questions)
Teacher 3	Listen to the audio, be interview, talk, song and answer questions on it	Based on topics, relate to past experiences, with good fluent proficiency	A text is given & while reading, asks questions after each paragraph, applying hots questions	Prepare mind map, write accordingly with Wh-questions, write based on experience with a good flow of grammar & lexical
Teacher 4	listen and sing songs with action	Group discussion, giving and sharing their opinions on a given topic, brainstorming and mind map to help them for the points they could speak about	process the information and ideas,	read and write short response
Teacher 5	listening to the speech, all the other friends will write feedback	impromptu speech	think and write their idea	Idea rush

Table 3 *(Continued)*

Teacher/Language skills	Listening skills	Speaking skills	Reading skills	Writing Skills
Teacher 6	use audio from the British Council websites to conduct listening activity because it covers many topics	impromptu speech, debate, and role-play interview	Authentic reading text from the British Council websites and text from websites like National Geography	essay and sometimes book or movie reviews.
Teacher 7	listen to songs, poems, and texts. Students then answer questions related to the listening audio.	Use role-play, group discussion and dialogues based on the topic	linear and non-linear text. Non-linear texts like table, mind map and graph.	Give the short answer and long answer response based on the topic and task given.
Teacher 8	-	Role-play	read and transfer information from non-linear to linear text and vice versa.	-
Teacher 9	students listen to the songs and sing to the lyrics (pronunciation)	share their personal experience related to failure in front of the class	sing while reading the lyrics. Discussion of new vocabulary and their meanings before singing the song	short responses about the songs and related issues found in the song
Teacher 10	listen and sing songs with action.	group discussion, talk about actual life events	Read and answer short-structured questions	read and write a short response

Table 3 *(Continued)*

Teacher/ Language skills	Listening skills	Speaking skills	Reading skills	Writing Skills
Teacher 11	listen to conversations and descriptions	presentation of PBL, projects	read a text and answer WH-questions and short structured questions	writing reports, letters, descriptions, story
Teacher 12	Questions in a textbook, listening activities from websites such as English teens, listen to songs and complete the lyrics by filling in the blanks	individual presentation on things they like to do or personal experience, events	Reading comprehension questions, short responses, guessing meanings	group writing, rearranging jumbled-up words to form a sentence, joining sentences, rearranging jumbled-up sentences to form a paragraph, using conjunctions and cohesive devices

Teacher 1 also mentioned that she usually shares the listening module for the listening skills so that the students can practice regularly at home at their own pace on weekends. As for the speaking skills, Teacher 1 instructs students to do role-play in groups and get the students to have dialogue in pairs to exchange ideas and talk about the topic.

> Teacher 1: For listening teachers will share the listening, module so that the students can have regular practice at home on weekends, and for speaking, the students will do role-play in groups and dialogue in pairs to talk about the topic discussed.

All the teachers agree that the purpose of integrating all four language skills is to assist the students in understanding meaning in a variety of familiar contexts. Students need to be exposed to deliver and communicate ideas; opinions based on familiar topics outlined in the unit plan. When students are exposed to reading activities, it allows expanding and exploring ideas for personal development. Teachers must prepare the activities to allow learners to appreciate and teach values and patriotism through language activities. All these aspects can be achieved through the tasks and activities planned for teaching and learning purposes. Only then can the curriculum develop the students to fulfil the requirements demanded by the workforce.

Assessment Tools. The data presented in Table 2 indicate teachers' ability to identify assessment tools and assessment methods that they can use to evaluate their students. Some of the assessment tools used include exercises, worksheets, role play, dialogues, pair work, mind map, discussion, debate presentation, reflections and class tasks in line with cross-curricular elements of the English language curriculum. In addition, teachers are aware of the assessment methods they can use to evaluate the students, namely portfolio assessment, authentic assessment, peer assessment and self-assessment. Finally, teachers can use different assessment methods to give learners a firm idea of the learning objective (Stiggins, 2005).

Challenges and Concerns Expressed by the In-service Teachers in Implementing CEFR-Aligned SBA. This section describes the challenges in-service teachers encounter during SBA implementation. Interview data shed light on some of the problems and challenges faced by the teachers in the implementation process. Some of the challenges and concerns are displayed in Table 4.

Table 4

Challenges and concerns of the teachers implementing CEFR-aligned SBA

Teacher/ concern	Time constraint	Students' negative attitude, poor attendance of students	The facilities, especially the audio for the listening activities	The understanding of SBA from the parents is too limited, parents' preference towards grades (traditional examination)
Teacher 1	✓	✓	✓	✓
Teacher 2	✓			
Teacher 3	✓	✓		
Teacher 4	✓	✓		
Teacher 5	✓	✓		
Teacher 6	✓			✓
Teacher 7		✓		✓
Teacher 8	✓			
Teacher 9		✓		
Teacher 10				
Teacher 11	✓	✓		
Teacher 12	✓			

Table 4 *(Continued)*

Teacher/concern	Student involvement in the class activities,	Teachers' workload	Lack of training	Availability of materials/ resources/ Access to Internet	Lack of confidence, motivation among teachers
Teacher 1					
Teacher 2	✓	✓	✓		
Teacher 3	✓	✓	✓	✓	
Teacher 4					
Teacher 5		✓			
Teacher 6			✓		✓
Teacher 7				✓	
Teacher 8		✓	✓		✓
Teacher 9					
Teacher 10					✓
Teacher 11					
Teacher 12					

Time Constraint. Table 4 clearly show the challenges, problems and concerns face by the teachers. Teachers know the importance of implementing the CEFR-aligned SBA for improving student proficiency but faced some constraints. Teachers 1 and 8 shared that time constraint impedes implementation of the CEFR-aligned SBA because she cannot complete the units stipulated in the syllabus. As for Teacher 2, he shared problems faced in terms of time constraints to carry out teaching and learning activities; students' involvement in the activities conducted, heavy workload that demotivates Teacher 2 to cover all the topics and lack of training and exposure to how CEFR-aligned SBA can be implemented successfully. Teacher 9 shares that her students are not

enthusiastic. Her students used to copy their friends' work and claimed they had attempted the tasks given.

Students' Attitude and Lack of Support from the Parents. Not only that, Teacher 1 mentioned that students' attitude towards SBA is negative as they feel SBA is not as important as the previous examination. As for Teacher 4 and 5, both divulged that students' attitude leads to negative opinions on school-based assessment. Students' poor attendance and low cooperation are the challenges faced by Teachers 4 and 5. When the teachers assign tasks, the students are reluctant to attend class, cooperate out of shyness and are not confident. Thus, it is very challenging to

assess such students, especially in the oral task. Besides that, SBA is time-consuming as it drains teachers' energy. Furthermore, she mentioned that parents do not support SBA as they lack understanding. Parents were very comfortable with the traditional examinations that give grades to students to measure student progress. Teacher 6 has attended various workshops and seminars based on CEFR and SBA, but she still lacks confidence in assessing the students by herself. It is because preparing the activities and tasks in the classroom takes much time and also because parents and students from Chinese schools do not understand the importance of SBA, so it is often not being emphasised. They are more concerned with high stakes examinations such as the SPM. As for Teacher 7, he teaches in a rural school. So, his students are very weak in English due to the lack of exposure. They also do not get much help from their parents, who are not so literate. They also do not have access to the Internet. Teacher 9 divulged that the student has a negative attitude towards the tasks she usually implements in the class. She also feels each worksheet might not cater to the individual's proficiency.

Availability of Resources to Implement CEFR-aligned SBA. One more problem was the facilities available, especially the audio availability for the listening activities. Teacher 1 said she could not conduct listening activities due to the unavailability of the audios needed for listening. Teacher 3 has a different view in terms of the challenges she faces. Teacher 3 mentioned

that she must use the materials required based on the student's ability. Most of the materials are extracted from workbooks, textbooks, and other relevant, authentic materials. However, the challenges are more as compared to the previous assessment. First and foremost, the textbook imposed by the curriculum development centre to use in classrooms does not reflect the students' ability. Teachers often refer to other simplified versions or better materials that suit the students' abilities. Teacher 3 believes that the textbook is a white elephant. Other than that, it is the time, the platform, facilities needed to implement the CEFR-aligned SBA, teaching workload, teachers who do not collaborate and share knowledge and students' negative attitude toward CEFR-aligned SBA.

Limited Knowledge to Implement CEFR-aligned SBA. According to Teacher 8, she is unprepared and not ready to implement SBA because she has a limited understanding of the rationale of implementing SBA. Teacher 8 shares that she also lacks confidence in conducting the assessment due to a lack of knowledge. The procedure of SBA implementation is remarkably complex as it involves much clerical work such as documentation, filing, and data entry. According to Teacher 10, he sometimes feels lost as this is a new evaluation system. Even though guidelines are given, not all can be applied 100% in the classroom setting. Teachers would usually adopt and adapt the best approach to get the desired results. Speaking lessons can be challenging as most

pupils are quite reluctant to participate in them.

The challenges and concerns expressed by the teachers in this study align with those in Darmi et al. (2017) as they showed teachers shared different views on CEFR; some of the teachers were uncertain how CEFR can assist in improving the proficiency courses, and some teachers disclosed positive attitude towards CEFR. It was reported that about 200 teachers in Malaysia agree that they are familiar with the CEFR concept (Uri & Aziz, 2018). However, this group of teachers also displayed a high level of anxiety and concern over CEFR implementation in Malaysia because they lacked information and were unsure of their roles in the changes (Don, 2015; Li, 2017; Omar & Sinnasamy, 2017; Lo, 2018). Overall, the in-service teachers faced some challenges and problems implementing CEFR to teach English; nevertheless, they also revealed an excellent familiarity with CEFR and moderate knowledge of CEFR-aligned SBA.

CONCLUSION

The main reason for introducing and implementing CEFR-aligned SBA was to facilitate and prepare the students to upgrade and improve their English proficiency to use and apply the language globally. The CEFR-aligned SBA aligns with government policy to ensure English language mastery among students and teachers and benchmark the English language curriculum. The findings of this study highlight the need for teachers to embrace assessment for learning and assessment as learning to complement the assessment of learning to ascertain the extent of student learning. Teachers also understood the requirement of the global world, which requires the students to have a good mastery of the English language that would enable them to function, and this could be realised through the adoption and reformation of the English curriculum and adoption of CEFR.

Furthermore, some teachers are aware of the integration of CEFR-aligned SBA. However, some also expressed their uncertainties of incorporating the CEFR-aligned SBA due to their inability to accept the new shift toward assessment for learning. Teachers' incompetence in understanding the revised CEFR-aligned SBA may contribute to hindering the smooth implementation of CEFR. Teachers' knowledge of the types of assessment tools to use for assessment can assist them in developing language skills among students. The finding also suggested that teachers provide oral and written feedback on students' work based on the CEFR-aligned SBA. Other factors that hinder smooth implementation of CEFR-aligned SBA include time constraints, teachers' workload, searching for simplified resources, lack of training and awareness that could hinder the whole process of implementing the CEFR-aligned SBA. Assessment of students should be ongoing to allow students to improve their performance. A new culture is now evolving, and the demand for education requires students' broad spectrum of competencies.

ACKNOWLEDGEMENT

REFERENCES

Anderson, R. S. (1998). Why talk about different ways to grade? The shift from traditional assessment to alternative assessment. *New directions for Teaching and Learning*, *1998*(74), 5-16. https://doi.org/10.1002/tl.7401

Ashraf, H., & Zolfaghari, S. (2018). EFL teachers' assessment literacy and their reflective teaching. *International Journal of Instruction, 11*(1), 425-436. https://doi.org/10.12973/iji.2018.11129a.

Azman, H. (2016). Implementation and challenges of English Language Education Reform in Malaysian primary schools. *3L: The Southeast Asian Journal of English Language Studies, 22*(3), 65-78. https://doi.org/10.17576/3L-2016-2203-05

Bourque, L., & Fielder, E. P. (2003). *How to conduct self-administered and mail surveys*. SAGE Publications Inc.

Box, C., Skoog, G., & Dabbs, J. M. (2015). A case study of teacher personal practice assessment theories and complexities of implementing formative assessment. *American Educational Research Journal, 52*(5), 956-983. https://doi.org/10.3102/0002831215587754.

Brown, D. (2004). Language assessment: Principles and classroom practice. Longman.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (Vol. 10). Pearson Education.

Chapman, D. W., & Snyder Jr, C. W. (2000). Can high stakes national testing improve instruction: Re-examining conventional wisdom. *International Journal of Educational Development, 20*(6), 457-474. https://doi.org/10.1016/S0738-0593(00)00020-1

Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education*, *24*(3), 301–304. https://doi.org/10.1080/0260293990240304

Cizek, G. (2010). An introduction to formative assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 3-17). Routledge.

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Pearson Education.

Darling-Hammond, L. (2012). Policy frameworks for new assessments. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 301-339). Springer. https://doi.org/10.1007/978-94-007-2324-5_6

Darmi, R., Saad, N. S. M., Abdullah, N., Behak, F. P., Zakaria, Z. A., & Adnan, J. N. I. (2017). Teachers' views on students' performance in English language proficiency courses via CEFR descriptors. *IJAEDU-International E-Journal of Advances in Education, 3*(8), 363-370. https://doi.org/10.18768/ijaedu.336688

Domingo, J., Martinez, H., Gomariz, S., & Gamiz, J. (2014). Some limits in peer assessment. *Journal of Technology and Science Education, 4,* 12-24. https://doi.org/10.3926/jotse.90

Don, Z. M. (2015). English language proficiency, graduate employability and the role of CEFR.

ASEAN Seminar 2015 on "Best Practices in English Teaching in ASEAN Universities". UM repository. http://repository.um.edu.my/98438/1/Asean%20Seminar%202015%20English%20language.pdf

Grob, R., Holmeier, M., & Labudde, P. (2017). Formative assessments to support students' competences in inquiry-based science education. *Interdisciplinary Journal of Problem-based Learning, 11*(2). https://doi.org/10.7771/1541-5015.1673

Hancock, C. R., & Brooks-Brown, S. (1994). *Teaching, testing, and assessing: Making the connection*. National Textbook Company.

Knight, P. T. (2002). *Being a teacher in higher education*. SRHE/OU Press.

Kruger, D., & Stones, C. R. (1981). *An introduction to phenomenological psychology*. Duquesne University Press.

Li, L. (2017). The role of anonymity in peer assessment. *Assessment & Evaluation in Higher Education, 42*, 645-656. https://doi.org/10.1080/02602938.2016.1174766

Little, D. (2013, October 29-30). *The Common European Framework of References for Languages: Purpose, origin, ethos and implications* [Paper presentation]. CEFR Conference: Towards Language Education Transformation in Malaysia. Putrajaya, Malaysia.

Liu, X., & Li, L. (2014). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education, 39*, 275-292. https://doi.org/10.1080/02602938.2013.823540

Lo, Y. Y. (2018). English teachers' concern on Common European Framework of Reference for Languages (CEFR): An application of CBAM.

Jurnal Kurikulum dan Pengajaran Asia Pasifik, 6(1), 46-58.

Malakolunthua, S., & Hoon, S. K. (2010). Teacher perspectives of school-based assessment in a secondary school in Kuala Lumpur. *Procedia Social and Behavioral Sciences, 9*, 1170-1176. https://doi.org/10.1016/j.sbspro.2010.12.302

Matsuno, S. (2017) Adoptability of peer assessment in ESL classroom. *Creative Education, 8*, 1292-1301. https://doi.org/10.4236/ce.2017.88091.

Ministry of Education Malaysia. (2013). *Malaysia Education Blueprint 2013-2025*. Kementerian Pendidikan Malaysia.

Ministry of Education. (2017). *English Language Curriculum Standard for Secondary School*. *Standard Document Curriculum and Assessment Form 3.* Kementerian Pendidikan Malaysia.

Mohtar, M. (2010). *The use of alternative assessment to sustain teaching and learning*. Penerbit Universiti Pendidikan Sultan Idris.

Mousavi, S. A. (2009). *An encyclopedic dictionary of language testing* (4th ed.). Rahnama Publication.

Omar, H. M., & Sinnasamy, P. (2017). Between the ideal and reality - Based oral English assessment. *The English Teacher, 38*, 13-30.

Pantiwati, Yuni, & Husamah, H. (2017) *Self and peer assessments in active learning model to increase metacognitive awareness and cognitive abilities. International Journal of Instruction, 10*(4), 45-57. https://doi.org/10.12973/iji.2017.10411a

Paulson, F. L., Paulson, P. R., & Meyer, C. (1991). What makes a portfolio a portfolio? *Educational Leadership*, *48*(5), 60-63.

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26*, 235-246. https://doi.org/10.1080/02602930120052396

Punch, M. (1994). Politics and ethics in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 83-97). Sage Publications.

Sidhu, G. K., Kaur, S., & Chi, L. J. (2018). CEFR-aligned school-based assessment in the Malaysian primary ESL classroom. *Indonesian Journal of Applied Linguistics, 8*, 452-463. https://doi.org/10.17509/ijal. v8i2.13311.

Singh, C. K. S., & Samad, A. A. (2013). Portfolio as an assessment tool and its implementation in Malaysian ESL classrooms: A study in two secondary schools. *Pertanika Journal of Social Sciences & Humanities, 21*(4), 1255-1273.

Singh, C. K. S., Lebar, O., Kepol, N., Rhaman, R. A., & Mukhtar, K. A. M. (2017). An observation of classroom assessment practices among lecturers in selected Malaysian higher learning institutions. *Malaysian Journal of Learning and Instruction, 14*(1), 23-61. https://doi.org/10.32890/mjli2017.14.1.2

Stiggins, R. J. (2005). *Student involved assessment for learning*. Merrill Prentice Hall.

Stillman, G. (2001). The impact of school-based assessment on the implementation of a modelling/applications-based curriculum: An Australian example. *Teaching Mathematics and Its Applications, 20*(3), 101-107. https://doi.org/10.1093/teamat/20.3.101

Toledo-Pereyra, L. H. (2012). Research design. *Journal of Investigative Surgery*, *25*(5), 279-280. https://doi.org/10.3109/08941939.2012.723954

Torrance, H. (1995). Teacher involvement in new approaches top assessment. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 44-56). Open University Press.

Uri, N. F. M., & Aziz, M. S. A. (2018). Implementation of CEFR in Malaysia: Teachers' awareness and the Challenges. *3L: The Southeast Asian Journal of English Language Studies, 24*(3), 168-183 http://doi.org/10.17576/3L-2018-2403-13

Weiss, D. M., & Belland, B. R. (2016). Transforming schools using project-based learning, performance assessment, and common core standards. *Interdisciplinary Journal of Problem-based Learning, 10*(2). https://doi.org/10.7771/1541-5015.1663

Teacher Readiness in Assessing Students for Malay Language Writing: An Exploratory Study

Rozita Radhiah Said*, Zuraini Jusoh, Azhar Md. Sabil and Shamsudin Othman

*Department of Language and Humanities Education, Faculty of Educational Studies,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*

ABSTRACT

This study explores the knowledge, understanding, and mastery of writing skills assessment among Malay language secondary school teachers in Malaysia. A total of 182 respondents from 91 secondary schools from seven different zones in Malaysia were selected using a purposive sampling technique. Survey design with a five-point Likert scale questionnaire instrument used in the study consisted of 117 items related to writing skills assessment. Statistical analysis is explained using standard deviation and mean score. The results of the study indicated that the determinants of the mastery level of writing skills assessment recorded the highest mean (M=3.92, SD=0.494). Then, it is followed by the second construct, which involves the implementation aspects of the evaluation was also rated highly (M=3.91, SD=0.482). The results also showed a significant and positive relationship between all respondents' knowledge and their understanding of writing assessment implementation and their mastery of writing skills assessment. The findings showed that the role of teachers as school-based appraisers is established and consistent with the guidelines outlined by the Ministry of Education Malaysia. Future research focusing on the implementation of writing skill assessment is suggested to ensure that the evaluation done is systematic and reliable.

Keywords: Knowledge, Malay language subject, secondary school, understanding, writing assessment

INTRODUCTION

Assessment of writing skills in the Malay language is a systematic process of collecting, interpreting, and responding to students' knowledge and experiences, which aims to understand the extent of students' knowledge, understanding, and abilities based on their learning. In Malaysia, the Secondary School Standards-based Curriculum (KSSM) for writing assessment standards of the Malay language was introduced in 2017, and it needs serious attention in investigating teachers' understanding of its implementation. In other words, ensuring that teachers have adequate knowledge and understanding of the new curriculum is crucial, especially since the new assessment system in KSSM gives teachers the autonomy to plan, administer, certify, and report student writing assessments. In general, student learning must be in line with what teachers are trying to assess. Teachers' understanding is directly proportional to the effectiveness of student learning.

However, society has limited information about how well teachers understand and master the assessment of writing skills. The issue mentioned above arose when the Curriculum and Assessment Standard Document was introduced to Malay language teachers in April 2016. It has been presented through courses and briefings conducted by the Ministry of Education Malaysia (MOE) to prepare for School-Based Assessment (SBA), such as the Secondary School Standards-based Curriculum (KSSM) guidelines in 2017.

In a short period of eight months, it is insufficient for the Malay language teachers to understand, research, appreciate, grasp its content, and fully master the writing skills used to assess the students. The short duration of the course affects teachers' knowledge related to its implementation, particularly on how the actual assessment is applied (Norazilawati et al., 2015).

Although several prior studies have been conducted on the assessment of writing abilities, such as Hashim (2009), Izam et al. (2012), Marohaini et.al (1997) and Marzni (2014)these studies have not been conducted in the context of current situations. Prior to 2017, writing skills were assessed centrally and fully by the Malaysian Examinations Board in public examinations held throughout the country. Assessment is limited to specific groups. However, beginning in 2017, the centralized public examination was phased out, and School-Based Assessment (SBA) was used at all lower secondary school levels in Malaysia. No matter how prepared or unprepared they are, all teachers are directly involved in assessment at school (Ministry of Education, 2017a, 2017b, 2017c). SBA empowers teachers and school administration with authority to plan, administer, issue certificates, and report on students' level of mastery of writing skills. Parents and the community now want to know how effectively the school, particularly the teachers, have acquired assessment knowledge after three years of implementation.

Adequate readiness and understanding of the evaluation method should be applied to teachers' prior preparation of the instrument, determining student's mastery level and interpreting their assessment scores (Lim et al., 2014). Curriculum change is a complex and challenging process that requires careful planning and sufficient time. Therefore, in the context of the recent developments in the assessment of the Malay language in Malaysian schools, this study aims to answer two main research questions. Firstly, this study attempts to determine the level of teacher's knowledge in regards to the performance standard for writing skills and implementation of writing skills assessment and the level of mastery in secondary school students' writing skills assessment. Secondly, to identify the relationship of all the three factors, knowledge, understanding of the appraisal, and the determining level of mastery in a writing skills assessment.

LITERATURE REVIEW

Changes in the educational curriculum will always occur to meet the dynamic demands of life. Malaysia, similar to other countries, also encountered internal and external issues and challenges due to the effects of globalization, liberalization, and the development of Information and Communication Technology (ICT). Thus, in Malaysia, the curriculum changed to a more holistic system focusing on School-Based Assessment (SBA). It enables the teachers throughout the year to monitor the learning and delivery of knowledge. Excellent examples of such assessment can

also be referred to in several countries, such as Singapore, South Korea, Japan, Finland, New Zealand, and Switzerland. They have taken earlier action in transforming their education system into 'school-based learning' to produce skilled students to compete globally. This type of learning experience involves teachers, parents, the community, the private sector, society, and friends (Ministry of Education, 2017d).

The implementation of School-Based Assessment (SBA) has been improved with the practice of Classroom Assessment (CA) for each subject, which is an alternative to the existing assessment and evaluation system. Through CA, teachers can track the effectiveness of the lesson and take action by replanning and modifying the lesson for the following teaching session. Teachers can also see the development of student learning as a whole because assessment occurs during daily activities in school and happens continuously (Heitink et al., 2016). Therefore, teachers will take subsequent action to improve the quality of pedagogy of the Malay language subject, especially in implementing the writing assessment. One compelling question is whether Malay language teachers are ready and have sufficient knowledge to make it successful. To what extent do teachers understand the meaning of SBA and CA, and how to implement them. Are Malay language teachers given adequate disclosure about how to evaluate writing skills? Do language teachers have the ability to assess student assignments. What are the matters, aspects, or skills to be assessed?

In the context of the challenges of implementing formative assessments, teachers were found to have low competencies and knowledge due to the limited duration of SBA courses conducted, which do not allow them to develop a clear understanding of how the assessment should be administered. It could be attributed to the lack of focus in terms of the SBA course content as the organizers only showed the teachers how to write Lesson Plans (LP) instead of guiding them on assessing students (Naim & Talib, 2014). What do the teachers need to master to make the writing assessment up to expectations? Knowledge related to the subject matter, understanding of assessment procedures, and teacher's mastery in determining level are closely related. A poor understanding of the assessment standards prevents teachers from performing accurate and appropriate assessments on the students' work. As a result, this situation is likely to raise uncertainty and create mistrust among parents and various parties towards the reporting of teacher assessment. Moreover, the quality of the evaluation may be questioned too. It is supported by the findings of a study conducted by Arsaythamby et al. (2015), who found that teachers face the challenge of doing self-assessment in the classroom, giving students an unfair and inaccurate score. However, writing skills are more structured and easier to assess because their scripts are readable, reviewed for verification, and can even be used as official records or documents for various purposes. The educational transformation of the examination centred on the assessment of the classroom is in line with the Malaysian Education Blueprint 2013-2025 (Ministry of Education, 2013), which reviews national exams and requires the percentage of questions to assess at least 50% of higher-order thinking skills (HOTS) in secondary schools by 2016.

In this regard, teachers need to equip themselves and always take a step forward by understanding the implementation of a system that mainly measures students' abilities. Teachers need to be knowledgeable, think critically, solve students' learning problems, have skills to access and analyze the information, and have effective oral and writing skills. Teachers should use versatile skills to teach new ideas, persuade others, record information, create imaginary worlds, express feelings, entertain others, heal psychological wounds, chronicle experiences, and explore the meaning of events and situations (Graham, 2019). For example, Singapore is implementing critical changes in terms of curriculum, pedagogy, assessment, theoretical-practice aspects of relationships, physical infrastructure to address concerns regarding skills that their teachers need to have (Lee, 2012),

" Addressing the concern that teachers themselves need to have the 21st centuryskills to teach those skills, the first of two pedagogical shifts is to increase emphasis on self-directed, inquiry-based, real-world learning ".

The practice of assessment among incompetent teachers will result in the teacher failing to recognize the student's true potential. Teachers are incapable of performing writing evaluations due to the lack of knowledge, understanding, expertise, and skills to assess students. As a result, teachers fail to monitor students' learning progress due to their lack of knowledge. This will have a negative impact on the mastery of students' communication skills in the future. Hence, the implementation of the assessment should be understood so that the assessment's accountability will not fail and can be fully implemented by the teacher (Naim & Talib, 2014).

In the context of writing assessment, writing assessment is an important language skill to improve language proficiency among secondary school students. All parties, especially the Malay language teacher, should master the techniques of teaching writing assessments effectively and strengthen the students' writing proficiency. Malaysia Education Development Plan (PPPM), in their 2013-2025 Blueprint highlights 11 Shifts and the Second Shift, focuses on ensuring that students need to master the skills in Malay language and English, and they are also encouraged to learn an additional language. The emphasis proves that the communication aspect of language among students is essential in the education system. As a result, teachers are not ready to conduct writing assessments. They have yet to fully master the assessment standards in the new document (KSSM), the Curriculum and Assessment Standard Document (DSKP), which is developed by the Curriculum Development Division, Minister of Education (MOE). The renewal of this document is in line with the new KSSM curriculum, which was introduced in 2017. Quantitative studies conducted by Arsaythamby et al. (2015) and Mei (2010) found that teachers face the challenge of scoring assessments in the classroom to the extent of giving students unfair and invalid scores. Teachers are found to be less prepared. Their knowledge of assessment is still at a low level as they are unclear on how to conduct an evaluation involving an assessment instrument.

In conclusion, Malay language teachers need to consider many factors in conducting SBA. Even though SBA is very good as a holistic instrument to assess students' efficiency and achievement, the issues raised regarding it may affect the implementation of SBA in school subjects in general, and the Malay language subject in particular. According to international study data, education based on the 'one-size-fits-all' model needs to provide different sets of interventions to suit different levels of school and student performance. Studies in Australia, New Zealand, and the United Kingdom have proven the advantages of SBA as an alternative in the current education system that measures students' academic achievement based on cognitive, affective, and psychomotor aspects (Ministry of Education, 2016c).

Past Studies on the Assessment of Writing Skills

An earlier study examined issues regarding the assessment of writing abilities from a variety of language scholars' perspectives. It is believed that this exposure would aid researchers in developing a knowledge of the subject and subsequently serve as a reference for assessing related concerns. In the Malay language education system, the essay is assessed by an inspector or by individual teachers who are tasked with the responsibility of checking students' work (Izam et al., 2012). Students' transcript essays will be reviewed first for the purpose of grading or marking on critical criteria, which will be evaluated using letters or numbers as deemed appropriate. After reading the student's transcript essay, the examiner will decide the student's score or will swiftly grade the student's essay. The purpose of scoring or grading essays that are completed quickly by reading is to obtain an overall result depending on the system defined previously. Assessment concerns in general, and assessment of essay writing in particular, remain a priority in the educational system. Alternative solutions are always sought in order to solve issues successfully.

According to Hashim (2009) and Malik et al. (2006), writing competency requires students to understand linguistic norms and structures in order for them to be utilized appropriately and comprehended accurately within a quantifiable context. To determine a person's level of competency, a rubric must be created that specifies precisely what is to be measured. For instance, in the lower secondary level in Malaysia, students' writing proficiency is geared toward meeting personal requirements in the areas of career education and daily tasks (Ministry of Education, 2016b). These two scholars were found to disagree when composing skills were assessed in SBA only using an impressionistic or holistic approach.

Hymes (1972 as cited in Pride, 2019) proposed that the writing skills assessment approach is aligned with the communicative idea which asserts that essays are language behaviors that serve as indicators of a person's linguistic fluency and should be evaluated directly using sociolinguistic, discourse, and grammatical proficiency criteria. In general, school-based grading systems lack defined communication criteria and ignore critical features of the language, most notably cohesiveness and coherence. This is confirmed by the findings of a study performed by many language scholars (Huzaina, 2007; Izam et al., 2012; Normah, 2006; Rohaya & Najib, 2008), who concluded that every instructor must grasp the writing talent. This immediately assists instructors in comprehending and developing a more equitable scoring scheme with a high degree of dependability. Not only should robust assessment tools conform to international best practices, but they should also be accurate and efficient for use in school-based assessment.

Marzni (2014) discovered that teachers were assessing students based on the end result rather than the complete writing process learned by students. This has

resulted in teachers being deeply involved in the test content throughout the writing assessment. The consequence is that teachers are more likely to assess students' factual knowledge alone, rather than their mastery of essay writing abilities or specific parts of the writing procedure. As such, when a teacher selects an assessment method, the teacher should first define the learning outcomes to be assessed and then match them to the appropriate assessment method (Izam et al. 2012), A language teacher should possess a range of knowledge, expertise, and skills that enable them to motivate students by raising awareness of the assessment process's benefits, advising on appropriate building materials, and leading the assessment process, thereby guiding students to comprehend the significance of coding results (Chan & Sidhu, 2012). Mutalib and Jamil (2012) corroborate this conclusion by claiming that incompetent teachers' assessment practices result in teachers failing to monitor student progress owing to a lack of expertise and an inability to generate fair results. Inadequate logic, planning, and failure to uncover the student's potential.

The Concept of Assessment for Learning in the Classroom

Formal or formative assessments have been practised for a long time in most countries around the world. The implementation, and the application of Assessment for Learning (AfL), involves continuous learning activities practised almost daily by students, peers, and teachers. It aims at gaining a holistic view of learning through various activities, such as dialogues, demonstrations, and observations (Graham, 2019). Formative assessment is defined as the integration of assessment processes into classroom learning. Assessment is a progressive process, and it occurs with the combination of four elements: teacher, student, evaluation, and context (Mandinach & Jackson, 2012). This statement is similar to the perspective of Pedder and James (2012), who asserted that the role of teachers and students is an essential aspect of teaching and assessment. They further added that assessment is carried out by collecting students' learning evidence, such as assessment tools and processes.

Based on the formative assessments of writing skills in Malaysia, evidence is collected via various assessment tools such as observation, questionnaire, written test, presentation, project, product, practical excursion, worksheet, writing, quiz, checklist, homework, peer review, daily work, scrapbooks, demonstrations, holistic 'rating' scales, portfolios, discussions, and simulations (Malaysian Examination Board, 2012; Board of Examination, 2017; Ministry of Education, 2017b). Assessment purpose is closely linked to its context, and that context needs to be considered in assessing assessment. The context involves internal authorities such as school administrators and external parties such as education policies. In brief, to ensure effective implementation of AfL in schools, planning, regulation, understanding of the content of focused language skills, an appropriate level of

thinking and coding needs to be developed. Since all of these elements are interrelated, a balanced proportion of all elements must be accentuated during the assessment (Ministry of Education, 2016a).

The definition of writing skills assessment in the context of Malaysian education needs to be proportionate to the guidelines that have been outlined in the KSSM. In other words, the assessment of writing skills being practised in Malaysia should obtain information from students based on what students know, ability, and practice. Through this process, teachers play a role in making professional decisions about student performance, which is the ultimate product of an educational programme. Formative assessments throughout the year must have clear goals, and teachers should design them according to the Student Learning Development Guide (PPPM), which contains official statements, bands, and descriptors for each subject so that learning can be implemented efficiently and effectively.

MATERIALS AND METHODS

Research Design

The researchers conducted an exploratory research to assess writing skills among teachers to gain an understanding of underlying reasons, opinions, and motivations. The findings can provide insights into the problem or help to develop ideas or hypotheses for potential quantitative research. A survey design using a questionnaire was utilized as the primary research instrument. Surveys are a tool that researchers often use in obtaining research data. The advantage of using the census as a research instrument is the uniformity factor in the direction of the question, and the same query used to be answered by all study participants.

Respondents

The need for a representative statistical sample in empirical research has created the demand for an effective method of determining sample size. To address the existing gap, this study used Krejcie and Morgan's (1970) table for determining sample size for a given population for easy reference. A total of 182 Malay language teachers with more than five years of teaching experience from 91 secondary schools across Malaysia participated in this study. Fifteen secondary schools in seven zones in Malaysia were selected using purposive sampling. Two teachers represented each secondary school. The seven zones are the southern zone, central zone, western zone, northern zone, Sabah zone, Sarawak zone, and Federal Territory zone.

Instrument of the Study

Table 1 shows the description of research instrument. There are four parts of the questionnaire in this study. Part A covers individual profiles of the respondents.

Meanwhile, the statements, as well as the construction of all the questionnaire items in Parts B, C, and D of this study, were adapted from the units contained in the assessment of DSKP, KPM, 21st

Century Skills assessment: PPPM 2013-2025, Secondary School Syllabus and Lesson Syllabus KBSM (2004), KSSM, and Curriculum Development Center (2017).

The questionnaire consisted of 117 items using a five-point Likert scale with 1= strongly disagree to 5 = strongly agree.

Table 1

Description of Research Instrument

Section	Component/ Construct	No of items	Statistics	Sources of reference
Part B	The Aspect of Knowledge in KSSM Document Writing Skill	20 items	frequency, percentage, standard deviation and mean score based on Pallant (2013)	Ministry of Education (2016b) Ministry of Education (2017a)
Part C	The Aspect of Understanding of Implementation of Writing Skills Assessment	53 items		Ministry of Education (2017b) and Ministry of Education, (2017c)
Part D	The Aspect of Determining Levels of Mastery in Writing Skills Assessment	44 items	Pearson's correlation Product-Moment	Ministry of Education (2017d)

Instrument Validity

Validity and reliability are two important factors to consider when developing and testing any instrument (e.g., questionnaire) for a study. In this study, validity refers to the degree to which an instrument accurately measures what it intends to measure by appointing two subject matter experts to review the instrument development and to assess content validity. Then, a construct validity test and re-test were implemented to determine which measurement method accurately represents the construct. Reliability refers to the concept of consistency and stability of an instrument constructed (Chin et al., 2019). Consistency refers to the high reliability of the instruments. Accordingly, to test the reliability of the study instrument, Cronbach's Alpha coefficient index is adequate. It is sufficient to test an instrument whereby a Cronbach's Alpha value is approaching a value of 1.00, indicating high reliability as shown in Table 2 (Cronbach, 1990).

Table 2

Cronbach's Alpha coefficient

Value	Levels
0.0-0.2	Low (modify items)
0.2-0.8	Moderate (modify some items)
0.8-1.0	High (items are acceptable)

Table 3 shows the overall reliability analysis in Cronbach's Alpha value index of the instrument based on a pilot study conducted by the researchers on 30 Malay language teachers at various schools in Petaling Jaya. Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. It is considered to be a measure of scale reliability.

Table 3

Reliability of instruments

Section	Component/ Construct	No of items	Alpha Cronbach
Part B	The Aspect of Knowledge in KSSM Document Writing Skill	20 items	0.928
Part C	The Aspect of Understanding of Implementation of Writing Skills Assessment	53 items	0.810
Part D	The Aspect of Determining Levels of Mastery in Writing Skills Assessment	44 items	0.940

The knowledge aspect of this questionnaire refers to teachers' knowledge of the philosophy of Malay language education, national curriculum definition, general and specific objectives of teaching and learning of writing skills in KSSM, teachers' knowledge of content assessment standards of writing skills, learning standards and performance standards in the DSKP, KSSM 2017. Meanwhile, the aspect of teachers' understanding of the implementation of assessment skills in writing refers to assessment methods, planning, implementation principles, administration and records, item development, script inspection, focus and theme assessment, and including materials and domain references. Finally, the aspect of determining the level of mastery of writing skills refers to teachers'

understanding of the determination of mastery levels in hierarchical order, six forms of performance standards, observation methods, observation methods, project work methods, or otherwise. Determination of the mastery level also refers to seven content standards for writing skills or teachers' determining student mastery level in professional judgment and determined by the teacher's discretion. This aspect also investigates whether the determination of student writing levels is measured based on all of the skills outlined in KSSM.

Analysis

This quantitative study employed a descriptive survey method to describe and summarise all information obtained from the study participants. Pearson correlation was performed to identify the relationship between the two variables' data in this study. Mean and standard deviation were used to measure the level at which the analysed data could meet the study's objective. The use of statistics in research is vital to explain the characteristics of a study population. According to Birenbaum et al. (2015) and Black (2015), descriptive studies are often aimed at providing a systematic explanation of the facts and characteristics of a population factually and accurately. Descriptive studies are useful when researchers collect data on an issue or phenomenon that cannot be directly observed.

RESULTS

The present study sought to discover how effectively the Malay language teachers implement writing assessment in terms of knowledge, understanding, and determination of students' level of mastery. The study is anticipated to provide an overview of the application of writing skills assessment in the classroom. The teacher's skills, knowledge, and actions are required to meet the criteria of the secondary school Malay language curriculum. It must be assessed in fulfilling the needs of the standard Malay language writing skills.

Table 4

The Level of Teacher's Knowledge for Assessment Implementation in Secondary School Writing Skills based on the Standard Secondary School Level of Writing Skills Performance

Questions	Construct	Mean	SD	Level
Part B	The Aspect of Knowledge of Writing Skills in KSSM Documents	3.857	0.434	High
Part C	The Aspect of Understanding the Implementation of Writing Skills	3.912	0.482	High
Part D	The Aspect of Determining Levels of Mastery in Writing Skills Assessment	3.923	0.494	High

As shown in Table 4, the results of the study indicated that the determinants of the mastery level of writing skills assessment recorded the highest mean M = 3.92 with SD = 0.494. Among the five items that showed the highest mean in this construct (Part D) is that the teachers understand and comprehend benchmarks arranged in hierarchically M = 4.83/ SD = 0.420. Then, it is followed by the teachers' understanding that there are two forms of performance standards which are general performance standards for each language skill and general performance standards across language skills with M = 4.80/ SD = 0.415. Other items and their mean and standard deviations are: Weak students need to be given guidance, and they need to be given reinforcement M = 4.76/ SD = 0.425. The teachers are also skilled in determining the overall mastery level of language skills by professional judgement according to the teachers' ability, M = 4.73/ SP = 0.486 through various methods M = 4.73/ SD = 0.486.

The second construct, which involves the implementation aspects of the evaluation was also rated highly (M =3.91, SD = 0.482). Among the five items that showed the highest mean in the construct (Part C) was an understanding of the implementation procedure of writing skills assessment in a planned but flexible manner M = 4.75/ SD = 0.463, according to the suitability and readiness of students M = 4.64/ SD = 0.532, basic principles inclusive, authentic and place principle (localized) M = 4.63/ SD = 0.534. The implementation also aims to develop students' potential and respond to KSSM's desire to produce students who master the skills of the 21st century at M = 0.4.63 / SD = 0.534.

The findings showed that the aspect of knowledge (Part B) is the third highest rated (M = 3.85, SD = 0.434). Among the five items that recorded the highest mean is the teacher knows KSSM that student is assessed continuously in PdPc, M = 4.83 /SD = 0.420, teachers also know all ten common objectives for KSSM with M = 4.83 /SD = 0.420. Teaching and learning strategies are clearly defined in the DSKP for teacher guidance M = 4.83/SD = 0.420. Similarly, the cross-curricular elements (EMK) are clearly shown in the DSKP for teacher guidance, M = 4.76/SD = 0.425.

Table 5 shows the results of Pearson product-moment correlation analysis which were performed to identify the relationships between the three aspects above. This method was chosen as suggested by Pallant (2013), who argues that Pearson product-moment can be used to observe the strength and direction of the relationship between the study variables.

Based on Table 5, there is a significant and positive relationship between respondents' content knowledge of writing skills (BB1) and understanding of the implementation of writing skills assessment (BC) among the students [r=0.710, n=182, p<0.000]. As the respondents' knowledge of the implementation in writing skills assessment increases, the respondents' understanding of the implementation in writing skills assessment increases too.

Table 5

The relationship between the content knowledge of writing skills, understanding the implementation of writing skill assessment, and determining the mastery level in the writing skills assessment

		BB1	BC	BD	BE
BB1	Pearson Correlation		0.710	0.710	0.022
	Sig. (2-tailed)		0.000	0.000	0.764
	N	182	182	182	182
BC	Pearson Correlation	710		0.814	0.076
	Sig. (2-tailed)	0.000		0.000	0.311
	N	182	182	182	182
BD	Pearson Correlation	710	0.814		0.059
	Sig. (2-tailed)	0.000	0.000		0.425
	N	182	182	182	182

BB1: Content knowledge of writing skills,

BC: Understanding of the implementation of writing skills assessment,

BD: Determining level of mastery in a writing skills assessment.

Significant $p<0.0.5$.

In addition, the findings also showed that there is a significant and positive relationship between knowledge (BB1) and the determinants of the mastery level in writing (BD) assessment [r=0.710, n=182, p<0.000]. As the respondents' knowledge of the implementation in writing skills assessment increases, the aspect of determining the mastery level of writing skills also increases. Similarly, the element of understanding (BC) and the determinants of the level of proficiency of the assessment of writing skills among outstanding teachers (BD) indicated a significant and positive relationship [r=0.814, n=182, p<0.000]. This means that if teachers' understanding in literacy skills increase, the aspects in the mastery level of teaching literacy skills also increase.

It can be seen that all of the three variables showed a significant and positive relationship based on the Pearson product-moment coefficient correlation values and the significant values set for social science studies. Overall, it can be concluded that there is a meaningful relationship between knowledge, understanding, and determination of mastery in the writing skills assessment. The relationship that exists between variables is a strong positive relationship. However, for teacher self-reflection in evaluating writing skills, there is no significant relationship between the knowledge, understanding, and determination of teachers' mastery level over the evaluation of writing skills.

DISCUSSION

Writing is a highly complex skill as learning how to write requires time and good instruction and to assess it (Coker et al., 2016; Vahapassi, 1988). Teachers conducting a classroom assessment is an essential factor in determining the success of educational reform (Norazilawati et al., 2015). With educational innovations based on school-based assessment practices, incompetent teachers accessing student's writing are most likely to affect students' learning performance. Different and inconsistent scoring or bias will result in insignificant scoring differences among 45 (Troia & Graham, 2017). The implementation of formative assessment is also considered to be complicated (Vingsle, 2014). Consequently, teachers will make a fair and accurate assessment result (Ministry of Education, 2014).

In light of the transformation, there are 11 shifts mentioned in the PPPM (2013-2025) that need to be implemented to bring about the change that everyone wants. The Government has established that quality is an essential element in achieving all the changes. Hence, quality must be given priority during its implementation. In the present study, the researcher focuses on the second shift, which is to ensure that each student has mastered the skills of the Malay language and is encouraged to learn additional languages. The second revision in chapter 4 of the PPPM (2013-2025) focuses on curriculum and assessment and mastery of language skills for the educational needs of students towards improving school performance. The students' reading, writing, and speaking skills in the Malay language are vital to produce a competent generation that is capable of speaking the language fluently in the future. PPPM (2013-2025) also states that the Malay language results show the highest passing rates in the public examinations, specifically UPSR, PMR, and SPM. Therefore, to strengthen the Malay language proficiency among students, the ministry has taken reasonable steps by introducing a standardized secondary level Malay language curriculum that covers all Malaysian public schools.

Chapter 4 of the PPPM (2013-2025) focusing on Student Learning, explains the curriculum developed for schools, which are the Written Curriculum, the Teaching Curriculum, and the Assessed Curriculum. The curriculum includes the knowledge gained, the skills developed, and the values instilled in the students. Written Curriculum refers to anything written in the curriculum outlined by KPM, which is the knowledge, skills, and values that make up the content of the curriculum that teachers need to teach. The ministry develops its aims to achieve international standards by using benchmarks from the higher education system. In this regard, using the teacher-referenced KSSM standard document, the DSKP containing Content Standards, Performance Standards, and Learning Standards, is the approach adopted by teachers to make it attainable and student-centred. Meanwhile, the Curriculum Assessment involved the knowledge, skills, and benefits of the students being assessed. These three dimensions are interrelated

with each other in classroom practice. To put it differently, Ishak (2011) stated that assessment is critical in determining how well a student understands the concept of learning and skills taught by a teacher. In the present study, the implementation of writing assessment by the teacher should include all three dimensions mentioned above.

The move towards assessment-based evaluation is part of the effort taken by the KPM to produce well-balanced students in the aspects outlined in the PPPM (2013-2025) and aligned with the country's National Education Philosophy. The basis of curriculum development aligned to the aspects that include the six student aspirations, 21st-century skills, national harmony, and communication skills. The transformation from exam-oriented evaluation to assessment-based evaluation that is happening in Malaysia is not isolated. Still, it is an effort to stay at par with others as the whole world of education changes towards a better education system in general. KSSM supports every aspect of equipping students with 21st-century learning skills as the new curriculum focuses on the importance of acquiring higher-order thinking skills (HOTS) among students. In Malaysia, the sudden shift from KBSM to KSSM concerns the Malay language teachers' readiness and ability to master the new assessment system. On top of that, it is believed that it takes a much longer time for teachers to make KSSM successful.

Paramasivam and Ratnavadivel (2018) mentioned that the lack of time for training before implementing the new curriculum affects the quality and teachers' understanding. Nevertheless, despite the concerns, the present study has revealed that the knowledge, understanding, and determining the level of mastery over the writing skills assessment based on band scoring as outlined in the Curriculum and Assessment Standard Document (DSKP), among the Malay language teachers are at a high level. This indicates that the Malay language teachers in Malaysia are showing positive signs of adapting to the change. In much the same way, Sabbir (2019) investigated the difference in English language assessment upon the introduction of KSSM, found that teachers gave positive feedback about the new assessment system. Apart from that, the present study's findings have also enlightened the ambiguity expressed in a previous study conducted by Jamil et al. (2017). Even though the researchers were uncertain of the success of the Standard-Based Secondary Curriculum (KSSM) as it was just implemented in that particular year, they hoped that the new curriculum would cause neither confusion nor hesitation among the Malay language teachers for its implementation in schools.

The findings of this study are consistent with previous studies about the pertinent role of teachers as assessors in assessment-based evaluation (Sabbir, 2019; Chin et al., 2019). The relationship among the aspects of teachers' knowledge, understanding, and mastery of the writing skills assessment exists. All the aspects are interconnected, signifying the need for teachers to acquire all three aspects to improve students'

performance. The mandate handed over to teachers at school to assess their own students' performance must be carried out fairly, wisely, and accurately. Thus, teachers need to equip themselves with adequate knowledge about the new assessment system, keep themselves revised and update any changes in the assessment in the future. This suggestion is in line with Rozita et al. (2019), who conducted a similar study of the relationship between knowledge and understanding of the Malay language subject, but among the SISC+BM officers.

It is strongly felt that the Malay language teachers should be guided and exposed consistently to the contemporary teaching method and learning according to the current demands and needs. Mentoring by the School Improvement Specialist Coaches (SISC+) has also been proposed to assist teachers. Efforts towards guiding the teachers are essential, as it is strongly believed that the success of a school depends mainly on the quality of its teachers. Implementing the newly-introduced writing skills assessment is not going to be an easy journey. Professional courses or training should also be provided to teachers from time to time for personal development. Besides that, the monitoring issue was raised by Chin et al. (2019) and Rietdijk et al. (2018) as one of the concerns regarding the implementation of KSSM in general. Similarly, this calls for the need to monitor the implementation of the AfL concept in the Malay language writing skills assessment in Malaysian schools. Through consistent monitoring, courses, and training, the

chances of teachers diverging from their established courses will be alleviated.

It should also be noted that AfL has become a norm in mainstream education around the world. Therefore, to be at par with others in terms of its implementation, acquiring knowledge and understanding of the assessment system is no longer an option for teachers, but mandatory if they do not want to be left behind. To produce students with 21st century learning skills, teachers will first need to equip themselves with 21st century teaching skills. AfL is a reformation towards the education of the 21st century. It is aimed at measuring student performance holistically and comprehensively, which is equal to the holistic characteristics of 21st century learners. The present method of evaluation is no longer exam-oriented but performance-based. In addition, students' success is measured beyond the grades. Competencies such as critical thinking, problem-solving, collaboration and teamwork, communication, and digital literacy are among the learning outcomes that need to be achieved. Teaching needs to be adjusted to match the outcomes. Since the process of learning matters, motivation, self-regulated learning, and progress monitoring are promoted during the teaching and learning process. Assessments are aligned with the teaching and learning outcome so that the measurement done is reliable. All these knowledge and skills must be acquired by teachers to ensure the accuracy and reliability of the assessment done. The impact of any misjudgment on students' performance is huge since the autonomy to

decide whether or not students can proceed to the next level of learning lies entirely on the teacher as the assessor in the school.

CONCLUSION

As school-based assessors, teachers are accountable for their student's achievement. Therefore, they need to ensure that their role is consistent and robust. Teachers must also persistently increase their knowledge in the field of their expertise. The procedure for implementing the assessment in the classroom must be clear, and teachers must adhere to the correct procedure. It is to ensure that the assessment grades genuinely reflect the talent and potential of the students. The main findings of this study showed that the role of teachers as school-based appraisers is established and consistent with the guidelines outlined by the Ministry of Education Malaysia. This provides input to various parties, especially to the Ministry of Education Malaysia (KPM), and acts as a call to action for the ministry to strive for a transparent and highly reliable implementation of assessments among teachers in the school.

Strictly speaking, a teacher must be skilled and educated in the art of composition before conducting an evaluation. The researchers emphasize the importance of developing an assessment system that considers the growth and development of the individuals being assessed on how a practical technique of assessment can be developed; and on the persons or agencies that may be engaged in the assessment. Objectivity in assessment must be studied since the current educational landscape continues to demonstrate something that is diametrically opposed to the goals and purposes of the assessment system in teacher education. This is quite concerning, as these issues, if not addressed, might jeopardize the goal and vision of national education. Last but not least, the findings of this study are anticipated to benefit all relevant parties, especially those in the field of Malay language teaching and learning, focusing not only on classroom assessment practices but also on other forms of assessment such as virtual assessment and benchmarking in an AfL context.

ACKNOWLEDGEMENT

REFERENCES

Arsaythamby, V., Hariharan, N. K., & Ruzlan M. A. (2015). Teachers' knowledge and readiness towards implementation of school-based assessment. *International Educations Studies*, *8*(11), 193-203. https://doi.org/10.5539/ies.v8n11p193

Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., Smith, K., Timperley, H., Volante, L., & Smith, C. W. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education, 13*(1), 117-140. https://doi.org/10.1177/1478210314566733

Black, P. (2015). Formative assessment - An optimistic but an incomplete vision. *Assessment in education: Principles, Policy & Practice*, *22*, 161-177. https://doi.org/10.1080/0969594X.2014.999643

Board of Examination. (2017). *The management guide of school based assessment.* Ministry of Education.

Chan Y. F., & Sidhu, G. K. (2012). School based assessment among ESL teachers in Malaysian secondary schools. *Journal of the Malaysian Education Dean's Council*, *9*, 1-18.

Chin, H., Thien, L. M., & Chew, C. M. (2019). The reforms of national assessments in Malaysian education system. *Journal of Nusantara Studies*, *4*(1), 93-111. https://doi.org/10.24200/jonus.vol4iss1pp93-111

Coker, D., Farley-Ripley, E., Jackson, A., Wen, H., MacArthur, C., & Jennings, A. (2016). Writing instruction in first grade: An observational study. *Reading & Writing: An Interdisciplinary Journal*, *29*, 793-832. https://doi.org/10.1007/s11145-015-9596-6

Cronbach, L. J, (1990). *Essential psycholigical testing.* Harper and Row.

Graham, S. (2019). Changing how writing is taught. *Research in education*, *43*(1), 277-303. https://doi.org/10.3102/0091732X18821125

Hashim, O. (2009). Writing skills enhancement assessment Malay secondary school level. *Educationist, 3*(2), 90-103.

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment forlearning in classroom practice. *Educational Research Review*, *17*, 50-62. https://doi.org/10.1016/j.edurev.2015.12.002

Huzaina, A. H. (2007). Application of peer review in improving writing skills. *Educational Problems*, *30*(1), 51-60.

Ishak, N. (2011, October 9-12). School based assessment as transformation in educational assessment. In *Keynote speaker at the International Seminar on Measurement and Evaluation (ICMEE 4)* (pp. 9-12). USM, Malaysia.

Izam M. T. N., Zamri M., Rahman, H. A., & Rozaiman, M. (2012). Reliability inspection authorship based on gender and holistic experiences among primary school teachers of English. *Malay Education Journal*, *2*(1), 43-64.

Jamil, H., Said, R. R., Sabil, A. M., & Kiram, N. A. M. (2017). The comparison of oral assessments aspect based on the KBSM descriptors with the learning standard in Standard Based Curriculum for Secondary Schools (KSSM). *International Journal of Education and Training*, *3*(2), 1-9.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*(3), 607-610. https://doi.org/10.1177/001316447003000308

Lee, S. K. (2012). *Preparing 21st century teachers: Singapore's approach*. International Summit On The Teaching Profession.

Lim, H. L., Wun, T. Y., & Chew, C. M. (2014). Enhancing Malaysian teachers' assessment literacy. *International Education Studies*, *7*(10), 74-82. https://doi.org/10.5539/ies.v7n10p74

Malaysian Examination Board. (2012). *Implementation of SBA for Improvement of UPSR and PMR*. Ministry of Education in Malaysia.

Malik, M. I. S. A., Zohir, A. M., & Hashim, O. (2006, Ogos 27-29). Writing proficiency level English proficiency level II primary school in peninsular Malaysia. In *Proceedings of the Regional Conference of Malay-2: Improvement of Professional Practice Teaching*. Universiti Sains Malaysia, Malaysia.

Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making.* Corwin. https://doi.org/10.4135/9781506335568

Marohaini, Y., & Zulkifli A. M. (1997). Essay writing skills teaching english in the secondary school: A qualitative assessment. *Pertanika Journal of Social Science & Humanities*, *5*(2), 65-72.

Marzni, M. M. (2014). *Teachers' qualitative perspectives on the implementation of assessment practices in the teaching of argumentative essay writing skills.* UTM Press.

Mei, T. A. (2010). *Are our teachers ready for school-based assessment? A management guide to school based assessment*. Gerak Budaya Enterprise.

Ministry of Education. (2013). *Malaysian education blurprint 2013-2025: Preschool to postsecondary education.* Ministry of Education.

Ministry of Education. (2014). *Upgrading school based assessment*. Ministry of Education.

Ministry of Education. (2016a). *Annual report 2015: Malaysian education blueprint 2013-2025*. Ministry of Education.

Ministry of Education. (2016b). *Domain for assessing writing in curriculum standard documents & secondary school assessment (DSKP), 21st Century Skills.* Ministry of Education.

Ministry of Education. (2016c). *Malaysian context on school based assessment (Ministry of Education perspective)*. Ministry of Education.

Ministry of Education. (2017a). *PPPM 2013-2025, Secondary Malay language syllabus in the Standard Secondary School Curriculum (KSSM)*. Ministry of Education.

Ministry of Education. (2017b). *The blueprint of Malay language education roadmap 2016-2025*. Ministry of Education.

Ministry of Education. (2017c). *Standard Documents & Secondary School Assessment (DSKP). 21st Century Skills.* Ministry of Education.

Ministry of Education. (2017d). *PPPM 2013-2025, Secondary Malay language syllabus in the Standard Secondary School Curriculum (KSSM).* Ministry of Education.

Mutalib, S. A., & Jamil, A. (2012). Use of formative assessment techniques subject Malay standard one: Case study. *Malay Education Journal*, *2*(1), 17-30.

Naim, H. A., & Talib, R. (2014, April 9-10). Challenges of school based assessment: Teachers' dilemma. In *Prosiding Seminar Antarabangsa Kelestarian Insan 2014* (INSAN2014). Johor, Malaysia.

Norazilawati, A., Noorzeliana, I., Sahandri, G. H. M., & Saniah, S. (2015). Planning and implementation of School-Based Assessment (SBA) among teachers. *Journal of Social and Behavioral Sciences, 211*, 247-254. https://doi.org/10.1016/j.sbspro.2015.11.031

Normah, O. (2006). *Assessment of direct writing in esl classroom in selected Malaysian secondary schools* (PhD dissertation). Universiti Putra Malaysia, Malaysia.

Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analsysis using SPSS.* McGraw-Hill International.

Paramasivam, T., & Ratnavadivel, N. (2018). Issues and challenges in managing curriculum change in primary schools : A case study of managing year four history curriculum in the district of Kuala Selangor , Malaysia. *Journal of Research, Policy & Practice of Teachers & Teacher Education*, *8*(1), 18-31. https://doi.org/10.37134/jrpptte.vol8.no1.3.2018

Pedder, D., & James, M. (2012). Professional learning as a condition for assessment for learning. In J.Gardner (Ed), *Assessment and learning* (pp. 33-48). Sage. https://doi.org/10.4135/9781446250808.n3

Pride, J. B. (2019). Sociolinguistics. In A. S. Abramson (Ed.), *Linguistics and adjacent arts and science* (pp. 1607-1628). Penguin Books. https://doi.org/10.1515/9783110811278-005

Rietdijk, S., Van Weijen, D., Jassen, T, van den Bergh, H., & Rijlaarsdam, G. (2018). Teaching writing in primary education: Classroom practice, time, teachers' beliefs and skills. *Journal of Educational Psychology*, *110*, 640-663. https://doi.org/10.1037/edu0000237

Rohaya, T., & Najib, A. G. M. (2008, November 25-27). Construction and validation of an instrument to measure the literacy level of secondary school teachers in Malaysia. In *Internal Working Paper: Postgraduate Education Research Seminar*. Universiti Teknologi Malaysia, Malaysia.

Rozita, S., Azhar, M. S., & Rasid, J. A. (2019). The relationship between knowledge and understanding factor of SISC+ Malay language officers and competency standards in personal, profession and social domains. *Asia Pacific Journal of Educators and Education, 34*, 49-63. https://doi.org/10.21315/apjee2019.34.3

Sabbir, F. (2019). Perceived view of teachers towards Pentaksiran Tingkatan Tiga (PT3) (Form Three Assessment) English Language : A case study. *Asian Journal of University Education, 15*(3), 34-44. https://doi.org/10.24191/ajue.v15i3.7819

Troia, G., & Graham, S. (2017). Use and acceptability of adaptations to classroom writing instruction and assessment practices for students with disabilities: A survey of Grade 3-8 teachers. *Learning Disabilities Research & Practice*, *32*, 257-269. https://doi.org/10.1111/ldrp.12135

Vahapassi, A. (1988). The domain of school writing and development of the writing tasks. In T. P. Gorman, A. C. Purves & R. E. Degenhart (Eds.), *The IEA study of written composition I: The international writing tasks and scoring scales* (pp. 15-40). Pergamon Press. https://doi.org/10.2307/358148

Vingsle, C. (2014). *Formative assessment: Teacher knowledge and skills to make it happen* (Licentiate dissertation). Umeâ University, New York.

Perspectives of Test Examiners of the Localized Speaking Assessment Framework: A Case Study in Vietnam

Thi Nhu Ngoc Truong¹*, Arshad Abd Samad² and Thi Thanh Phan³

¹University of Economics Ho Chi Minh City, 59C Nguyen Dinh Chieu Street, Ward 6, District 3, 700 000 Ho Chi Minh City, Vietnam
²Faculty of Educational Studies,Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
³University of Labor and Social Affairs (Campus 2), 1018 To Ky Street, Tan Chanh Hiep Ward, District 12, 700 000 Ho Chi Minh City, Vietnam

ABSTRACT

The present study explores the test examiners' perspectives on the role and qualitative aspects of the current localized speaking assessment framework used in Vietnam. A case study with two experienced test examiner-cum-English lecturers was conducted. Inductive content analysis was used to analyze the qualitative data findings obtained from individual semi-structured interviews. Drawbacks, merits, and standardization issues of the current localized speaking assessment frameworks, i.e., the Vietnamese Standardized Test of English Proficiency (VSTEP), were also discussed, especially in comparison to internationally recognized examinations and frameworks such as the International English Language Testing System (IELTS) and the Certificate in Advanced English (CAE) as well as the Common European Framework of Reference for Languages (CEFR). The study informed both English educators and policymakers to improve localized speaking assessment to suit the local teaching needs while still meeting the requirements of widely accepted international proficiency tests.

Keywords: CEFR, speaking assessment, speaking skill, test examiners, VSTEP

INTRODUCTION

Assessing oral production is often a challenging task as the nature of language comprises explicit knowledge, which

students learn through formal schooling, and implicit knowledge when they are exposed to multimedia sources and real-life communicative settings. Moreover, during oral tests, students have to process information to use grammar, vocabulary, and phonology appropriately and may also be called upon to demonstrate sociolinguistic competence (Canale & Swain, 1980; Liontas & Siegel, 2019). All these expectations in speaking assessment represent a challenge for the students to produce native-like speech (Seifoori & Vahidi, 2012). A major concern for language examiners, thus, revolves around the need for explicitly delineated objective criteria for marking oral skills that take into consideration all the aspects of effective speaking ability.

Efforts to develop and improve criterion-referenced assessment for the speaking skill have been highlighted by language scholars (Liu & Jia, 2017). International assessments and frameworks such as the Common European Framework of Reference (CEFR) for languages have been very influential in this respect in the past few years. Interestingly, there have been increased calls for more localized tests and assessment frameworks in language assessment practices to meet the demands of various groups of learners in EFL countries, for example, the General English Proficiency Test (GEPT) in Taiwan (Wu, 2012) and the Fudan English Test in China (Fan & Ji, 2014). In line with this trend, the Vietnamese Standardized Test of English Proficiency

(VSTEP), approved in 2015, has been gradually used by many local educational institutions to replace international English proficiency exams in Vietnam (Nguyen et al., 2020) and is considered an alternative to international tests such as TOEIC, PET, KET, and IELTS (T. N. Q. Nguyen, 2018).

Since VSTEP was introduced, many training programs have been conducted for test examiners, writers, and validators. A few studies have reported the effectiveness of those training programs on test validity or ratings (T. N. Q. Nguyen, 2018; Nguyen et al., 2020; T. P. T. Nguyen, 2018). Nevertheless, none of these studies have focused on the qualitative aspects and practices of the VSTEP speaking test from the practitioners' perspective. This study, therefore, reports on the views of two experienced test examiners who are English university instructors in Vietnam regarding the current speaking assessment practice in using the existing localized speaking test, especially within the prevailing context of internationalization. In particular, this study seeks to answer the following questions:

1. What are the participants' perceptions of the VSTEP speaking test?
2. What are the participants' perceptions of the VSTEP speaking assessment practice?
3. What do the participants think about standardization in speaking assessment practices in Vietnam?

LITERATURE REVIEW

English Language Assessment Practices in the Vietnamese Context

Overview of the English Language Assessment Practices in Vietnam. English is a mandatory subject in the Vietnamese educational system for all academic levels and a compulsory national examination for high school students to enter university (Hoang, 2010). English language assessment in Vietnam has undergone three main phases (Vu, 2016). During the pre-scientific phase in the 1990s, pre-constructed test papers were mainly designed by lecturers in top universities; around 100 to 200 preconstructed mock test papers for each subject, including English, were released to the public for students to review before the official exam dates (Vu, 2016). Selected universities, designated by the Ministry of Education and Training (MOET), subsequently chose a set of test papers randomly, edited and censored them to produce the official test paper. These pre-constructed test papers were designed to narrow the curricular content as an early form of standardization for the whole country; however, they became counterproductive because they encouraged teachers' teaching to the tests and learners' rote learning.

The second period, 1996 to 2007, was directed towards standardization for reliability (Vu, 2016). Electronic marking with closed-ended questions was first piloted in 1996 for national exams to avoid raters' subjectivity and errors in scoring. In 2002, the multiple-choice question university entrance exam was promulgated under the three policies: same paper, exam date, and results, which all universities used for admission decisions. To enhance the English teaching and learning quality and meet the challenges of globalization, in 2008, the MOET approved the National Foreign Language Project 2020, aiming to produce a skilled workforce able to communicate competently, independently, and confidently in a multicultural and multilingual environment (The Prime Minister of the Socialist Republic of Vietnam, 2008). A standardized speaking test and assessment framework that meets international standards while being localized to suit the national needs was essential to fulfill the goal.

The National Foreign Language Project 2020 marked the third stage, standardization for reliability and validity (Vu, 2016). This project adopted the CEFR and proposed the six-level foreign language competency framework for Vietnam in 2012, called the Common European Framework of Reference - Vietnam (CEFR-V). The six levels of competency in the CEFR-V, parallel to those of the CEFR, were localized to orient English curriculum design and assessment (Hoang, 2010; Le et al., 2017; T. Nguyen, 2017; T. N. Q. Nguyen, 2019; Pham & Bui, 2019). Accordingly, CEFR-V was introduced at the primary and secondary school levels. Meanwhile, at the tertiary level, the foreign language curriculum is decided by each institution following guidelines provided by the government. Following this policy, Vietnamese students

can choose either a domestic (i.e., VSTEP) or an international English language proficiency test (e.g., FCE, IELTS, and TOEFL) to take as long as they obtain at least level 3 of the CEFR-V or B1 on the CEFR to graduate (Le et al., 2017; Pham & Bui, 2019).

Vietnamese Standardized Test of English Proficiency (VSTEP). The first localized proficiency test, i.e., VSTEP 3-5, was designed by Vietnam National University in 2012, assessing four skills: listening, speaking, reading, and writing (T. N. Q. Nguyen, 2019). After three years of planning, designing, and piloting, in 2015, the MOET approved the official utilization of the national standardized VSTEP.3-5 and the CEFR_V as the benchmark for English language assessment nationwide (Nguyen et al., 2020). Following the VSTEP.3-5 test format (assessing English proficiency levels 3, 4, and 5 of CEFR-V, equivalent to levels B1, B2, C1 of CEFR), other variants of VSTEP were designed such as VSTEP.1 (i.e., level 1 or A1-CEFR), VSTEP.2 (level 2 or A2-CEFR), and even level 6 (or C2-CEFR) which is supposed to be beyond the English capacity of the majority of Vietnamese people (T. N. Q. Nguyen, 2019). VSTEP tests were designed with a globalized quality and localized to meet the national standards. In effect, these tests are considered a reliable instrument to measure the English ability of Vietnamese adult learners from different professions and levels of qualification (T. N. Q. Nguyen, 2019).

Because of its more comprehensive range of users compared to other VSTEP tests, in this paper, we focused on VSTEP.3-5, which is more common for most employees and students in Vietnam. The test aims to test interaction, discussion, problem-solving, and presentation skills and includes three parts: social interaction (comprising 3-6 questions about two different topics), solution discussion (requiring students to select, present, and defend their solution to a given situation from three suggested solutions), and topic development (requiring students to ask questions about a given topic using prompts to develop their ideas) (MOET, 2015) (see Appendix B for a VSTEP sample test). VSTEP.3-5 scores are measured on a scale from 0 to 10, based on five marking criteria: grammar (range and accuracy), vocabulary (range and control), pronunciation (individual sounds, stress, and intonation), fluency (hesitation and extended speech), and discourse management (coherence, cohesion, and thematic development) (MOET, 2015)**.**

Factors Relevant to Oral Performance

Speaking assessment practices can be affected by diverse factors such as task and interlocutor characteristics, test validity and reliability, assessment criteria (Fan & Yan, 2020; Kang & Wang, 2014), rater effects (McNamara et al., 2019), and rater training (Kang et al., 2019). Studies have shown that interaction tasks involving interactions with examiners can be unnatural compared to paired or oral group tests, in which the test taker interacts

with another test candidate (Brooks, 2009; Winke, 2013). O'Sullivan (2002) found an acquaintanceship effect in an experimental study with 32 Japanese students for decision making, personal information exchange, and narrative tasks, subsequently confirmed by Norton (2005) in the document analysis of 15 transcribed recordings of pairs of candidates for the FCE test in the UK. These studies indicate that subjects achieved higher scores when collaborating with a friend rather than a stranger. Interaction effects between the gender of the interlocutor and acquaintanceship were also found for grammatical accuracy (O'Sullivan, 2002).

Ahmadi and Sadeghi (2016) found that accuracy, fluency, and complexity differed across three tasks (monologue, interview, and group discussion), and accuracy was significantly correlated with the analytical and holistic assessment scores. The score differences between these two assessment methods were also documented in prior studies (Namaziandost, 2019; Namaziandost et al., 2019). Moreover, because candidates' oral performance is assessed through speech features, dependent on the test purpose and construct, these features may have different score weightings (Plough, 2018). Also, various tasks require different responses and rating scales, affecting the test-takers performance (Chalhoub-Deville & Wigglesworth, 2005). For instance, responsive and interactive tasks require the test-taker to interact more with an interlocutor peer or test examiner than the imitative, intensive, and extensive tasks (Brown & Abeywickrama, 2010).

Despite these multi factors, the basics of speaking assessment involving scales, raters, and methods should be considered to ensure the reliability and validity of speaking assessment (Ginther, 2020). While methods and scales are objective and predetermined, rater characteristics and rater bias were reported to be inconsistent over time (Lumley & McNamara, 1995). Moreover, raters' elicitations of demonstrating speaking competence, structuring talk sequences, and questioning techniques lead to variations in the impressions of candidates' ability (Brown, 2000). However, regular rater training can improve rating accuracy and minimize rating bias (Bijani, 2018; Kang et al., 2019). Surprisingly, there were no significant differences between non-native and native speakers as assessors in the outcome scores in some studies (Rossiter, 2009; Zhang & Elder, 2014), although EFL teachers viewed native speakers and their pronunciation as ideal models (Walkinshaw & Duong, 2012).

To sum up, although numerous factors such as rating scales, task characteristics, and rater effects have been reported in previous studies as influential variables to speaking assessment quality (Fan & Yan, 2020; McNamara et al., 2019; O'Sullivan, 2002), very few studies provide insights into test examiners' perspectives about these factors. Besides, qualitative aspects of VSTEP have not been extensively reported in the literature (Nguyen, 2015). Thus, exploring expert raters' perceptions of VSTEP and its speaking assessment practice can provide initial insights into influential

variables and issues that have not been reported in previous studies on speaking assessment in Vietnam (Nguyen et al., 2020; T. P. T. Nguyen, 2018).

METHODOLOGY

Participants

Vietnamese university lecturers were invited through a TESOL network in Vietnam to participate in the study. Two female English lecturers aged 35 and 38 were subsequently recruited based on their experience in teaching and testing as VSTEP speaking examiners. Using the pseudonyms Anna and Jane, Table 1 provides demographic data of the participants.

Table 1

Participants' demographic information

Name	Institution	Years of working as an English lecturer	Years of working as a general English-speaking examiner.	Years of working as VSTEP speaking examiner	Familiarity with speaking assessment frameworks
Anna	Private University	9	7	1	IELTS, CEFR & VSTEP
June	Public University	8	7	1	IELTS, CEFR & VSTEP

Data Collection

Data were collected using in-depth semi-structured interviews. The first and third authors contacted participants via phone calls to introduce themselves, explain the purpose of the study, and schedule meetings. The participants who agreed to participate in the study signed an informed consent form. The authors agreed on an interview protocol, and the third author interviewed the participants using Skype video calls after informing them the interviews would be recorded. The interview questions involved teaching and testing experiences, perception of the localized VSTEP speaking test and assessment practice, and views regarding standardization of speaking assessment practices in Vietnam (see Appendix A). All their personal information, participation in the study, and recorded interviews were kept confidential.

Data Analysis

Data were gathered, collected, transcribed, and analyzed using inductive content analysis guidelines suggested by Creswell's (2002) guidelines. The researcher organized the qualitative data through open coding and created categories for abstraction. Accordingly, the researcher clarified the

content by writing notes and headings during reading and rereading. The final coding scheme comprises inductive codes. When common patterns were found within and across cases, the researcher identified disconfirming cases and patterns before checking and rechecking codes with data and clustering them into categories. The researcher continued revising and refining the category system, and within each category, the researcher searched for sub-topics, including contradictory viewpoints and new insights. Suitable direct quotes from the interviews were used to illustrate, support, validate the findings (Thomas, 2006).

Reliability and Credibility

Our findings were based on raw data. We employed reliability procedures, including conducting multiple transcripts reviews to reduce mistakes in the participants' narratives of their experiences (Creswell, 2007). Multiple authors were involved in the coding process. Our positionality was employed as a form of reliability (Merriam & Tisdell, 2015). As the researchers, we were aware that reflexivity affected how we made meaning of the participants' worldviews. The position of the first and third authors as full-time university English lecturers and speaking examiners in Vietnam also provided access to and acceptance by our participants.

RESULTS AND DISCUSSION

This section presents and discusses the findings in three categories: perceptions of

the VSTEP speaking test, rater training and styles, and beliefs about standardization in speaking assessment practices in Vietnam.

Perceptions of the VSTEP Speaking Test

Both examiners agree that the format of the VSTEP speaking test is well-organized and localized, with three main parts varying from basic to a higher level of task difficulty (social interaction, solution discussion, and topic development). Also, participants shared similar opinions in that the assessment criteria for VSTEP are detailed, although the test lacks natural interaction and a high level of reasoning skills.

Localization and Authenticity. Anna believed that the VSTEP test was localized and reliable. In addition, the language of instruction in the VSTEP test is easy to understand as test writers considered different language backgrounds, especially students from low to high levels of English.

> *I think the test is somehow reliable, and it is localized. However, the sentences and questions in VSTEP are very short though the wording of the task requirements is clear. It may be because the test writers think that candidates' general English proficiency is not high, so when they write the test, they aim at students of the average English proficiency level. (Anna)*

Besides, both Jane and Anna found the three parts in a VSTEP speaking test similar to those in the international standardized

test, e.g., IELTS. However, they both opined that VSTEP topics were sensitive to the Vietnamese contexts for international exchange inside the country. For Anna, this might be a unique feature of VSTEP. Anna also elaborated that VSTEP topics had higher authentic features involving real-life situations.

> *The task in part 2 is designed to suit the Vietnamese context, more practical, authentic, and applicable, compared to part 2 in IELTS, which is related to personal topics. Part 2 in VSTEP requires test-takers to explain a problem to a person and the choice they go for to solve the problem…This test is more useful in real life than IELTS because students need to communicate and discuss and persuade others regarding practical problems. (Jane)*

> *[In] IELTS, students just talk about a given topic like a book they prefer… However, sometimes, many students do not like reading books [and] may not have any ideas to talk about...For this reason, I think the authenticity of the IELTS test is not high. In VSTEP, students are presented with three options, and at least students get interested in one of the three given options. So, the authenticity is high. (Anna)*

However, Anna felt that test questions written by non-native speakers were still less reliable and natural than those written by native speakers. She justified her view by stating that she sometimes found grammar

and spelling mistakes in the VSTEP speaking tests written by Vietnamese test writers compared to the questions written for international standardized exams.

> *Regarding the wording of the test questions and description of tasks, I think as VSTEP tests are written by the Vietnamese, they may not sound as good as those in IELTS written by native English speakers…Sometimes, there are some typing or spelling mistakes, lacking the verb "to be" or auxiliary verbs in VSTEP tests. (Anna)*

Detailed Assessment Criteria. Both participants have positive attitudes towards the VSTEP rating scale at the macro level, stating that the marking rubrics are clear and detailed. The new criterion, which they did not find in other familiar frameworks such as CEFR and IELTS, concerns discourse management, including coherence, cohesion, and theme development. When comparing VSTEP with IELTS, Anna highlighted the importance of the discourse management criterion. She explained that this criterion enabled her to give an accurate assessment of other criteria.

> *In IELTS, there are no discourse management criteria, so I think this is a drawback in IELTS as discourse management is very important. If we give an accurate assessment of students' discourse management which comprises thematic development, coherence, and cohesion, in VSTEP we can mark the remaining criteria accurately. However,*

if we give the wrong assessment of students' discourse management, we may not accurately assess other criteria such as grammar, vocabulary, pronunciation, and fluency. (Anna)

Likewise, Jane considered discourse management the "new assessment criterion," although she did not highlight its importance. She added that VSTEP assessment criteria were designed for analytical assessment, enabling her to assess students' performance accurately because "it is possible to give scores on a scale from 1 to 10 more precisely." However, when comparing tasks in VSTEP with those in the CAE exam, Jane mentioned that VSTEP speaking tasks were less interactive. For example, in the VSTEP part 1, although the candidate interacted with the assessor, the interaction was not natural.

[T]he CAE test is more interactive… because in some tasks, candidates interact with each other, and they have to discuss and share opinions. Although part 1 of the VSTEP speaking test is titled Social Interaction, the examiners ask only one candidate for a speaking test session, and the candidate shares their personal experience. Thus, they are not actually interacting socially with the examiner. (Jane)

Sharing the same opinion, Anna emphasized that interaction criterion be included for VSTEP because this could

ensure authenticity and identify candidates memorizing prepared notes.

[In] CEFR, they have the interaction criterion to avoid students' preparation of tasks in advance…In CEFR, the interaction task is designed to test students' natural interaction with others...I think interaction should be a criterion in VSTEP. (Anna)

However, Anna cautioned that the interlocutor's characteristics could affect the test taker's performance in paired oral tasks. For example, she explained, "sometimes one student says something the other student cannot grasp the main idea due to bad pronunciation, and this creates some difficulty for the test-taker."

In terms of reasoning skills, Anna and Jane posited that the purpose of part 3 in both VSTEP and IELTS is similar, i.e., assessing reasoning skills and a higher level of linguistic ability. However, they commented that VSTEP part 3 was more manageable than IELTS part 3 because candidates were given a mind map with three provided ideas as prompts to enable test-takers to elicit their ideas. Meanwhile, in IELTS part 3, candidates must discuss follow-up topics at a more macro level without hints, clues, or prompts, requiring a higher level of cognitive thinking and knowing a wide range of social and educational topics.

In part 3 of the IELTS test, candidates will be asked deeper argumentative questions to share their views on

more macro issues... Part 3 of VSTEP speaking is similar when candidates are asked about macro follow-up questions but only after being given a topic with a concept map. The suggestion on this mind map facilitates the candidate to answer and give them hints… In my opinion, part 3 of the two tests is quite similar because both assess the linguistic ability and reasoning ability of candidates. (Jane)

About task 3 in VSTEP speaking test, there is a mind map of ideas, so for candidates who need time to think about ideas, they can still base on 3 suggested ideas to come up with their ideas. (Anna)

In general, both participants had positive views about the VSTEP speaking test and its assessment criteria, which contribute to content validity information of the VSTEP test that was previously validated for the reading and writing skills (T. N. Q. Nguyen, 2018 & T. P. T. Nguyen, 2018). Furthermore, because candidates could apply the solution situation in VSTEP topics to the real-world context, sociocultural expectations were considered in constructing the VSTEP tests. However, as Anna revealed that VSTEP oral tests written by non-native speakers contain errors and do not sound natural in terms of wording, it seems that Anna may view native speakers as ideal models of standard English, consistent with EFL teachers' beliefs about native speakers reported by Walkinshaw and Duong (2012). Besides, because paired tasks and the assessment of

interaction are not included, participants indicated that the authentic interaction level in the one-to-one interview in VSTEP speaking tasks was less natural, which coincides with Brooks' (2009) and Winke's (2013) empirical findings on the superiority of the interactive nature in paired tasks.

Also, cognizant that the interlocutor's characteristics could influence the examinee's performance, as reported in previous studies (Norton, 2005; O'Sullivan, 2002), Anna may mean that various tasks should be included to cancel out weaknesses of each task. However, a need for different task types may mean different task-specific assessment scales, as Chalhoub-Deville and Wigglesworth (2005) posited. Thus, VSTEP designers should consider which tasks can bring more positive washback effects to improve the current oral test. Finally, the provision of prompts in VSTEP task 3 indicates that the VSTEP test considers EFL learners' characteristics and difficulties in language processing, which may also explain why the requirement of the reasoning skill is not so high in Vstep oral tasks.

Rater Training and Styles

Intensive Rater Preparation. Participants affirmed that training is an essential element for effective assessment. For example, Jane shared that to become an official VSTEP speaking examiner, she had to complete "a two-week training program," including "120 offline periods for writing and speaking assessment" and "240 periods for online studies" and mark at least ten students' oral performances together with

an experienced examiner. She added that the scores difference between her and another rater's marking "should not exceed two scores." Attending the same program, Anna reflected that the VSTEP training equipped her with useful assessment knowledge. She also observed a change in her assessment style because the detailed descriptions of the VSTEP rating scale that she was trained with rendered the assessment procedure more logical and transparent to her.

> *I think I learned many new things when participating in training programs. Before, I only gave a subjective assessment based on my experience, but when I attended the training, I was given the rating scale with detailed descriptions for every criterion. I think the speaking assessment becomes clearer and logical. (Anna)*

Positive Assessment Style and Rater Drifts. Both Jane and Anna disclosed that they based their assessment style on the 'can-do mindset,' i.e., the candidate's actual oral production guided in the rater training. Although the participants did not explain why they adopted the 'can-do mindset,' it can be inferred that the 'can-do statements' describing the proficiency levels in CEFR-V may be transformed into the 'can-do mindset' assessment, i.e., the positive assessment style for the VSTEP oral test.

> *[We] do not deduct students' marks but assess them based on the "can-do mindset." (Anna)*

For example, when a test-taker does not perform very well in part 1, but in part 2 and 3, they can perform well, I mark [their] performance based on what they have performed and what they can answer. I don't deduct scores. (Jane)

However, although the VSTEP assessment rubrics were designed to assess students' oral performance more analytically, Jane shared that she often used a holistic instead of an analytic approach to assessing overall performance.

> *[M]ost often, I have a holistic assessment of students' performance after performing all three tasks. But this is not mentioned in the rating scale. (Jane)*

Likewise, Anna reflected that although the marking rubric was extensively descriptive, she usually did not have sufficient time to refer to the rubric during oral exam marking because she had to listen to the test taker's responses. Hence, she relied on her memorizing the general description for each band score and her subjective experiences to mark her students' responses.

> *[T]here is not enough time to simultaneously listen to students' performance and read the band descriptions to give them scores…I remember the general description of each band score, and based on my personal experience, to give students marks. (Anna)*

Besides, another reason for the change from an analytic to a holistic scoring approach after time elapsed from the training is that both participants found that several sub-criteria in the rubric were not clearly described.

> *[T]he descriptions in the band scores in some criteria are sometimes overlapped, such as band 5 and 6. Sometimes, I don't know whether to give the student 5 or 6 scores for their performance. (Anna)*

> *The scores in the middle range like 4,5,6,7 and band descriptors for these scores easily confuse test examiners… I am sometimes confused because I do not know which score in the rating scale I should go for. (Jane)*

Thus, it appears that from both participants' perspectives, training English lecturers to become test examiners was a necessary step towards standardization in speaking assessment practices, which echoes findings from previous studies that rater training mitigates rater bias and improves rating consistency (Bijani, 2018; Kang et al., 2019). However, the participants' assessment style changes are quite surprising because both Anna and Jane stated that they highly valued the detailed description of the VSTEP rating scale, which enabled them to give objective assessments to test takers. Just as Lumley and McNamara (1995) observed, this "rater drift," however, is somewhat reconciled through moderation at the end of the grading process. As task types are related to analytic and holistic approaches (Ahmadi &

Sadeghi, 2016), and rating rubrics can affect effective assessment (Fan & Yan, 2020), more training on the differences among criteria descriptors for each band score and assessment approaches for different task types can lead to improved standardization in speaking assessment practices.

Beliefs about Standardization in Speaking Assessment Practices in Vietnam

Regarding the necessity for standardization in speaking assessment practices at the national level, Anna and Jane shared a similar viewpoint that standardization was indispensable to ensure fairness, equity, and consistency among universities.

> *Standardizing speaking assessment practices is necessary to ensure equity and fairness. Right now, each educational institution has its way of assessing its students. Thus, this lacks synchronization and accuracy in assessing students. (Anna)*

> *If talking about standardization in speaking assessment practices, the promulgation of general regulations for one common framework for all educational institutions to adapt to their context can be a good choice. (Jane)*

However, both were cautious about the inherent difficulties of unifying speaking assessment practices for all local educational institutions due to differences in university contexts, learner proficiency level, and training programs.

Although it is necessary to standardize speaking assessment practice, I think it is also difficult…because students at public and private universities have different proficiency levels. (Anna)

[S]tandardizing the speaking skills assessment framework by applying the VSTEP framework…may not be necessary for some non-public organizations because they can follow the international frameworks which are more suitable for their teaching and learning context or the needs of overseas cooperation and study. (Jane)

In responding to which assessment framework should be used to standardize speaking assessment practices at the national level, Anna believed that VSTEP should be used because it was suitable "for most working people, secondary and tertiary students." However, she recommended that "English majors should study IELTS, and non-English majors should take VSTEP" because topics of VSTEP were "localized." Likewise, Jane suggested that students who planned to study overseas "should study IELTS [which] can benefit them in the long term" because VSTEP was not globally recognized. However, if students did not intend to study abroad, VSTEP could be a better choice because of "its low cost."

In general, participants hold a balanced view towards standardization in speaking assessment practices because proficiency evidence can be proved by either local or international standardized assessment dependent on the test-taking purposes and the training institutions. Since the locally produced VSTEP has not yet gained international recognition, both favored IELTS for overseas studies and academic advancement. Taking stock of the current speaking assessment practices in Vietnam, if VSTEP is to be gradually globally recognized just as other locally standardized proficiency tests, e.g., GEPT in Taiwan (Wu, 2012), validation of VSTEP speaking test and addressing challenges related to its speaking assessment rubrics is necessary.

IMPLICATIONS

To sum up, participants expressed the need to include local content in the test design, the interaction criterion in the rating scale, the importance of receiving training, and the necessity to balance standardization in speaking assessment. The participants emphasized a positive perspective towards completing the test task (as seen in their can-do mindset) and believed that assessing actual speaking ability should not be clouded by students first understanding foreign and unfamiliar contexts. VSTEP seems an appropriate assessment tool that considers localized contexts besides meeting localized objectives, especially for local employees and non-English majors. However, teachers continue to refer to the IELTS as it seems that the VSTEP has not yet received global acceptance. Despite this, the VSTEP is still a successful test. It measures language ability based on internationally accepted criteria (as indicated by its close and careful association to the CEFR), and many characteristics make VTEP a practical speaking test.

Notably, test task characteristics are important concerns raised by the two participants. Different test tasks will elicit various kinds of language as numerous factors can make speaking a complicated activity involving a high cognitive level of information processing and knowledge and consequently a difficult skill to assess. For example, more prompts were provided in the VSTEP than other tests as observed by one participant, which is understandable in an EFL context as it can encourage speech production. Although prompts make the task easier, they allow for greater speech to be produced and assessed. On the other hand, the same participant felt that interaction was not emphasized in the VSTEP. This concern needs to be addressed by exposing learners to a wide range of communicative situations and engaging them in various test tasks. Thus, English lecturers can familiarize learners with information processing in retrieving necessary core linguistic knowledge to solve the tasks. Examples of test tasks include discussing a situation, role plays, talking about a past event, solving a problem, and other real-life communicative tasks. Also, including various speech functions such as comparing, describing, expressing opinions, and persuading can increase the task difficulty, differentiate proficiency levels. Interestingly, however, there is mention of discourse management as a new criterion in VSTEP speaking assessment, which can clearly and eventually lead to a greater focus on interaction. It is also worth noting that including different speaking test

tasks can help balance out the advantages and disadvantages of each task, and test designers might have considered this by including paired tasks when composing VSTEP speaking tests.

Finally, specialized training for the test examiners must be continually provided as, without it, test assessors may find it hard to assess students objectively. Speaking assessment is very demanding on the test examiners, especially when the test examiner holds dual roles as a grader and an interlocutor because this adds to the cognitive load they face. Thus, examiners may skip or ignore the details when the scoring criteria are extensive and switch to global and sometimes subjective assessments. Despite being trained, rater drift and rater variability cause concern as the effect of training may not last long. In this respect, paired and group tasks should be considered as one-to-one oral interviews have been criticized for failing to evaluate all aspects of oral proficiency (Ockey, 2018). Besides, it should be emphasized that there is no best practice for speaking assessment practices because different tasks with various difficulty levels are designed to suit diverse purposes ranging from personal, easy, concrete to non-personal, difficult, and abstract topics. Suppose the VSTEP speaking test is comparable to other international tests. In that case, test designers should ensure that test tasks follow a justifiable order of difficulty comparable to international frameworks (e.g., CEFR) and include a wide range of tasks and updated topics. Also, the predictive validity of

VSTEP in terms of language achievement should be examined.

CONCLUSION

Adopting an existing assessment wholesale may be easy, especially if it has been internationally recognized and accepted. However, as is the mantra today, assessment is not necessarily just for the sake of assessment and should also encourage learning. Hence, national and localized examinations such as the VSTEP are not surprisingly slowly becoming a more common occurrence. From the two participants' perspectives, VSTEP has positive features (e.g., localized topics, availability of prompts, and detailed assessment criteria) and drawbacks (e.g., the lack of high interactivity, extensive and overlapping criteria descriptions, and not being completely free from grammatical errors). Nevertheless, the interviewed participants believe that standardization in speaking assessment practices was essential to ensure assessment fairness, equity, and consistency, especially among local educational institutions. To achieve this, though, a balanced view of standardization in assessment practices at the national level should be adopted as various institutions had different training and educational purposes, and learners also had various study intentions. Furthermore, although VSTEP was designed based on an internationally accepted framework, it has yet to receive complete local acceptance, let alone global recognition. Thus, caution should be taken when imposing standardization practices using VSTEP for all local educational institutions.

The research findings provide useful information about the drawbacks and merits of VSTEP and localized speaking assessment practice for English test examiners and administrators. Vietnamese test examiner-cum-English lecturers' positive attitude and critical evaluation on VSTEP play an initiative role in inspiring other EFL countries to create their own localized English proficiency tests which are equivalent to other international standardized English proficiency tests in terms of quality and validity suitable for national or domestic use. However, due to the limited number of participants, not all Vietnamese VSTEP test examiners' views were represented. Therefore, future research should include more VSTEP test examiners from public and private institutions to confirm the findings. We also note the special need for further studies on the VSTEP in terms of methods. For example, recent studies to validate VSTEP speaking tests only used the inter-rater reliability method to determine reliability. Hence, to prove its reliability and validity and consequently gain larger global acceptance, future studies could consider other methods like discourse analysis to confirm test validity and test score reliability. Besides, future studies can also use generalizability theory (G-theory) to validate the test as G-theory allows the researcher to determine relevant facets that are related to the assessment context (Lynch & McNamara, 1998) and their relative effects on test scores (Bachman et al., 1995; Brennan, 1992).

REFERENCES

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, *13*(4), 341-358. https://doi.org/10.1080/15434303.2016.1236797

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257. https://doi.org/10.1177/026553229501200206

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education, 5*(1), Article 1460901. https://doi.org/10.1080/2331186X.2018.1460901

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341-366. https://doi.org/10.1177/0265532209104666

Brown, H. D. (2000). *Principles of language learning and teaching*. Addison Wesley Longman Inc.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (Vol. 10). Pearson Education.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47. https://doi.org/10.1093/applin/1.1.1

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*(3), 383-391. https://doi.org/10.1111/j.0083-2919.2005.00419.x

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*. Prentice-Hall.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications.

Fan, J., & Ji, P. (2014). Test candidates' attitudes and their test performance: The case of the Fudan English Test. *University of Sydney Papers in TESOL*, *9*, 1-35.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, *11*, Article 330. https://doi.org/10.3389/fpsyg.2020.00330

Ginther, A. (2020). Assessment of speaking. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-8). Wiley. https://doi.org/10.1002/9781405198431.wbeal0052.pub2

Hoang, V. V. (2010). The current situation and issues of the teaching of English in Vietnam. 立命館言語文化研究 [*Ritsumeikan Language and Culture Studies*], *22*(1), 7–18.

Kang, O., & Wang, L. (2014). Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Research Notes, 57*, 40-49.

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481-504. https://doi.org/10.1177/0265532219849522

Le, V. C. (2017). English language education in Vietnamese universities: National benchmarking in practice. In E. S. Park & B. Spolsky (Eds.),

English education at the tertiary level in Asia: From theory to practice (pp. 283–292). Routledge. https://doi.org/10.4324/9781315391588-11

Liontas, J. I., & Siegel, M. (2019). Cultural perspectives in teaching speaking. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-8). American Cancer Society. https://doi.org/10.1002/9781118784235.eelt0696

Liu, L., & Jia, G. (2017). Looking beyond Scores: Validating a CEFR-based university speaking assessment in Mainland China. *Language Testing in Asia*, *7*, Article 2. https://doi.org/10.1186/s40468-017-0034-3

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180. https://doi.org/10.1177/026553229801500202

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment*. Oxford University Press.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

MOET. (2015). *Decision No. 729/QD-BGDDT: Approving VSTEP.3-5 format based on the Vietnamese version of the Common European Framework of Reference for Languages*. Thukyluat.vn. https://thukyluat.vn/vb/quyet-dinh-729-qd-bgddt-de-thi-danh-gia-nang-luc-su-dung-tieng-anh-tu-bac-3-den-bac-5-416b4.html

Namaziandost, E. (2019). The assessment of oral proficiency through holistic and analytic techniques of scoring: A comparative study.

Applied Linguistics Research Journal, *3*(2), 70-82. https://doi.org/10.14744/alrj.2019.83792

Namaziandost, E., Banari, R., & Momtaz, S. (2019). Evaluating oral proficiency skill through analytics and holistic ways of scoring. *Humanities & Social Sciences Reviews*, *7*(5), 424-433. https://doi.org/10.18510/hssr.2019.7547

Nguyen, A. T. (2015). Towards an examiners training model for standardized oral assessment qualities in Vietnam. *Malaysian Journal of ELT Research*, *11*(1), 41-51.

Nguyen, T. (2017, March 26-29). *Vietnam's National Foreign Language 2020 Project after 9 years: A difficult stage* [Paper presentation]. The Asian Conference on Education & International Development. Japan.

Nguyen, T. N. Q. (2018). A study on the validity of Vstep writing tests for the sake of regional and international integration. *VNU Journal of Foreign Studies*, *34*(4), 115-128. https://doi.org/10.25073/2525-2445/vnufs.4285

Nguyen, T. N. Q. (2019). Vietnamese standardized test of English proficiency: A panorama. In L. I. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 71–100). Routledge. https://doi.org/10.4324/9781351254021-4

Nguyen, T. N. Q., Nguyen, T. Q. Y., Tran, T. T. H., Nguyen, T. P. T., Bui, T. S., Nguyen, T. C., & Nguyen, Q. H. (2020). The effectiveness of Vstep.3-5 speaking rater training. *VNU Journal of Foreign Studies*, *36*(4), 99-112. https://doi.org/10.25073/2525-2445/vnufs.4577

Nguyen, T. P. T. (2018). An investigation into the content validity of a Vietnamese standardized test of English proficiency (Vstep.3-5) reading test. *VNU Journal of Foreign Studies*, *34*(4), 129-142. https://doi.org/10.25073/2525-2445/vnufs.4286

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal*, *59*(4), 287-297. https://doi.org/10.1093/elt/cci057

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277-295. https://doi.org/10.1191/0265532202lt205oa

Ockey, G. J. (2018). Oral language proficiency tests. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-5). American Cancer Society. https://doi.org/10.1002/9781118784235.eelt0234

Pham, T. N., & Bui, L. T. P. (2019). An exploration of students' voices on the English graduation benchmark policy across Northern, Central and Southern Vietnam. *Language Testing in Asia*, 9, 15. https://doi.org/10.1186/s40468-019-0091-x

Plough, I. C. (2018). Speaking assessment for high-stakes testing. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-6). American Cancer Society. https://doi.org/10.1002/9781118784235.eelt0235

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, *65*(3), 395-412. https://doi.org/10.3138/cmlr.65.3.395

Seifoori, Z., & Vahidi, Z. (2012). The impact of fluency strategy training on Iranian EFL learners' speech under online planning conditions. *Language Awareness*, *21*(1-2), 101-112. https://doi.org/10.1080/09658416.2011.639894

The Prime Minister of the Socialist Republic of Vietnam (2008). *Decision No 1400/QD-TTg*. https://vanbanphapluat.co/1400-qd-ttg

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, *27*(2), 237-246. https://doi.org/10.1177/1098214005283748

Vu, T. P. A. (2016, October 13-14). *25 years of language assessment in Vietnam: Looking back and looking forward* [Paper presentation]. International Conference on English Language Assessment. Hanoi, Vietnam.

Walkinshaw, I., & Duong, O. T. H. (2012). Native- and non-native speaking English teachers in Vietnam: Weighing the Benefits. *TESL-EJ*, *16*(3).

Winke, P. (2013). The effectiveness of interactive group orals for placement testing. In K. McDonough & A. Mackey (Eds.), *Second language interaction in diverse educational contexts* (pp. 247-268). John Benjamins.

Wu, J. R. W. (2012). GEPT and English language teaching and testing in Taiwan. *Language Assessment Quarterly*, *9*(1), 11-25. https://doi.org/10.1080/15434303.2011.553251

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306-325. https://doi.org/10.1080/0969594X.2013.845547

APPENDICES

Appendix A

Semi-structured interview questions

1. How many years have you been working as an English instructor/ speaking examiner?

2. Have you ever taken VSTEP before?

3. Have you ever participated in any speaking assessment training programs?
   If yes, can you share with me your experiences of participating in those programs?

4. What speaking assessment frameworks are you familiar with?

5. What do you think about the VSTEP test and assessment criteria? How would you compare VSTEP with other tests and assessment frameworks such as CEFR or IELTS (e.g., marking criteria and test components)? What difficulties and benefits do you think test examiners may have when using localized VSTEP rating rubrics?

6. What aspects of VSTEP speaking assessment do you think need changing?

7. What do you think about standardization in speaking assessment practices for all the Vietnamese educational institutions?

8. Do you have any other suggestions to improve the current speaking assessment practices in our country?

Appendix B

A sample VSTEP speaking test

Part 1: Social interaction (3 minutes)

Let's talk about your free time activities.
- What do you often do in your free time?

- Do you watch TV? If no, why not? If yes, which TV channel do you like best? Why?

- Do you read books? If no, why not? If yes, what kinds of books do you like best? Why?

Let's talk about your neighborhood.
- Can you tell me something about your neighborhood?

- What do you like most about it?

- Do you plan to live there for a long time? Why/why not?

Part 2: Solution discussion (4 minutes)
Situation: A group of people is planning a trip from Da Nang to Hanoi. Three means of transport are suggested by train, by plane, and by coach. Which means of transport do you think is the best choice?

Part 3: Topic development (5 minutes)
Topic: Reading habits should be encouraged among teenagers

SOCIAL SCIENCES & HUMANITIES
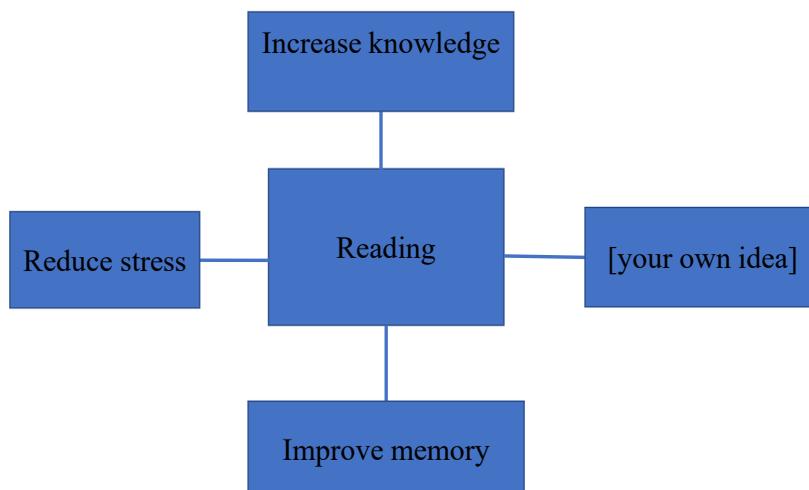
Framing the English Language (EL) CEFR-informed Curriculum Structure: The UKM Experience

Normazidah Che Musa, Azwan Shaiza Nizam, Wan Nur'ashiqin Wan Mohamad and Zarina Othman*

Pusat Pengajian Citra Universiti, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

ABSTRACT

One of the main concerns that have been raised in the realm of the graduate employability workforce in Malaysia is for graduates to be competent in speaking English. The concern on the graduates' standard in speaking English has been amplified to meet a global standard. It has triggered the Malaysian Ministry of Education to adopt the CEFR benchmark. This paper presents the framework of a CEFR-informed curriculum for Universiti Kebangsaan Malaysia (UKM) English Language (EL) courses at UKM and its basis. The paper outlines the method used in framing the curriculum structure based on the ADDIE model. It particularises the curriculum structure into four (4) main implementation phases that address the different proficiency levels in the targeted CEFR levels.

Keywords: ADDIE model, CEFR, curriculum design, English language, higher education

INTRODUCTION

The Roadmap for 2015-2025 in the English Language (EL) Education Reform in Malaysia is an important declaration of the nation's aim to improve the students' proficiency in the EL substantially. This roadmap formalises the Ministry of Education's (MoE) ongoing efforts to generate graduates with good command of the EL (Chonghui, 2019). The roadmap is introduced to address a fundamental

problem faced by Malaysian graduates today.

> "They have to cope with a rapidly changing and increasingly globalised world and job market which requires them to communicate effectively in English at a much higher level than before. They need to be aware of the wider world in which they are growing up, and they need sufficient support to enable them to achieve the necessary levels of English proficiency" (Don & Abdullah, 2019).

Central to this roadmap is the adoption of the CEFR framework as a point of reference. CEFR is introduced in the Roadmap to relate the English proficiency level of Malaysian graduates with its international relevance. It calls for the review of the existing curriculum in UKM to be framed against CEFR taking into consideration the Malaysian English Language Education Reform Roadmap 2015-2025.

For these graduates to be marketable and accepted as part of a global workforce, one of the important attributes required is communicating competently in the second language, i.e., English. It is thus central that the Malaysian graduates equip themselves with the required competence and readiness to face the demands of the global workforce. In light of this, the need to ensure graduates' competence in speaking English is intensified. The goal of English Language (EL) education at the university is to prepare graduates with the adequate English Language skills and competencies.

Studies carried out in Malaysia indicate that competency in English is a substantial factor in graduate employment. Ahmad and Zainol (2011) reported that proficiency in EL is one of the requirements for managerial posts in five-star hotels, for instance. In a market research on 295 Malaysian employers by Zubairi et al. (2011), they reported that 80% of the respondents agreed that competence in English is equally or more important than content knowledge or professional skills. This confirms the need to focus on enhancing graduates' English communication skills and competencies besides enriching their content knowledge.

The Malaysian employers, in general, perceive that the universities have not, to a certain extent, provided ample opportunities for students to develop abilities critical to the labour market. They elaborate that low proficiency in the EL and the lack of soft skills, including creativity, communication and critical thinking, are among the reasons fresh graduates lack the readiness to enter the workforce (NST Education, 2019).

In a related study, Pandian and Balraj (2013) examined final year students' readiness to enter the global workplace in the digital age. The study's findings showed that students have difficulties speaking, reading, and writing fluently in English. These weaknesses will hinder the students' professional development in their future careers, especially if they cannot participate in the English language dominated marketplace such as in international affairs and in business communication worldwide.

To conclude, these ongoing discussions on graduates' workplace competency suggest that Malaysian graduates generally require specific language training to equip themselves with expected workplace competency.

One of the aims in the EL roadmap is to adopt a CEFR approach to achieve a level of competency of international standards. The introduction of the CEFR framework into the Malaysian Education Blueprint aimed to develop student's English language competency to be at par with global competency. Several studies investigated the challenges in implementing a CEFR-aligned curriculum in Malaysia (Uri & Aziz, 2018, Darmi et al., 2018; Sidhu et al., 2018). They revealed that Malaysians have a limited understanding of the framework to adopt the fundamental shift into the CEFR. It may largely be due to educators having limited knowledge and a lack of exposure to the CEFR.

In light of all these, there is the need to benchmark the standard of the EL courses at the tertiary level of education with the Common European Framework of Reference before the students enter the workforce. Universiti Kebangsaan Malaysia (UKM) believes that framing the EL courses against the CEFR framework would help produce graduates with good mastery and competency in using English. In designing the curriculum, it is important to coordinate the implementation phases by, for instance, considering several factors that include the university's aspiration mainly to generate UKM graduates who have attained employability readiness in the real working world in the local and international arena.

As such, Arslan and Özenici (2017), in their study, had proposed a possible EFL curriculum design in line with the principles on the CEFR (Common European Framework of Reference for Languages) for tertiary education level. In addition, they had posed several elements to be considered in designing and developing key components of a CEFR-based EFL curriculum. These elements included the development of four language skills in order for language learners to be able to communicate, the consistency of content for learning and teaching EFL skills with real-life situations, the employment of communicative language teaching methods, strategies, and techniques, and the use of alternative testing and assessment.

Statement of the Problem

The impetus for this curriculum review is the pressing concerns on the need to produce graduates with English language competence meets the international standards. Abdullah et al. (2015) pose that the common issue among graduates is that they do not meet the level of English language competence required by potential employers. This apprehension led to the EL reform Roadmap 2015–2025 for the Malaysian universities to implement a CEFR aligned curriculum for English language education. As such, the language competency unit at UKM had embarked on this curriculum change by reframing the existing EL curriculum structure to be aligned to the CEFR.

A more pressing need for this curriculum change is based on the outcome of the university's oral performance assessment. Hazita et al. (2018) designed to evaluate students' oral competency before they graduate from the university. The results of the performance evaluation indicated that most of the students did not attain the targeted CEFR level of C1 (proficient users) as stipulated in the Malaysian EL Roadmap. The English language proficiency courses aim for the undergraduates, who are beginner, elementary and intermediate users of English to attain a higher proficiency level by at least one band at the end of their studies. This target has not been achieved for the two consecutive years as indicated in the results of the Competency Based English Test (CBET) conducted at the university. This is based on the 'UKM CBET' report findings for the first two consecutive years (Hazita et al., 2018). The report states that almost 70 % of the test takers are at CEFR B1 level which is categorised as lower independent users. This calls for a revision of the English language curriculum to address this issue to improve students' language competency. These two factors significantly emphasise the need for a curriculum review of the EL courses offered at UKM.

The English Language Education Reform: The Roadmap 2015-2025`

The Roadmap is a comprehensive and holistic plan that emphasises the skills and abilities required by the students to become independent users of the EL (Yusof, 2015, p. ix). The Roadmap is a timetabled implementation plan for the systemic reform of EL education in Malaysia. It aims to transform the existing EL education system from preschool to tertiary level education and teacher education (Don & Abdullah, 2019).

Studies on the Policies and Implementation of CEFR in Malaysia

Common European Framework of Reference (CEFR) for languages was established in 2001 by the Council of Europe to provide a common basis for elaborating language syllabuses, curriculum guidelines, examinations, and textbooks for countries in Europe (Council of Europe, 2017). It is not exclusively tailored to one specific language. Rather, it is a framework that can easily be adapted to teaching and learning any specific language. Its main objectives are to promote plurilingualism in the European context of its multitude of languages and cultures.

In Malaysia, Uri and Aziz (2018) carried out a study on teachers and the Ministry of Education officials' views on the adopted approach of CEFR onto the Form 5 (Upper school Secondary level) English syllabus and assessment. The study showed that, in general, the teachers were positive to adopt the CEFR framework and the English Language Roadmap in their school curriculum. They agreed that there is a need to improve the English proficiency of the students in order to be at par at the global level. However, most of the teachers also expressed that they were not ready

to adopt the CEFR framework in their teaching despite having attended the CEFR familiarisation workshops and other CEFR related training; the initial moves taken by the ministry.

The UKM initiative in reviewing and redesigning the English curriculum has primarily identified several reasons contributing to implementing CEFR in Malaysia. Bearing in mind the 'vagueness' that many language teachers similarly may have regarding the implementation of CEFR, the need to 'familiarise' the teachers with the CEFR framework was a priority. No doubt, the main challenge in the Malaysian context is the ability of the council and Ministry of Education to produce resources locally according to the local contexts since CEFR is rather 'new' in Malaysia. Teachers' limited knowledge, lack of adequate training and a low level of awareness about CEFR may also hinder the implementation of the CEFR process. It is also worth to note that several teacher elements such as the teachers' English proficiency, resistance, and lacking CEFR expertise to construct and produce local CEFR aligned textbooks can add to the challenges in the implementation.

Since the EL Education Reform in Malaysia 2015-2025 was launched, the ministry has made it compulsory for all EL teachers in universities to undergo its CEFR familiarisation training. Therefore, Darmi et al. (2017) carried out a study to understand EL teachers' views on the EL proficiency courses in a local university in Malaysia. This study examined teachers' views on students' performance in the existing EL courses based on the global CEFR descriptors. The study found that most of the students did not achieve a clear understanding of a variety of texts in particular. Furthermore, the study revealed that although students were not able to write clearly on different topics, in terms of communicative ability, they were, however, capable to communicate fluently and spontaneously to provide a clear explanation on specific areas of concern.

On another note, Sidhu et al. (2018) investigated the implementation of the CEFR-aligned school-based assessment in primary ESL classrooms in five schools located in Damansara, Malaysia. Their study revealed that the implementation of school-based assessment left much to be desired and was far from formative assessment. Though teachers expressed rather positive opinions on the assessment, they lacked the full understanding of the method and admitted possessing limited knowledge of the revised CEFR-aligned ESL curriculum. In addition, the UKM initiative had identified two areas of concern regarding the students' proficiency profiling. First, students of the lower proficiency and the intermediate proficiency levels based on the CBET results indicate the lack of ability to attain the mastery or satisfactory level (Hazita et al., 2018). Second, other than in speaking, students are also found to have the lack of ability to understand the main ideas of complex texts based on teachers' feedback that their students were less able to produce clear and detailed texts.

METHODOLOGY

In developing a new structure for an English Language CEFR-informed curriculum, UKM adopts the ADDIE model (Morrison, 2010). The term ADDIE is an acronym for Analyze, Design, Develop, Implement, and Evaluate, phases of the instructional design process. Each phase reflects the important components in the process of instructional design.

The model is most commonly used in instructional design to create "instructional" experiences that make knowledge acquisition more efficient and effective (Aldoobie, 2015; Drljaca et al., 2017; Hsu et al., 2014). It offers a systematic approach in curriculum design and implementation, which can be applied to different contexts of study and modes of instruction—be it face to face or online—and duration of instruction (Aldoobie, 2015; Razali & Shahbodin, 2015; Zhang, 2020). Despite being highly structured, this model allows flexibility in implementing the processes (Balanyk, 2017). The flexible and systematic characteristics of the model became the main reason for UKM to use ADDIE as a guiding framework in developing the new English Language CEFR-informed curriculum.

Developing CEFR-informed Curriculum using the ADDIE Model

The ADDIE model framework categorises five steps of the instructional design process (Dick et al., 2011; Gustafson & Branch, 2011). This model postulates a 5-step process in curriculum design, consisting of the Analysis stage, followed by Design, Develop, Implement and Evaluate. Hsu et al. (2014) used this model to develop and evaluate an online continuing education curriculum for a hospital in Taiwan. The study found that the model is useful and practical for course development projects as it helps describe what happens and prescribes what needs to happen (Hsu et al., 2014). Drljaca et al. (2017) used the ADDIE model to prepare the teaching materials for online courses. The study detailed the five stages suggested by the model and emphasised the iterative process of each stage. Zhang (2020) implemented the ADDIE Model in developing a college online English learning community to improve the effective interaction between teachers and learners, learners and learners and teachers and teachers. By applying the model in the development of the community, Zhang states that the five stages of this model are interconnected and are to be used as a non-linear cyclic mode (Zhang, 2020). These characteristics of the model are illustrated in Figure 1.

The application of this model on the development of the new English Language curriculum indicates that each stage comprises processes and goals which require careful and detailed planning as they are interconnected and affect the other stages. These moves are necessary to ensure that the curriculum developers in UKM can produce a sound curriculum for the students. Adapting the ADDIE model, the UKM method in framing the CEFR informed structure is described in Table 1.
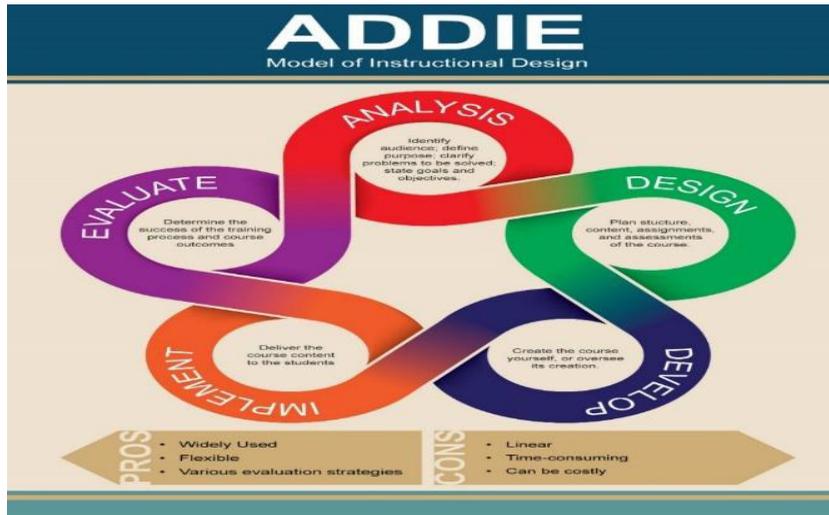
Figure 1. The ADDIE Model

Table 1

The UKM Method in framing the CEFR informed structure

Stage	Steps taken	Description	Justification
Analysis	Distribution of student questionnaires on perceptions towards the existing courses.	This questionnaire comprises items on teaching and learning processes, language skills, course materials, and assessments.	To identify student's perceptions of the courses offered, their preferences and their needs on language.
	Discussion on Teacher feedback on the courses.	Similar items in the student questionnaire were discussed.	To identify teachers' best practices, challenges in teaching the courses, feedback on course improvement.
	Discussion with faculties.	Discussions include the language skills needed by students in their academic lives and the workplace.	To identify the language needs of students from the different faculties.

Table 1 *(Continued)*

Stage	Steps taken	Description	Justification
Design	Setting up a curriculum task force to plan the new structure.	Comprising language instructors.	To develop the EL curriculum and to plan for implementation stages.
	Setting up committees for the English Language courses.	New courses were introduced in the new structure.	To design courses.
	The implementation of the design stage.	A series of training workshops on CEFR Familiarization and Constructive Alignment.	To provide instructors with the input for them to design the new courses.
Development	Preparation of the course information; *proforma*	To be submitted to the Centre for Learning Accreditation UKM	To be reviewed and approved.
	Mapping the curriculum structure.	Comparison of *proforma* across courses.	To ensure alignment of learning outcomes.
	Presentation of the proposed curriculum structure.	To the board of English Language Enhancement Programme Initiatives (the committee members of "*Initiatif Pengukuhan Bahasa Inggeris*)	To obtain feedback on the new CEFR informed curriculum structure.
	Development of course materials, evaluation tasks and assessment scales.	Based on the CEFR Book of New Descriptors.	To design course materials, evaluation tasks and assessments based on feedback.
	Submission for Senate Approval	Presentation to University Senate for approval	To obtain Senate approval
	New CEFR Curriculum Roadshow.	Presentation to faculties	To inform the faculties on the new CEFR informed curriculum structure.

Table 1 *(Continued)*

Stage	Steps taken	Description	Justification
Implementation	The commencement of the new EL structure in the new Academic session.	UKM 2019/2020 Academic Session.	To implement the new CEFR informed curriculum structure
Evaluation	Internal Audit: at the end of the semester.	1. Student and teacher feedback surveys. 2. Feedback on students' performance based on students' end of the semester grade.	To gain students' and teachers' perspectives and performance on the courses.
	External Audit: from stakeholders and industries	Presentations in seminars and conferences	To gain feedback from industries on the relevance of the courses to the workplace

THE NEW UKM CEFR-INFORMED ENGLISH LANGUAGE (EL) CURRICULUM

In the effort for English Language **(**EL) education to meet the standards and benchmarks stipulated in the CEFR, UKM designed a new "CEFR-informed" EL curriculum structure. This structure is designed to situate CEFR into the existing curriculum taking into consideration four crucial facets, which are:

i.   the former EL curriculum structure
ii. UKM students' language proficiency at the point of entry into the university
iii. the policies on EL credit requirements as stipulated by the Ministry of Higher Education (MOHE) and the university
iv. the principles and benchmarks of the CEFR

Following the national requirement for the public universities in Malaysia to offer five credits in a total of English courses to be taken, UKM offers the students a 2 +2 + 1 credit of English courses to be taken throughout their studies in the university. Forty notional hours are allocated for each EL credit. It makes a total of two hundred learning hours for each track to justify the number of hours needed for a student to improve their proficiency to the next upper band level. These courses are spread throughout their studies, from Year 1 to Year 3, to provide students with the EL continuous learning experience before starting their industrial training and practical sessions.

The Needs Analysis

At the initial stage of the new curriculum design, a survey was conducted with the

students and 30 language instructors to gain feedback on the language courses. This survey was emailed to all students who took the English language courses. A total of 801 (44.5%) undergraduate students responded to the survey.

The Likert–scale survey was divided into several sections; students' perception of language skills learnt in the course, soft skills acquired, course materials used and suitability of the course assessment. The same questions were given to the language instructors in open-ended questionnaires to gather more qualitative responses from the teachers. This feedback provides a needs analysis on the strengths and weaknesses as the basis in designing the current curriculum. The overall results indicated that students preferred courses to enhance their oral communication skills and confidence-building. In addition, the instructors' feedback indicates more emphasis on the required communication skills for students to function in academic and workplace settings.

These two findings relate to the call from the MOHE to promote a higher level of EL communication competencies amongst university students. It is aimed to prepare the graduates to be compatible with the global challenges and competitive job markets. With this vision, UKM highlights the concern to ensure that the level of language competency taught in the university meets international standards. The CEFR benchmark is used as a guiding tool in the new EL curriculum design for UKM. The CEFR's principles on focusing on the real-time communication needs of the learners provide the impetus for the design of courses in the new curriculum. The CEFR Global Scale is also used as a point of reference for describing the students' existing and targeted proficiency levels.

UKM embedded courses are interactive and integrated into nature, focusing on enhancing the students' communication skills applicable in real-life contexts in the new curriculum. The curriculum is then designed to focus on these two main pillars; i) Academia English (in comparison to Academic English) and ii) Workplace Communication English with 'speaking' as the core emphasis of each course module. The new curriculum structure is depicted in Table 2.

Table 2

The UKM CEFR-Informed Curriculum Structure

MUET Band	Year 1	Year 2	Year 3	Target
1 & 2	*Breakthrough English* CEFR A2	*Academic Interactions* CEFR B1-B2	*Let's Get Talking* CEFR B2	A2 to B2
	Bridging English CEFR A2-B1			

Table 2 *(Continued)*

MUET Band	Year 1	Year 2	Year 3	Target
3	*Academic Interactions* CEFR B1-B2	*Pro-Talk English* CEFR B1-B2	*Corporate Storytelling* CEFR B2	B1 to B2
4	*Academic Literacy* CEFR B2	*Speak to Persuade* CEFR B2-C1	*Professional Communication* CEFR C1	B2 to C1
5 & 6	*Page to Stage* CEFR C1-C2	*Advanced Communication Project* CEFR C2		C2

There seems to be an intersection of targeted CEFR levels in certain courses and tracks from the general perspective. Although the CEFR levels indicated are the same, the courses use different sets of tasks, activities and assessments. Even though the CEFR levels targeted for each course seem to intersect, the courses are designed to develop communicative competencies. The intersecting CEFR levels between courses in the same track are intentional. The progress is made visible at the 1) assessment scales, 2) course objectives, and 3) materials used.

The Malaysian University English Test (MUET)

The new curriculum takes into account students' level of EL proficiency upon entering the university. The Malaysian University English Test (MUET) bands have become a benchmark for designing the package of courses offered. Based on the students' MUET results, the students of the same proficiency levels are placed into respective phases. Each phase will offer the students the EL courses that suit their levels of proficiency and existing abilities. In determining the parallelism of competencies between MUET bands and CEFR levels, the benchmarking matrix of MUET results, CEFR and major EL competency tests are referred to. The matrix is as stated in Table 3.

The student database shows that the majority of the students entered the university with MUET Bands 3 to 5. The university also takes in several students with a MUET Band 2 level. There is also a small number of students in the range of Bands 5 and 6. Observations from the EL instructors and the scores obtained in oral tasks in the existing courses indicate that the students' oral competencies require substantial improvement compared to reading and writing. In framing the new EL curriculum, adaptations to the mapping guide were made based on these reasons:

Table 3

Mapping of MUET Results against CEFR, IELTS, and TOEFL (Malaysian Examination Syndicate, 2018)

MUET (Malaysian University English Test)			CEFR	IELTS (International English Language Testing System)	TOEFL (Test of English as a Foreign Language)	English Level
Range score	Average score	Band	Band	Band	Band	Proficiency
260-300	264.39	6	C2	>8.0	110-120	Advanced
220-259	232.90	5	C1	7.0-8.0	94-109	
180-219	202.60	4		6.0-6.5	60-93	Upper intermediate
140-179	163.40	3	B2	5.5	46-59	
100-139	125.90	2	B1	4.0-5.0	31-45	Lower Intermediate
<100		1	A2	<4.0	<30	Elementary

1. The Band 5 and 6 students are grouped as they are closely described as proficient and independent users of English in the CEFR Global Scale. They are to aim further to reach higher CEFR levels of English competencies.
2. The Band 4 students are considered independent users of English, and the curriculum aims to enhance their oral competencies further;
3. The Band 3 students need more effort and scaffolding to increase their oral competencies and confidence in using the language.
4. Those students below Band 3 require extra formal learning hours to be closely guided to build their confidence in oral competencies mainly.

Table 4

The Adaptation of Mapping Between MUET Results and CEFR Levels

MUET results	CEFR levels range based on oral competencies	English proficiency level
Band 6	B2–C1	Advanced
Band 5		
Band 4	B1–B2	Lower Intermediate to Upper Intermediate
Band 3	B1	Lower Intermediate

Table 4 *(Continued)*

MUET results	CEFR levels range based on oral competencies	English proficiency level
Band 2		
Band 1	A2–B1	Elementary to Lower Intermediate

Table 4 illustrates the adaptation of the mapping of MUET results against CEFR based on students' oral competencies:

Four pathway tracks are designed using this mapping as a guide. In each track, a series of courses are pitched at either mastery level or one level higher. In line with the aspirations of the Malaysian EL Roadmap 2015-2025 (Don et al., 2015), students can show incremental improvement and complete the tracks. Each track offers continuous progress of learning outcomes, tasks and assessments based on the CEFR descriptors (the 'can do' statements) to facilitate language acquisition and learning. This gradual progress is illustrated in Table 5.

Table 5

The UKM EL Curriculum Tracks

Track I	Track II	Track III	Track IV
Low Proficiency A2	Lower Intermediate Proficiency B1	Intermediate Proficiency B2	High Proficiency
MUET Band 1& 2	MUET Band 3	MUET Band 4	MUET Band 5 & 6

Implementation Phase for Low Proficiency Level Track I

This track represents the courses that are offered to students with MUET Bands lower than 3. These bands are equivalent to the CEFR range of A2 (basic user) to B1 (intermediate user) English. Table 6 illustrates the courses that are offered in Track I.

Table 6

Low Proficiency Level Track I

Low Proficiency Level Track I				
Course code and name	LMCE1042 *Breakthrough English*	LMCE1052 *Bridging English*	LMCE1062 *Academic Interactions*	LMCE 3051 *Let's Get Talking*
CEFR Level	A2	A2–B1	B1–B2	B2
Credit Hour	2	2	2	1

Source. Citra UKM (2019)

In this track, students have an additional two credit hours as compared to the other three tracks. The first two courses offered in this track are designed as introductory courses and expose students to university academic culture. Students begin their first year taking *Breakthrough English* in the first semester, a foundation level course that aims to enhance their competencies on vocabulary and language structures used in familiar and basic situations to build their confidence in using the language specifically. In line with the CEFR principle of empowering learners in using the language, this course allows students to attain a basic mastery of English. The learning outcome (LO) of this course is for students to communicate and interact with confidence on familiar topics that encompass receptive, production and mediation skills. It addresses the concerns raised regarding the common issue of the lack of confidence in using English among many university graduates in Malaysia.

The following course, Bridging *English,* is pitched at a high CEFR A2 that intersects with a lower CEFR B1. This course bridges basic English and using English in an academic setting. This course offers classroom tasks where students engage in writing and basic research tasks focusing on reading comprehension to establish their academic learning styles.

Academic Interactions is designed to assist students in engaging with EL texts, continuing the skills emphasised in *Bridging English.* It aims towards students achieving at least a high CEFR B1 to lower CEFR B2 at the end of the course. The classroom tasks are designed to enhance students' receptive skills of reading and interacting with texts, emphasising the production skills of communicating and collaborating in group discussions.

The students' last course in this track is *Let's Get Talking*, a communication course that prepares students to communicate more competently in English before they attend their internship training in Year 3 onwards. This course is pitched at CEFR B2 level. It consolidates the skills acquired in the earlier courses and allow students to perform communicative tasks in multiple authentic contexts.

Implementation Phase for Lower Intermediate Level Track II

Track II offers EL courses designed for students who obtained MUET Band 3, as summarised in Table 7.

Table 7

Lower Intermediate Level Track II

	Lower Intermediate Level Track II		
Course code and name	LMCE1062 *Academic Interactions*	LMCE2082 *Pro-Talk English*	LMCE3071 *Corporate Storytelling*
CEFR Level	B1–B2	B1–B2	B2
Credit Hour	2	2	1

Source. Citra UKM (2019)

The first course in this track is *Academic Interactions*, is the same course offered at the end of Low Proficiency Level Track I. It indicates the intersect between Track I and II. Following *Academic Interactions*, students proceed to Pro-*Talk English* (an abbreviation for *Professional Talk in English*) which is pitched at CEFR B1 level to prepare students for workplace communication. This course focuses on the workplace communication genre, further enhancing the skills acquired in the previous course. Students write emails, conduct meetings and pitch ideas in given workplace simulations. These are in line with the concerns raised on graduates' lack of workplace communicative abilities by industries.

Corporate Storytelling introduces students to the current trend of corporations informing the public about their products and values. It is pitched at CEFR B2, where students' communicative skills are reinforced in for public viewing group discussions and oral presentations. In addition, it will allow students to use a wider selection of word choices in their ability to review and present.

Implementation Phase for Intermediate Proficiency Level Track III

Track III comprises courses for students who obtained MUET Band 4 or those in a high CEFR B1 level and low CEFR B2 levels of proficiency, as illustrated in Table 8.

Table 8
Intermediate Level Track III

Intermediate Level Track III			
Course code and name	LMCE 1072 *Academic Literacy*	LMCE 2092 *Speak to Persuade*	LMCE 3071 *Professional Communication*
CEFR Level	CEFR B2	CEFR B2-C1	CEFR C1
Credit Hour	2	2	1

Source. Citra UKM (2019)

The first English course, *Academic Literacy,* aims to equip students with the language to function in the academic setting and become independent English users (CEFR B2 level). This course requires students to analyse multiple forms of texts used in their different fields of study, focusing on mediation class activities and assessments. Students are expected to

engage with texts critically, analytically selecting and using the information in forum discussions as their classroom activities and assessments. These activities reflect the tasks commonly used in the students' faculties, and the students can transfer the language and skills when dealing with other content courses. In addition, these tasks are designed to empower students as agents

of learning, where they take charge of the selection and analysis of texts, a principle strongly emphasised in CEFR.

In the students' second year of study, they proceed to *Speak to Persuade,* a public speaking course. This course targets students to achieve a high CEFR B2 level and move towards being proficient users of the language (CEFR C1). The course aims to empower students' confidence level in communicating a persuasive speech. This task is highly relevant, not just for academic purposes in the university but also in the workplace and other contexts. Students are given the responsibility to chart and mediate the speaking tasks, from selecting topics, researching information, and constructing the meanings in their speeches. These activities allow for creative and transactional language use, as suggested by CEFR, as the students use information in persuading the audience through their speeches.

In *Professional Communication*, students are exposed to the language and tasks that reflect the workplace contexts to prepare them for industrial training, internships and communication with people outside the campus. The focal point of the course is for students to apply the language and knowledge learned in the previous courses in a video production task. To complete this assessment task, students collaborate to uncover workplace issues and scenarios and produce videos to raise awareness. These activities should allow students to use analytical discourse in the video production process.

Implementation Phase for Upper Intermediate Level Track IV

The courses in Track IV are designed for students with MUET Bands 5 and 6 to enhance students' ability to communicate competently in any given situation. In addition, the activities in these courses allow the students to showcase their competence at CEFR C1 level and beyond. Table 9 details the list of courses under Upper Intermediate Level Track IV:

Table 9

Upper-Intermediate Level Track IV

Upper Intermediate Level Track IV		
Course code and name	LMCE 1082 *Page to Stage*	LMCE 2013 *Advanced Communication Project*
CEFR Level	C1–C2	C2
Credit Hour	2	2

Source. Citra UKM (2019)

In *Page to Stage*, students experiment with the nuances of EL in pronunciation, intonation and language patterns through reading a novel and portraying a character

through a play using literature. The course combines literature appreciation and interpretations into a stage play. An *Advanced Communication Project* course follows it. Students identify a project where they are expected to write a project proposal, a progress report and display their EL presentation skills in front of a large audience beyond the classroom walls. Students are encouraged to participate in any existing project within the faculties and campus. The track enhances students' appreciation and awareness of the language. It also opens the opportunity for a wider range of language use and knowledge transfer.

This curriculum structure is designed to integrate the CEFR elements in the Malaysian EL curriculum to align with the international standards outlined by the Malaysian EL Roadmap (Don et al., 2015). Therefore, it is significant to produce a Malaysian workforce to perform and to compete at global platforms.

CONCLUSION

UKM adopted the ADDIE model (Morrison, 2010) in restructuring its new English Language CEFR-informed curriculum. The primary objective of this curriculum is to elevate students' confidence in using English in various given contexts regardless of their levels of proficiency upon entry to the university. There were several immediate challenges faced when the curriculum was rolled out for the two semesters. One of the challenges faced during the curriculum implementation was the monitoring of

each course in each track. It was crucial to ensure that all instructors embraced the understanding of the new CEFR-informed curriculum. It is acknowledged that this is a gradual process to get the 'buy in' from the instructors to adopt CEFR in the delivery of their courses. The continual monitoring of the courses during the implementation phases of the curriculum enables the significant "hiccups" to be addressed immediately. In moving forward to strengthen the curriculum, UKM will initiate an ongoing review of each course towards CEFR—aligned tasks and activities to achieve the course objectives. The results from the ongoing reviews can be used to remap the materials and assessment scores to represent "true" CEFR can-do statements. This exercise is a crucial process to the next evaluation of the curriculum involving stakeholders from the industry. This exercise in framing the new EL in UKM within the CEFR is part of producing world-class graduates. It is framed based on the Malaysian EL teaching and learning experience. In the long run, this can be a 'mould' of reference for a Malaysian CEFR informed curriculum for tertiary education.

ACKNOWLEDGEMENT

their ongoing support and CITRA-2020-001 grant. Special thanks to Prof Dato' Ir Dr Mohd Marzuki Mustafa, the former Deputy Vice-Chancellor of Academic Affairs UKM and Prof Dr Hazita Azman, the Chair for UKM English Language Enhancement programme, for their initiatives, enthusiasm and guidance throughout this exercise. Thanks should also go, of course, to UKM for giving us this opportunity and supporting this initiative.

REFERENCES

Abdullah, M. H., Ho S. W., & Wong B. E. (2015). Delivering the ELEX Package: An English language experience approach for developing undergraduates' language proficiency. *Journal of Language and Communication, 2*(2), 283-296.

Ahmad, R., & Zainol, Z. A. (2011, March 14-16). What it takes to be a manager: The case of Malaysian five-star resort hotel. In *2nd International Conference on Business and Economic Research (2nd ICBER 2011)*. Langkawi, Malaysia.

Aldoobie, N. (2015). ADDIE model. *American International Journal of Contemporary Research, 5*(6), 68-72.

Arslan, A., & Özenici, S. (2017). A CEFR-based curriculum design for tertiary education level. *International Journal of Languages' Education and Teaching, 5*(3), 12-36. https://doi.org/10.18298/ijlet.1778

Balanyk, J. (2017, March 6-8). Developing English for academic purposes MOOCS using the ADDIE Model. In *Proceedings of INTED2017 Conference* (pp. 6514-6522). Valencia, Spain. https://doi.org/10.21125/inted.2017.1506

Chonghui, L. (2019, June 9). Boosting English standards. *The Star Online*. https://www.thestar.com.my/news/education/2019/06/09/boosting-english-standards#8HRz7DmHz13e3YvC.99.

Citra UKM. (2019). *Panduan Prasiswazah Pusat Citra Universiti: Sesi Akademik 2019-2020* [Undergraduate guide book. Academic session 2019-2020]. Pusat Citra UKM.

Council of Europe. (2017). *Common European Framework of References for Languages: Teaching, learning and assessments*. Cambridge University Press.

Darmi, R., Saad, N. S. M., Abdullah, N., Puteh-Behak, F., Zakaria, Z. A., & Adnan, J. N. I. (2017). Teachers' views on students' performance in EL proficiency courses via CEFR descriptors. *International E-Journal of Advances in Education, 3*(8), 363-370. https://doi.org/10.18768/ijaedu.336688

Dick, W., Carey, L., & Carey, J. O. (2011). *The systematic design of instruction* (7th Ed.). Pearson.

Don, Z. M., & Abdullah, M. H. (2019, May 2019). The reform of English language education in Malaysia. *Free Malaysia Today*. https://www.moe.gov.my/en/menumedia/printed-media/newspaper-clippings/the-reform-of-english-language-education-in-malaysia-free-malaysia-today-22-mei-2019

Don, Z. M., Abdullah, M. H., Abdullah, A. C., Lee, B. H., Kaur, K., Pillai, J., & Hooi, M. Y. (2015). *English language education reform in Malaysia: The roadmap 2015-2025*. Ministry of Education.

Drljaca, D., Latinović, B., Stankovic, Z., & Cvetković, D. (2017, April 21). *ADDIE model for development of e-courses*. In *International Scientific Conference on Information Technology and Data Related Research* (pp. 242-247). Singidunum University, Serbia. https://doi.org/10.15308/Sinteza-2017-242-247

Gustafson, K. L., & Branch, R. M. (2011). What is instructional design? In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and issues in instructional design and technology* (pp. 16-25). Pearson

Hazita, A., Aziz, M. S. A., Ismail, K., Wahi, W., Shamsuddin, C., Ya'acob, A., Mohamad, W. N. W., Othman, S., & Wong, F. F. (2018). *CBET: Competency based test CBET 2017 and 2018 results report* [Technical Report: KRA-2017-00]. Universiti Kebangsaan Malaysia.

Hsu, T. C., Lee-Hsieh, J., Turton, M. A., & Cheng, S. F. (2014). Using the ADDIE model to develop online continuing education courses on caring for nurses in Taiwan. *Journal of Continuing Education in Nursing*, *45*(3), 124-131. https://doi.org/10.3928/00220124-20140219-04

Morrison, G. R. (2010). *Designing effective instruction* (6th ed.). John Wiley & Sons. https://doi.org/10.4135/9781452232379.n4

NST Education. (2019, June 19). Preparing undergraduates for the workplace. *New Straits Times*. https://www.nst.com.my/education/2019/06/497514/preparing-undergraduates-workplace

Pandian, A., & Balraj, S. (2013) Ready for the workplace? English language and literacy skills among public university students in Malaysia. *Malaysian Journal of Languages and Linguistics, 2*, 56-71. https://doi.org/10.24200/mjll.vol2iss1pp56-71.

Razali, S. N., & Shahbodin, F. (2015). The development of an online project based collaborative learning using ADDIE Model. *Procedia - Social and Behavioral Sciences, 195*, 1803-1812. https://doi.org/10.1016/j.sbspro.2015.06.392

Sidhu, G. K., Kaur, S., & Chi, L. J. (2018) CEFR-aligned school-based-assessment in the Malaysian primary ESL classroom. *Indonesian Journal of Applied Linguistics, 8*(2), 452-463. https://doi.org/10.17509/ijal.v8i2.13311

UKM. (2019). *Bicara Perdana Naib Canselor 2019: Kualiti teras kebitaraan university* [Vice Chancellor's Inaugural Speech. Quality as the basis for university excellence]. The National University of Malaysia.

Uri, N. F. M., & Aziz, M. S. A. (2018). Implementation of CEFR in Malaysia: Teachers' awareness and the Challenges. *3L: The Southeast Asian Journal of English Language Studies. 24*(3), 168-183. https://doi.org/10.17576/3L-2018-2403-13

Yusof, K. M. (2015). In *English language education reform in Malaysia: The Roadmap 2015-2025*. Ministry of Education.

Zhang, J. (2020). The construction of college English online learning community under ADDIE Model. *English Language Teaching, 13*(7), 46-51. https://doi.org/10.5539/elt.v13n7p46

Zubairi, A. M., Sarudin, I., Nordin, M. S., & Ahmad, T. B. T. (2011). *English language competency for entry level employment: A market research on industrial needs*. Prestariang Systems Sdn Bhd.

The Use of Rhetorical Strategies in Argumentative Essays

Zulaikha Khairuddin[1], Noor Hanim Rahmat[2], Maizura Mohd Noor[1] and Zurina Khairuddin[3]*

[1]*Akademi Pengajian Bahasa, Universiti Teknologi MARA, 40000 Shah Alam, Selangor, Malaysia*
[2]*Akademi Pengajian Bahasa, Universiti Teknologi MARA, Cawangan Johor, Kampus Pasir Gudang, 81750 Masai, Johor, Malaysia*
[3]*Faculty of Languages and Communication, Universiti Sultan Zainal Abidin, 21300 Kuala Nerus, Terengganu, Malaysia*

ABSTRACT

The most challenging skill perceived by students when they learn the English language is the writing skill. This recent study would like to identify the rhetorical strategies used by good writers and poor writers. Two participants were selected, and written essays was the instrument employed for this study. Both participants were required to write an essay on 'Should examinations be abolished?' The essays written were analysed using a coding technique. The findings indicated that both writers utilised the three elements, Logos, Ethos and Pathos, differently. Both were considerate to the readers when they wrote the essays and presented their message, which was also heavily emphasised. However, they did not focus on their roles as writers. Based on the findings, it can be concluded that teachers need to help students familiarise themselves with rhetorical strategies. As for students, they should be aware of the rhetorical strategies to enhance their writing skills to write argumentative essays.

Keywords: Argumentative essay, Malaysian students, rhetorical strategies, think-aloud protocol, thinking process

INTRODUCTION

In learning English, students cannot avoid learning writing skills. A study found that students' perceptions towards learning writing skills are challenging (Badiozaman, 2017). They perceived learning writing skills as challenging and difficult because they were aware that need to consider many elements to write a good essay. Among the elements that they need to consider are syntax, semantic, and pragmatics (Mubarak, 2017). Students face difficulties in learning writing skills due to a low English proficiency. According to Pablo and Lasaten (2018), students with low level of English proficiency would perform poorly in their writing tasks. Poor writers and good writers can be defined in terms of the strategies used by both writers. A poor writer writes low-quality texts due to their inability to detect the errors, while a good writer uses their rhetorical and linguistics knowledge to write a better text (Ferrari et al., 1998). Maharani et al. (2018) further define poor writers as having a lower degree of awareness, belief and proficiency than good writers who have better purposes of learning language, motivation, degree of awareness, and learning style.

In the genre of writing, students have issues in writing argumentative essays because the essay's structure is different compared to the narrative, descriptive and compare and contrast essays (Amer, 2013). Therefore, for students to write a good argumentative essay, they need to consider suitable writing strategies. There are a few types of writing strategies: rhetorical strategies, meta-cognitive strategies, cognitive strategies, communicative strategies, and social/affective strategies (Mu, 2005). However, the most suitable writing strategies to be used in writing an argumentative essay are rhetorical strategies. It is because rhetorical strategies are used to ensure that the writers reach out to the readers, and the writers need to consider three elements when using rhetorical strategies which: Pathos, Logos and Ethos (Ramage et al., 2018).

To help the students to write good essays, the teachers need to guide their students with the correct and suitable writing strategies without ignoring the objective of the genre. As mentioned earlier, in teaching how to write argumentative essays, teachers should teach rhetorical strategies because rhetorical strategies are the used to persuade the readers to agree with the arguments made in the essay (Çam, 2015). However, the teachers themselves face difficulties and challenges in teaching argumentative essays. Students are still not equipped with the knowledge of writing where they lack vocabulary, do not master grammar or rule of syntax, the organization, and the mechanics of writing (Sujito & Muttaqien, 2016; Jumariati & Sulistyo, 2017). These issues should be addressed to help the students have better writing skills in writing argumentative essays. Thus, there is a need to investigate rhetorical strategies in argumentative essays (Cahyono, 2001).

Studies have been conducted on rhetorical strategies and argumentative essays; however, these studies focused

on non-Malaysian contexts in Germany (Wachsmuth et al., 2018) and Indonesia (Sujito and Muttaqien, 2016). Although Rahmat (2020) investigated the writers in a Malaysian university, she focused on the issues and challenges experienced by these writers in writing argumentative essays. She found that some of the challenges were: writer's anxiety, lack of opportunities to write, lack knowledge on punctuation (full stops, question marks, exclamation marks, and commas), language use (using synonyms/antonyms) and other writing skills (spelling, summarising, paraphrasing, in-text citation, and end-of text citation). Sujito and Muttaqien (2016) also identified differences in fast learners, medium learners and slow learners using rhetorical patterns in argumentative essays: fast learners managed to determine more critical ideas and give reasoning more logically than medium and slow learners. It shows that studies on rhetorical strategies in argumentative essays in Malaysian universities, particularly among writers with different proficiency levels, are scarce. Hence, the main objective of this study was to explore the use of rhetorical strategies by poor and good writers in writing argumentative essays. The research question for this study was: What rhetorical strategies are used by good and poor writers of argumentative essays?

Literature Review

Arndt (1987), Wender (1991), Victori (1995), Riazi (1997) and Sasaki (2000) argued for different strategies in writing essays in the context of language learners.

These studies mainly suggested that language learners employed different writing strategies, categorised in different categories except Wenden (1991) and Riazi (1997), who categorised the different writing strategies from a theoretical stance. However, Mu (2005) argues that the different categories could highly likely be confusing, particularly for language learners. Mu (2005) also suggests that effective writers employ the taxonomy of ESL writing strategies: rhetorical, meta-cognitive, cognitive, communicative and social/affective strategies. To reiterate, students must master relevant writing strategies suitable for different genres of writing. This study employed the theory of rhetorical strategy proposed by Ramage et al. (2016), and the supporting theories employed were Mu (2005), Larenas et al. (2017) and Nimehchisalem (2018).

Rhetorical strategies are related to Aristotelian theories of Logos, Ethos and Pathos. These three elements commonly focus on how a speech should be conducted (Lutzke & Henggeler, 2009). According to Aristotle's theory of Logos, Ethos and Pathos in a speech, Logos focuses on the message that speakers would like to convey. Speakers need to ensure that the message is clear and easy to understand by the audience. Ethos plays the same important role when speakers deliver their speeches as they need to establish their role and credibility in delivering the information or knowledge. Speakers need to ensure that the audiences believe in what they say. They need to build trust and rapport

with the audience. Lastly, Pathos serves a significant role for speakers because they need to grab the audience's attention to comprehend the message delivered. They can also relate the information provided with their schemata. It is crucial because if the audience cannot relate to the information, they may ignore the speech. These three concepts, Logos, Ethos and Pathos, apply to writers too. When writers write an essay or a composition, they need ensure that the message they want to convey is well-delivered to the readers. At the same time, they need to develop their credibility in writing good arguments supported with strong evidence. It is one of the ways writers initiate their role as credible and trustworthy writers. They also need to ensure that the readers understand the writing or essay. Hence, writers are suggested to write matters related to readers' background knowledge. From these explanations, the elements of Logos, Ethos and Pathos can be applied to both spoken and written forms of communication.

Mu (2005, p.3) and Mu and Carrington (2007, p.2) define rhetorical strategies as "strategies that writers use to organise and to present their ideas in writing conventions acceptable to native speakers of that language". Mu (2005) proposes four sub-strategies for rhetorical strategies: organisation, use of L1, formatting/ modelling and comparing. Mu further defines an organisation as 'beginning/ development/ ending' while L1 is defined as 'translating generated idea into ESL'. Modelling is then defined as 'genre

consideration,' and 'different rhetorical conventions' is the definition of comparing. Larenas et al. (2017) further add the sub-strategies: organising ideas, code-switching and translating. They also found that their participants employed different strategies before and after process-based writing intervention (thinking aloud protocol).

To reiterate, Logos, Ethos, and Pathos are applicable to be used in writing argumentative essays. Ramage et al. (2016) classify Logos as logical appeal, Ethos as ethical appeal and Pathos as emotional appeal. According to them, these three elements are called the rhetorical triangle (Figure 1).

Figure 1 describes that the three main elements, Logos, Ethos and Pathos, are interconnected suggesting that the triangle may not be complete if one element is missing. Therefore, the rhetorical triangle is symmetrical: all three strategies are significant in writing argumentative essays. Wachsmuth et al. (2018) suggest that mastering the rhetorical strategies would help writers persuade and convince the readers better. Hence, all these elements are crucial to be considered by writers. Wachsmuth et al. (2018) argue that writers synthesise the text using these three elements: selecting content in argumentative discourse units, arranging the structure and phrasing the style. Despite not using Logos, Ethos and Pathos elements (Ramage et al., 2016), and Abdullah et al. (2014) illustrate the importance of knowing and using rhetorical strategies in writing academic research as academic research has a similar
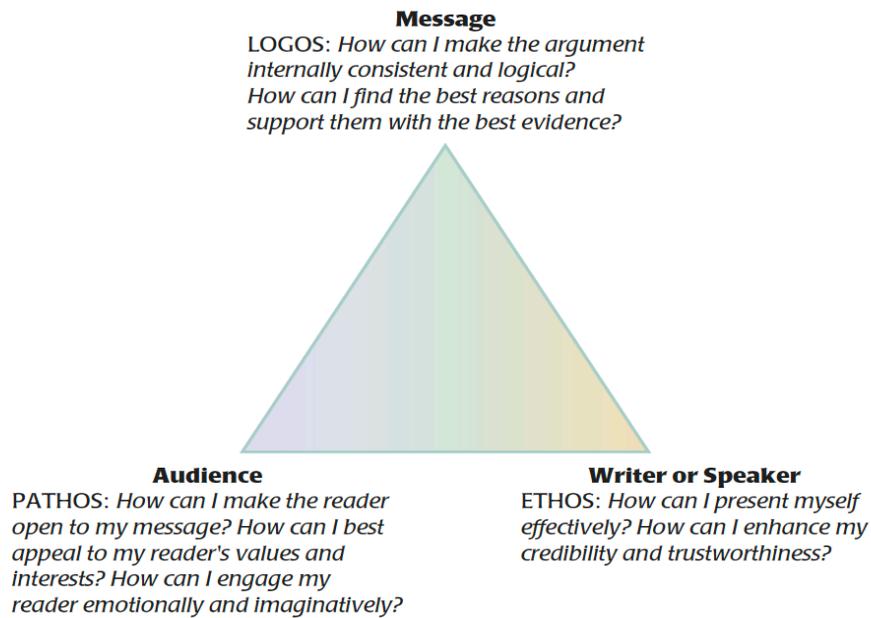
Figure 1. Rhetorical Triangle (Ramage et al., 2016, p. 55)

structure to argumentative essays (Ozfidan & Mitchell, 2020). However, focusing too much on one strategy may sway writers from their focus or purpose of writing argumentative essays. Therefore, they need to cover all three main elements and strategies in writing argumentative essays.

Ramage et al. (2016) stated that Logos, or the message, needs to be consistent and logical when the writer explains their writing. At the same time, the ideas need to be justified with strong support, and consequently, the ideas will indirectly appeal to the readers' needs. Abdullah et al. (2014) assert that writers need to ensure that the goal or purpose of writing essays is achieved where most statements must be well-explained, elaborated and supported with credible evidence. Abdullah et al.

(2019) and Campbell and Filimon (2018) supported this by suggesting students argue their ideas in their essays with strong support and evidence. For instance, Abdullah et al. (2019) found that writers who use more citations would use more rhetorical strategies than those who use fewer citations. In other words, when writers use more citations, they may be able to persuade readers to agree with their points and arguments, and indirectly, they use rhetorical strategy, logos to appeal to the readers to understand and agree with the arguments provided. Hence, they suggest that rhetorical strategy, logos, is used in writing argumentative essays as writers want readers to be attracted to read the essay and agree with the arguments made. Sujito and Muttaqien (2016) and Ahmad et al. (2019)

also discovered that students could not use the concepts of coherence and cohesion because they could not follow the flow of the argumentative essays. Consequently, they suggest that students will need some improvement, especially in connecting the ideas to apply the concepts of coherence and cohesion. Wachsmuth et al. (2018) suggest writers should use 70% Logos rhetorical strategy in their essay because messages are the most important part of writing the argumentative essay. They need to be logical in delivering the message and use their reasoning skills correctly.

The second element, Ethos, requires writers to be credible, and at the same time, they need to be seen as reliable and fair (Ramage et al., 2016). Nimehchisalem (2018) further defines Ramage, Bean and Johnson's ethos as good sense, goodwill, good morals. Even though the writers would like to persuade the readers to agree with their arguments, they also need to highlight the alternative views where readers could judge. Ethos is where writers need to ensure their reputation as wise and credible writers in delivering their arguments or thoughts. Wachsmuth et al. (2018) further argue that writers should use 10% Ethos in writing their argumentative essays.

Pathos requires writers to identify the intended audience before they write (Ramage et al., 2016). For example, if the marketing team wants to write an advertisement, they need to ensure that the words and phrases capture the readers' or the consumers' attention. The same method needs to be employed by writers,

such as students, who write argumentative essays. Abdullah et al. (2014) discover that writers need to consider the element of readership where the focus would be on the audience: whether they can comprehend and understand the information. In other words, students need to make their lecturers understand their arguments, and they also need to guarantee that the lecturers can relate to the arguments delivered with their schemata. The students can grab the teachers' attention by relating the points or arguments with teachers' values, beliefs and experiences. Wachsmuth et al. (2018) assert that writers need to include 20% Pathos in argumentative essays.

In writing argumentative essays, writers need to know how to utilise rhetorical strategies as the general structure of argumentative essays is almost similar to academic writing. Writers need to provide arguments where readers can comprehend and be attracted to the arguments provided. Rhetorical strategies used in argumentative essays are similar to the ones used by advertising companies for certain brands. According to Moore (2020), the most popular brands globally, such as Coca-Cola, Nike, John Deere, and Nivea, were listed as the 2019 world's most valuable brands. He discovered that the brands use rhetorical strategies in promoting and marketing their products. The most used strategy is Pathos, followed by Ethos and Logos. These brands used Pathos the most is because this element or strategy focuses on the audience, and it goes back to the purpose of the brand advertisement,

which is to promote their brand. In relation to writing argumentative essays, this is similar to studies done by Nguyen (2019) and Varpio (2018), which found that writers focused on the audiences when writing their argumentative essays. Similarly, writers need to use rhetorical strategies when writing argumentative essays. They need to understand the goal, task and targeted audience as it will help the writers achieve their purpose of writing argumentative essays. Thus, rhetorical strategies are suggested to be used in the teaching and learning of writing argumentative essays.

METHODOLOGY

Research Approach and Instrumentation

The research approach used in this research was a qualitative research approach utilising students' writing (essays) as the means to collect the data. In order to collect the

data, two (2) students were approached and requested to write an essay entitled 'Should examinations be abolished?'. After the students wrote the essay, the essays were analysed using a 'Rhetorical Strategies' descriptor which was adopted and adapted from Mu (2005), Ramage et al. (2018), Larenas et al. (2017) and Nimehchisalem (2018), as discussed in Rhetorical strategies are used in arguments to persuade the readers by using logical reasoning which affects the audiences' ethics and emotions (Wachsmuth et al., 2018). There are many rhetorical strategies used in argumentative essays; however, in this research, only three main elements were chosen: logos which focuses on the message, ethos which refers to the writer's credibility and pathos refers to the audience's emotions (Ramage et al., 2018; Nurjanah, 2016). The rhetorical strategies adopted and adapted are as follows:

Rhetorical strategy	Sub-strategy	Description of the strategy
Logos	ASWD	Arguments are supported and well-developed
	AUPA	Arguments are used to persuade the audience
	ASDA	Arguments are suitable to be comprehended by different audiences
	AIE	Arguments are stated implicitly or explicitly
	ALTV	Alternative views are given in explaining the arguments
	CGRE	The claim is supported with good reasons and evidence
	EUE	Evidence is used effectively
	CRDCE	Claims, rebuttals, and data are provided clearly and efficiently
	IOA	Ideas are organised accordingly (Beginning-Development-Ending)
	PRC	Problematic arguments are immediately changed to memorable arguments

Rhetorical strategy	Sub-strategy	Description of the strategy
Ethos	WC	The writer is credible enough to evaluate the topic/the issue
	WT	The writer is trustworthy to the intended audience
	TAS	The tone used by the writer is appropriate and suitable for the audience to understand the information/points/ideas
	WE	The writer has expertise in the field of discussion
	WPBV	The writer provides both views on the issue/the topic discussed (agreement vs disagreement)
	TLSGI	Text length and sentence complexity give impact to the presented arguments and the audience
	LWS	Language choices and word choices are suitable to be used
	PAC	Problematic arguments are immediately changed to memorable arguments
	CS	The writer is uses code-switching in explaining the information/points/ideas
	GA	Genre awareness can be identified in the text
	DRC	Different rhetorical conventions are used in the text
	TFGMW	The text focuses on good sense, good morals and goodwill
Pathos	TCV	The text gives the audience to connect with their values
	TPB	The text persuades the audiences to evaluate the arguments based on their beliefs
	TPOA	The text provides opportunities for the audience to make assumptions
	TTPA	The text focuses on the task, purpose, and audience

These sub-strategies were utilised when analysing the writings of the two participants, where a frequency of 880 was found.

Sample

The two participants for this study were purposively chosen as they were in their second semester, where they had learnt how to write argumentative essays in one of their courses during their first semester. The course required students to argue a topic with valid evidence by citing previous studies. The selection of participants was also based on their proficiency level, where their Malaysian University English Test (MUET) result was considered. Those with Band 4 and above are considered good writers, while those with Band 3 and below are considered poor writers. Their results in the courses as recommended by their lecturer were also a part of the selection criteria. Those who obtained B and below are considered poor writers, while those

obtaining B+ and above are considered good writers. Therefore, to be considered good writers, students must meet both selection criteria. In this research, the participants were required to write an argumentative essay entitled, 'Should examinations be abolished?' utilised by Zainuddin and Rafik-Galea (2016).

FINDINGS

The extracts in this section were taken verbatim where they were written without making corrections. The frequency of the strategies used by the writers was taken into consideration to determine whether a strategy is most often or least often used.

The Rhetorical Strategies Used by Good and Weak Writers in Argumentative Essay

Logos (Message). For Logos, it was found that both writers mostly used ASWD, AUPA, ASDA and AIE sub-strategies when writing the argumentative essays with the frequency of 88, 70, 86 and 55, respectively. In contrast, the least used sub-strategies for Logos are ALTV, CGRE, EUE, CRDCE and IOA, where the frequencies are 5, 10, 12, 13 and 25, respectively.

The Most Used Strategies.

Extract 1: ASWD - Arguments are supported and well-developed.

For the ASWD sub-strategy, the good writer used the strategy 56 times while the poor writer only used it 32 times.

Participant	Transcriptions
Good writer:	Examinations have been a practice in schools and universities throughout the centuries, as most of people have been through the examinations for years, people have been through the sleepless nights of preparation and memorizing facts about their subject, and it is believed that exams help students to enhance their knowledge capability
Poor writer:	Examination has been a method to measude one capability to understand and remember what has been taught by the teacher

Based on Extract 1, both writers did support and develop their arguments. However, it was not supported with strong evidence, and at the same time, it can be seen that the poor writer wrote a shorter sentence than the good writer. Therefore, the data could be interpreted to show that these two writers only used their background knowledge in supporting their ideas in writing argumentative essays. However, they should support their ideas with facts such as previous research or statistics.

Extract 2: AUPA - Arguments are used to persuade the audience

The good writer used the AUPA sub-strategy 51 times, while the poor writer only used it 19 times.

Participant	Transcriptions
Good writer:	According to Ferrer (2016), one of the major benefits of examinations is that they encourage students to learn.
Poor writer:	According to Vasugi (2019), school should abolish the exam because student will continue to get stressed even after the exam.

For the second extract, both writers used their arguments to persuade the readers or audiences by giving strong support with an in-text citation. Extract 2 suggests that writers were aware that they needed proof or strong evidence to support their points and ideas, indicating that they were familiar with the argumentative essay convention.

Extract 3: ASDA - Arguments are suitable to be comprehended by different audience

For the ASDA sub-strategy, the good writer used the strategy 55 times while the poor writer only used it 31 times.

Participant	Transcriptions
Good writer:	Apart from that, examinations will be able to help a student to know where their weaknesses is and it will motivate them to improve more on the next trial.
Poor writer:	Student make bad choice during and before examination by staying up late and neglect their daily need such as eating and drinking.

It can also be seen in Extract 3 that both writers used an argument that was suitable to be comprehended by different audiences. In addition, they gave examples that may be related to the audience for a better understanding of the argument or content they were making. It implies that writers consider their audiences or readers when writing the essay to ensure that the message or information is well-delivered.

Extract 4: AIE - Arguments are stated implicitly or explicitly

The good writer used the AIE sub-strategy 52 times, while the poor writer only used it 3 times.

Participant	Transcriptions
Good writer:	The reason why exams is a good way of assessments is that examinations will be able to improve the position of a weak student because of the disciplinary that they will apply to begin their success through the examination and after a few of trials and errors, they will not stand in the middle again.
Poor writer:	First and foremost, I think that examination should be abolish because exam does not show one's capabilities.

Based on the extract above, the good and poor writers managed to give arguments implicitly and explicitly based on their opinions. This extract indicates that good writers understood that their opinion should be explained further as it will help readers

have a better understanding of the points stated.

The Least Used Strategies.

Extract 5: ALTV - Alternative views are given in explaining the arguments

The good writer used the ALTV sub-strategy four times, while the poor writer only used it once.

Participant	Transcriptions
Good writer:	There are a few pros and cons of having examinations such as it can motivate students to study hard and it is a good way of assessments, however: there is also cons of having examinations such as affect a person's mental health.
Poor writer:	In my opinion, I partially agree that examination should be abolish because exam does not show ones capabilities, the pressure of performing well and exam make people better at the subject.

Extract 5 refers to whether the authors posted the ideas for both sides of the argument. Based on the transcript, both writers did discuss positive and negative points of the topic where both writers wrote them at the end of their introduction as their thesis statement. Therefore, this extract could be interpreted to show that both writers know the main structure of an argumentative essay where they included the thesis statement, which would help them to avoid writer's block.

Extract 6: CGRE - Claim is supported with good reasons and evidence

CGRE sub-strategy was employed 6 times by the good writer and 4 times by the poor writer.

Participant	Transcriptions
Good writer:	According to Ferrer (2016), one of the major benefits of examinations is that they encourage students to learn.
Poor writer:	According to Sani (2019), the student who have gone through the school system and graduated lacked lack the soft skill and critical thinking.

Extract 7: EUE - Evidence is used effectively

For the EUE sub-strategy, the good writer used the strategy 6 times while the poor writer also used it 6 times.

Participant	Transcriptions
Good writer:	According to Kocayörük and Telef (2015), exams can damage the happiness of a student. It is after seen in the community where a student intelligence symbolise the family's name. If they perform badly in their examinations, they will be seen as a disgrace to their family.

Poor writer: According to Ferrer (2016), examination is a great way to show their capabilities in the classroom.

Extract 8: CRDCE - Claims, rebuttals and data are provided clearly and efficiently

The good writer used the CRDCE sub-strategy 6 times, while the poor writer used it 7 times.

Participant	Transcriptions
Good writer:	According to Talib et al. (2018), exams are typically seen as a good way of testing course knowledge
Poor writer:	According to IT Learning and Development (2017) although the line between different form of assessments, test and evaluation are not always clear.

Based on Extracts 6, 7 and 8, both writers also used the evidence effectively to support their claims and arguments from reliable resources by giving reliable in-text citations even though some in-text citations did not follow the correct format. Both writers understood they needed to ensure that they support their statements with evidence and not solely based on their prior knowledge or opinions, especially when writing an argumentative essay. In other words, both writers were aware that they needed to argue and persuade readers by showing their credibility in providing the information.

Extract 9: IOA - Ideas are organised accordingly (Beginning – Development - Ending)

For the IOA sub-strategy, the good writer used the strategy 15 times, while the poor writer only used it 10 times.

Participant	Transcriptions
Good writer:	Thus, students are well-informed not only in the subjects that they are interested in, but also in the subjects that they find it difficult with.
Poor writer:	In conclusion, I partially agree that examination should be abolish.

Extract 9 shows that both authors organised their ideas accordingly, using discourse markers to guide the readers or the audiences to understand the text better. Thus, it implies that both writers understood discourse markers and knew that the markers would help readers comprehend the messages better.

Ethos (Writer). For Ethos, it was found that both writers mostly used three sub-strategies: WC (87 times), WT (88 times) and TAS (88 times), while the least used sub-strategies are WE (3 times), WPBV (5 times), TLSGI (4 times) and LWS (1 time).

The Most Used Strategies.

Extract 10: WC – Writer is credible enough to evaluate the topic/the issue

For the WC sub-strategy, the good writer used the strategy 57 times, while the poor writer only used it 30 times.

Participant	Transcriptions
Good writer:	However, some people believe that examination should not be treated as a tool to define a person's level of intelligence.
Poor writer:	This can be proven when the teaching become unbalance between teaching in a standard way and teaching more to the student skill.

Based on Extract 10, both writers showed that they were credible writers who wrote their ideas and opinions. However, it would be better to support their arguments and opinions with strong evidence such as past studies. It indicates that the writers focused only on their opinions showing their credibility in evaluating the issue. However, it may not satisfy the readers' needs to agree with the writers.

Extract 11: WT - Writer is trustworthy to the intended audience

The WT sub-strategy was used 57 times by the good writer, while the poor writer only used it 31 times.

Participant	Transcriptions
Good writer:	Thus, students are well-informed not only in the subjects that they are interested in, but also in the subjects that they find it difficult with.

Participant	Transcriptions
Poor writer:	Beside that, I think that examination should be abolish because examination cause pressure for the student to perform well.

From extract 11, both writers would also be considered trustworthy based on the given opinion on their discussed idea.

Extract 12: TAS – Tone used by the writer is appropriate and suitable for the audience to understand the information/points/ideas.

The good writer used the TAS sub-strategy 57 times, while the poor writer only used it 31 times.

Participant	Transcriptions
Good writer:	Examinations will encourage students to study and learn in a subject that they need to.
Poor writer:	It is another way of showing how much progress they have done and measure their understanding and the ability to apply it in the exam.

Based on Extract 12, the tone used by the good and poor writer is suitable for an argumentative essay as both gave relevant opinions on the topic.

Extracts 11 and 12 suggest that both writers considered their readers when writing the essay. The writers would ensure that messages are well-delivered to the

readers by explaining their opinions related to the readers' experiences and considering the tone of their writing.

The Least Used Strategies.

Extract 13: WE – The writer has expertise in the field of the discussion

For the WE sub-strategy, the good writer used the strategy 3 times while the poor writer did not employ the sub-strategy.

Participant	Transcriptions
Good writer:	There are a few pros and cons of having examinations such as it can motivate students to study hard and it is a good way of assessments, however: there is also cons of having examinations such as affect a person's mental health.

However, in terms of the writer's expertise, it can only be seen in the essay written by the good writer where the writer was clear with the topic discussed and understood the task and the issue that the writer had to complete.

Extract 14: WPBV - Writer provides both views on the issue/the topic discussed (agreement vs disagreement)

The WPBV sub-strategy was employed 3 times by the good writer, and the poor writer only used it twice.

Participant	Transcriptions
Good writer:	Examinations have been a practice in schools and universities throughout the centuries, as most of people have been through the examinations for years, people have been through the sleepless nights of preparation and memorizing facts about their subject and it is believed that exams help students to enhance their knowledge capability. However, some people believe that examination should not be treated as a tool to define a person's level of intelligence.
Poor writer:	In my opinion, I partially agree that examination should be abolish because exam does not show ones capabilities, the pressure of performing well and exam make people better at the subject.

Both writers also provided ideas on the issues of the topic discussed, referring to Extract 14.

Extract 15: TLSGI - Text length and sentence complexity give impact to the presented arguments and the audience

For the TLSGI sub-strategy, the good writer used the strategy 4 times, while the poor writer did not utilise it.

Participant	Transcriptions
Good writer:	The reason why exams is a good way of assessments is that examinations will be able to improve the position of a weak student because of the disciplinary that they will apply to begin their success through the examination and after a few of trials and errors, they will not stand in the middle again.

Based on Extract 15, it can only be seen that the idea in the good writer's transcription was written with a more complex sentence and lengthier.

Extract 13, Extract 14 and 15 indicate that only the good writer managed to fulfil these rhetorical strategies. However, the good writer managed to show her understanding of the task leading her to compose better ideas and points. Furthermore, the good writer also wrote longer sentences compared to the poor writer. It shows that a good writer has knowledge of different types of sentences which helps her write and combine ideas by writing more complex sentences and consequently help readers understand the issue better.

Pathos (Audience). It was found that both writers mostly used two sub-strategies: TCV (83 times) and TPB (83 times), while the least used sub-strategies are TPOA (55 times) and TTPA (49 times).

The Most Used strategies.

Extract 16: TCV - The text allows the audience to connect with their values

For the TCV sub-strategy, the good writer used the strategy 53 times, while the poor writer utilised it 30 times.

Participant	Transcriptions
Good writer:	Even though examinations stand as good as it seems, however, there is a negative effect from having examinations.
Poor writer:	They will worry about the score and also they will push themselves harder when they do not score well.

Based on Extract 16, both writers were writing by considering the readers' or the audiences' values.

Extract 17: TPB - The text persuades the audiences to evaluate the arguments based on their beliefs

The good writer also used the TPB sub-strategy 53 times, while the poor writer utilised it 30 times.

Participant	Transcriptions
Good writer:	One of the negative effects of having examinations is examination can cause mental health.
Poor writer:	This can be seen when the student get ther test paper, they can identify their weakness.

Other than that, in Extract 17, both writers also considered their readers' and audience's beliefs in composing their ideas.

Extracts 16 and 17 imply that both writers attempted to argue and persuade the audience and readers by highlighting the ideas and points that meet readers' values and beliefs. By doing this, the writers could assist the readers to understand the issue better as they could relate their values and beliefs to the issue.

The Least Used Strategies.

Extract 18: TPOA - The text provides opportunities to the audiences to make assumptions

The good writer also used the TPOA sub-strategy 49 times, while the poor writer only utilised it 6 times. It shows a big difference in the number of times both writers used the strategies.

Participant	Transcriptions
Good writer:	Exams can push students to be mentally ill because during day and night, they will have to study hard and everytime, only one job is reading, studying, reading and studying continuously.
Poor writer:	Student make bad choice during and before examination by staying up late and neglect their daily need such as eating and drinking.

Good and poor writers allowed the readers and audience to make assumptions based on the readers' point of view. The extract suggests that both writers let the readers decide whether to agree or disagree with the statements given. Consequently, readers were allowed to evaluate the issue independantly with some attempt of persuasion from the writers.

Extract 19: TTPA - The text focuses on the task, purpose, and audiences

The good writer used the TTPA sub-strategy 49 times, while the poor writer did not employ the strategy when writing.

Participant	Transcriptions
Good writer:	Students do not realise that pushing themselves to study hard will not do them good as they can perform poorly during the real examinations as they lack of focus and concentration during the test.

Extract 19 indicated that only the good writer focused on task, purpose and audience in writing the essay. It suggests that a good writer would consider all the three aspects when writing argumentative essays as it would help her be on track with the topic while writing the essay.

Discussion

Based on the findings of this current study, the researchers discussed the results according to each element of rhetorical strategies, i.e., logos (message), ethos (writer) and pathos (audience). For the first element, logos, writer must focus on the message, the issue or the argument that she

or he would like to highlight (Ramage et al., 2016). There are four Logos rhetorical strategies that were most used by the writers which were:

i. ASWD - Arguments are supported and well-developed
ii. AUPA - Arguments are used to persuade the audience
iii. ASDA - Arguments are suitable to be comprehended by different audience
iv. AIE - Arguments are stated implicitly or explicitly

Both the good and poor writers managed to support and develop their arguments. They were able to persuade their audiences by making sure that their arguments were comprehensible to different audiences. It is in accordance with research was done by Abdullah et al. (2019), Campbell and Filimon (2018) and Wachsmuth et al. (2018), wherein their research, ELL students, manage to write arguments using suitable strategies. It also suggests that both writers employed four of the Logos rhetorical strategies successfully, and they emphasised making their argument appealing to the readers as suggested by Ramage et al. (2016) by providing relevant evidence and information that is relatable to the reader. Both writers could also be interpreted as emphasising Logos as asserted by Abdullah et al. (2014): writers' arguments need to be well-elaborated and supported with substantial evidence.

There are five rhetorical strategies that the writers least used in writing their argumentative essays, which were:

i. ALTV - Alternative views are given in explaining the arguments
ii. CGRE - Claim is supported with good reasons and evidence
iii. EUE - Evidence is used effectively
iv. CRDCE - Claims, rebuttals and data are provided clearly and efficiently
v. IOA - Ideas are organised accordingly (Beginning – Development - Ending)

Despite not using much of these strategies, they elaborated their ideas from different perspectives. Both writers managed to support their arguments with reliable sources, even though there was an error in writing the in-text citation done by the poor writer. Both writers also managed to write their argumentative essays coherently and cohesively by using appropriate discourse markers, which is not parallel with the findings found by Ahmad et al. (2019), Rahmat (2020) and Sujito and Muttaqien (2016), where students were not able to follow the concept of coherence and cohesion. However, there are some grammatical errors done by both writers where the poor writer made more errors compared to the good writer.

Although both writers only utilise 40% of Logos rhetorical strategies in writing the argumentative essays, more emphasis on the message can still be seen. It could be interpreted similarly to what Wachsmuth

et al. (2018) suggest: the Logos element needs to be used more than other elements, suggesting that the good writer wrote the argumentative essay focusing more on the message to be delivered. Some differences between the poor and good writers were also found in the number of times some sub-strategies were used. For instance, there is a difference of 24 times between the writers for the Logos ASWD sub-strategy, where the good writer was found to utilise the sub-strategy more. Nonetheless, the difference is similar for other sub-strategies: AUPA and ASDA. One sub-strategy with a high difference in number is AIE, where the good writer used the sub-strategy 17 times more than the poor writer. However, the difference is small in the least used strategies, suggesting that both writers used these sub-strategies similarly.

The next element is Ethos, where the writers need to ensure that the audiences would be on the writers' side by showing their credibility (Ramage et al., 2016).

Both writers mostly used the following Ethos rhetorical sub-strategies:

i. WC - The writer is credible enough to evaluate the topic/the issue
ii. WT - The writer is trustworthy to the intended audience
iii. TAS - The tone used by the writer is appropriate and suitable for the audience to understand the information/points/ideas

It indicates that writers understood their roles in ensuring that the audiences believed in their arguments by showing that they were credible and trustworthy. At the same time, when reading the essay, the writers' tone was suitable to persuade the audiences to agree with the writers' arguments. It is similar to the finding of Nguyen (2019), whose study found that students from Thailand understood their roles as a writer.

However, it is found that the least Ethos rhetorical strategies used were:

i. WE - The writer has expertise in the field of the discussion
ii. WPBV - The writer provides both views on the issue/the topic discussed (agreement vs disagreement)
iii. TLSGI - Text length and sentence complexity give impact to the presented arguments and the audience

In providing the ideas for the issues for both views, both good and poor writers wrote it in the introduction as asserted by Ramage et al. (2016), who emphasise that writers need to provide readers with the opportunity to make their judgement by providing alternative views. However, in portraying the expertise and writing lengthier and complex sentences, these strategies were only shown by the good writer.

Both writers utilised 30% Ethos rhetorical strategies, which is different from the only 10% suggested by Wachsmuth et al. (2018). Vast differences were found in some sub-strategies used between the poor and good writers. For instance, there is a difference of 26 times between the writers

for the WT and TAS sub-strategies under the Ethos element, where the good writer was found to utilise both sub-strategies more. It could be interpreted as the good writer managing to show his/her trustworthiness as a writer as Ramage et al. (2016) and Nimehchisalem (2018) argue the importance of the writer's ethical appeal. Nevertheless, the difference is similar for other sub-strategies. However, in the least used strategies, the difference is how the good writer used the three sub-strategies: WE, TLSGI and LWS, but the poor writer did not, suggesting the relative difference in awareness of the strategies and the skills the different writers may possess. In other words, the poor writer's proficiency in writing argumentative essays is more prominent from his/her use of the Ethos rhetorical strategies or lack of, as he/she did not manage to ensure his/her credibility in arguing their thoughts well (Ramage et al., 2016). Similarly, Sujito and Muttaqien (2016) also found that writers' proficiency affects their ability to write argumentative essays well as they may not be able to incorporate rhetorical strategies successfully.

Finally, the third element is pathos, focusing on the readers as the audience. The most used Pathos rhetorical strategies were:

i. TCV - The text gives the audiences to connect with their values
ii. TPB - The text persuades the audiences to evaluate the arguments based on their beliefs

Both writers successfully utilised these sub-strategies well due to their awareness that other people would read their essays. In other words, they are aware of the need to provide better understanding for their intended readers, which is parallel to Abdullah et al.'s (2014) assertion that writers need to consider who their readers are. Hence, when they wrote their essays, they considered the audiences' values and beliefs so that the content would be suitable for intended readers (Ramage et al., 2016).

Nevertheless, the least used rhetorical strategies were:

i. TPOA - The text provides opportunities to the audiences to make assumptions
ii. TTPA - The text focuses on the task, purpose, and audiences

Despite Wachsmuth et al. (2018) suggesting that 20% of Pathos rhetorical strategies need to be used when writing argumentative essays, both writers utilised 30%. The poor and good writers were found to be different in their use of the sub-strategies under Pathos. For example, both writers used TCV and TPB sub-strategies in relatively similar ways. However, there is a difference in the number of times they employed the sub-strategies: 30 and 53, respectively. Both good and poor writers gave the audiences opportunities to make assumptions when they read the essay. It is important to consider readers when writers write a composition (Varpio, 2018; Ramage et al., 2016; Abdullah et al., 2014).

However, only a good writer focuses on the task, purpose, and audience when writing the essay. This is similar to what Sujito and Muttaqien (2016) argued, where poor and good writers use rhetorical strategies differently, with good writers able to be more critical and logical compared to poor writers. The findings suggest that both writers are different where the good writer used more TPOA sub-strategies than the poor writer six times. Other than that, the good writer used TTPA sub-strategies, and the poor writer did not, suggesting differences in knowledge and skills between the two writers.

CONCLUSION

In conclusion, based on the results, both good and poor writers were considerate towards the readers when they wrote the essays, and the writers were also focusing on the message they wanted to deliver. It means that as long as the message was delivered clearly to the audiences, the writers considered the essays as good essays. However, the writers were not paying attention to showing their credibility as writers when writing the essays. Nonetheless, the writers must ensure that they must consider all the three elements of logos (message), ethos (writer) and pathos (audience) if they want to write a better argumentative essay.

This study implicates teachers and students in their teaching and learning experience of writing. Teachers are encouraged to expose the students to correct rhetorical strategies to be used when writing argumentative essays. It will allow students to learn and consequently master how to use the strategies when writing and improve their critical thinking skills as the rhetorical strategies are related to thinking skills. The teaching of rhetorical strategies to students directly implicates the students writing experience. Students are encouraged to not only be familiar with rhetorical strategies but also master them and their usage. It is especially important for writing argumentative essays for the message to be delivered and explained successfully. Hence, this current research would suggest having more in-depth data by utilising interviews as the instrument for future research. Students must master and employ rhetorical strategies when writing essays regardless of the audience they are writing for (Warschauer, 2010). According to Nimehchisalem (2018), students' use of these strategies is personal and subjective, which could be different for each student. It suggests that teachers play a crucial role in designing activities that expose students to writing strategies. As students develop their own writer's profile, they would try different strategies to become effective writers of the English language.

ACKNOWLEDGEMENT

REFERENCES

Abdullah, I. S. M. M. A. B., Kamaruddin, H., Hashima, N., Yaacob, A., Rashid, S. A., & Desa, H. (2014). Rhetorical structures in academic research writing by non-native writers. *International Journal of Higher Education, 3*(1), 29-38. https://doi.org/10.5430/ijhe.v3n1p29

Abdullah, I. S. M. M. A. B., & Desa, H. (2019). Interpreting analysis on rhetorical

strategies modeling in computer science research articles. In L. Wang & L. Yu (Eds.), *Computer Architecture in Industrial, Biomechanical and Biomedical Engineering*. IntechOpen. https://doi.org/10.5772/intechopen.85679

Ahmad, M., Shah, S. K., & Mushtaq, M. (2019). Cohesion in argumentative writing: A case study of Pakistani essay writers. *Dilemas Contemporáneos: Educación, Política y Valores, 6*, 1-26.

Amer, M. (2013). The effect of explicit instruction in expository text structure on the writing performance of Arab EFL university students. *Arab World English Journal*, *4*(1), 224-238.

Arndt, V. (1987). Six writers in search of texts: A protocol-based study of L1 and L2 writing.

ELT Journal, 41(3), 257-267. https://doi.org/10.1093/elt/41.4.257

Badiozaman, I. F. A. (2017). 'I'm getting there': An investigation of academic writing developments of undergraduate students in a foreign branch campus in Malaysia. *Issues in Language Studies, 6*(1), 40-60. https://doi.org/10.33736/ils.476.2017

Cahyono, B. Y. (2001). Research studies in second language writing and in contrastive rhetoric. *k@ta, 3*(1), 39-52. https://doi.org/10.9744/kata.3.1.39-52

Çam, M. S. (2015). Aristotle is still persuading us: Content analyses about the rhetorical structure of magazine advertisements in Turkey.

International Journal of Multidisciplinary Thought, 5(2), 203-213.

Campbell, Y. C., & Filimon, C. (2018). Supporting the argumentative writing of students in linguistically diverse classrooms: An action research study. *RMLE Online*, *41*(1), 1-10. https://doi.org/10.1080/19404476.2017.1402408

Ferrer, D. (2016, October 9). *Pros and cons of tests and exams.* demscandoit. https://demscandoit.wordpress.com/2016/10/09/pros-and-cons-of-tests-and-exams/.

Ferrari, M., Bouffard, T., & Rainville, L. (1998). What makes a good writer? Differences in good and poor writer's self-regulation of writing. *Instructional Science, 26*, 473-488. https://doi.org/10.1023/A:1003202412203

IT Learning and Development. (2017). *Differences between testing, assessment, and evaluation.* http://tutorials.istudy.psu.edu/testing/testing2.html

Jumariati, J., & Sulistyo, G. (2017). Problem-based writing instruction: Its effect on students' skills in argumentative writing. *Arab World English Journal*, *8*(2), 87-100. https://doi.org/10.24093/awej/vol8no2.6

Kocayörük, E., & Telef, B. B. (2015, September 7-11). *The relations between attachment to parents and school happiness: Positive and negative experiences as a mediator* [Paper presentation]. European Educational Research Association Conference (ECER 2015 Education and Transition)*,* Budapest, Hungary.

Larenas, C. D., Leiva, L. R., & Navarrete, M. O. (2017). Rhetorical, metacognitive, and cognitive strategies in teacher candidates' essay writing. *PROFILE Issues in Teachers Professional Development, 19*(2), 87-100. https://doi.org/10.15446/profile.v19n2.60231

Lutzke, J., & Henggeler, M. F., (2009). *The Rhetorical Triangle: Understanding and using logos, ethos,*

and pathos. Indiana University. https://www.lsu. edu/hss/english/files/university_writing_files/ item35402.pdf

Maharani, S., Fauziati, E., & Supriyadi, S. (2018). An investigation of writing strategies used by the students on the perspective language proficiency and gender. *International Journal of Multicultural and Multireligious Understanding, 5*(5), 185-190.

Moore, S. (2020). Rhetorical strategies of Forbes' 2019 most valuable brands. *Elon Journal of Undergraduate Research in Communications, 11*(1), 64-73.

Mu, C. (2005). A taxonomy of ESL writing strategies. In *Redesigning pedagogy: Research, policy, practice* (pp. 1-10). QUTePrints.

Mu, C., & Carrington, S. (2007). An investigation of three Chinese students' Engish writing strategies. *TESL-EJ, 11*(1), 1-23.

Mubarak, A. A. A. (2017). An investigation of academic writing problems level faced by undergraduate students at Al Imam Al Mahdi University- Sudan. *English Review: Journal of English Education*, *5*(2), 175-188. https://doi. org/10.25134/erjee.v5i2.533

Nguyen, T. T. (2019). A case study of teacher feedback on Thai university students' essay writing. *GEMA Online® Journal of Language Studies*, *19*(2), 121-138. https://doi.org/10.17576/gema-2019-1902-08

Nimehchisalem, V. (2018). Pyramid of argumentation: Towards an integrated model for teaching and assessing ESL writing. *Journal of Language and Communication, 5*(2), 185-200.

Nurjanah, S. (2016). *A depiction of EFL learners' rhetorical strategies in their argumentative essays: A systemic functional linguistics perspective (A case study at IAIN Syekh Nurjati Cirebon)* (Bachelor Thesis). Syekh Nurjati State Islamic Institute Cirebon, Indonesia.

Ozfidan, B., & Mitchell, C. (2020). Detected difficulties in argumentative writing: The case of culturally and linguistically Saudi backgrounded students. *Journal of Ethnic and Cultural Studies, 7*(2), 15-29. http://dx.doi.org/10.29333/ejecs/382

Pablo, J. C. I., & Lasaten, R. V. C. (2018). Writing difficulties and quality of academic essays of senior high school students. *Asia Pacific Journal of Multidisciplinary Research*, *6*(4), 46-57.

Rahmat, N. H. (2020). Gender differences in rhetorical problems. *European Journal of Literature, Language and Linguistics Studies, 4*(1), 59-74.

Ramage, J. D., Bean, J. C., & Johnson, J. (2016). *Writing arguments: A rhetoric with readings*. Pearson Longman.

Ramage, J. D., Bean, J. C., & Johnson, J. (2018). *The Allyn & Bacon guide to writing*. Pearson.

Riazi, A. (1997). Acquiring disciplinary literacy: A social-cognitive analysis of text production

and learning among Iranian graduate students of education. *Journal of Second Language Writing, 6* (2), 105-137. https://doi.org/10.1016/S1060-3743(97)90030-8

Sani, R. (2019, February 20). Demand for soft skills in the workplace. *New Strait Times.* https://www. nst.com.my/education/2019/02/461884/demand-soft-skills-workplace.

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing, 9*(3), 259-291. https://doi.org/10.1016/S1060-3743(00)00028-X

Sujito, S., & Muttaqien, W. M. (2016). Rhetorical pattern in argumentative essay writing by EFL students of Iain Surakarta. *LINGUA: Journal of Language, Literature and Teaching, 13*(2), 157-168. https://doi.org/10.30957/lingua.v13i2.172

Talib, A. M., Alomary, F. O., & Alwadi, H. F. (2018). Assessment of student performance for course examination using Rasch measurement

model: A case study of information technology fundamentals course. *Education Research International, 2018*, 1-8. https://doi.org/10.1155/2018/8719012

Varpio, L. (2018). Using rhetorical appeals to credibility, logic, and emotions to increase your persuasiveness, *Perspectives on Medical Education, 7*(3), 207-210. https://doi.org/10.1007/s40037-018-0420-2

Vasugi, M. (2019, April 14). Should examinations be abolished? *The Star.* https://www.thestar.com.my/news/education/2019/04/14/should-examinations-be-abolished/

Victori, M. (1995). *EFL writing knowledge and strategies: An interactive study* [Unpublished Doctoral dissertation]. Universitat Autonoma de Barcelona.

Wachsmuth, H., Stede, M., Baff, R. E., Al-Khatib, K., Skeppstedt, M., & Stein, B. (2018, August 20-26). Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3753-3765). New Mexico, USA.

Warschauer, M. (2010). Invited commentary: New tools for teaching writing. *Language Learning & Technology, 14*(1), 3-8.

Wenden, A. L. (1991). Metacognitive strategies in L2 Writing: A case for task knowledge. In J. E. Alatis (Ed.), *Georgetown University Round Table on Languages and Linguistics 1991* (pp. 302-322). Georgetown University Press.

Zainuddin, S. Z., & Rafik-Galea, S. (2016). Effects of training in the use of Toulmin's model on ESL students' argumentative writing and critical thinking ability. *Malaysian Journal of Languages and Linguistics (MJLL), 5*(2), 114-133. https://doi.org/10.24200/mjll.vol5iss2pp114-133

A Systematic Review of Chinese Language Learning Strategies in the Past Decade (2011-2020)

Xi Mizhe[1,2], Ng Chwee Fang[1]*, Mohd Azidan Abdul Jabar[1] and Ilyana Jalaluddin[1]

[1]*Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[2]*Faculty of Literature and Communication, Xianyang Normal University, 712000, Weicheng District, Xianyang City, Republic of China*

ABSTRACT

Empirical studies and literature on Chinese language learning strategies (CLLS) in China and abroad have outlined theoretical introductions and case descriptions for nearly a decade. Reportedly, studies on CLLS indicated the following characteristics: The study respondents were primarily international students in China gearing towards regionalisation, nationalisation, or localisation. Furthermore, the qualitative study method followed an empirical, comprehensive, and descriptive learning strategy, such as observations and interviews. However, although the factors influencing learning strategies were gradually becoming diversified, several study limitations were identified (uneven regional studies, insufficient research samples, single research methods, and lack of theoretical paradigms in training learning strategies). Hence, researchers needed to conduct in-depth studies and deeply perceive CLLS to promote Chinese learning and teaching.

Keywords: CLLS, influencing factors, nationalisation, regionalisation, strategy training

INTRODUCTION

Studies on learning strategies began in the 70s. Influential scholars, such as Stern (1983), Rubin (1987), Oxford and Nyikos (1989), Oxford (1990), and O 'Malley and Chamot (1990), defined, categorised, and summarised second language learning strategies from different perspectives. Regardless, learning strategies in this

context primarily emphasised theoretical illustrations. Consequently, more empirical studies have emerged with the gradual progression and soundness of theoretical principles. Although research on foreign language learning strategies contributed to a broader comprehension of the relevant challenges, past studies were limited to European languages following the Roman alphabet (English, French, and Spanish). Unsurprisingly, CLLS-oriented research remained lacking. Concerning language type and internal composition, the Chinese language (Mandarin) was an ideographic language system involving unique symbols as opposed to Indo-European languages. Hence, the process of Chinese language acquisition varied from the phonetic system of the English language.

Foreign scholars researched the learning strategies of Chinese characters in the 1980s (Hayes, 1988), whereas a doctoral thesis studied the implementation of reading strategies in Chinese (Everson, 1986). In this vein, the studies above pioneered research on CLLS. To date, more CLLS-based studies, specifically the comprehensive study of learning strategies (Cao, 2010; Chen, 2008; Lin & Lv, 2005; Wu, 2007) and the utilisation of single-skill learning strategies by international students (Jiang & Zhao, 2001; Qian, 2006; Zhou, 2004), have gained due attention. Also, with the proposal of China's "One Belt and One Road" (Hong & Jieyan, 2013) strategy, China's influence in the international community is gradually enhanced, and the number of Chinese learners both in China and abroad

is increasing year by year. Therefore, the research achievements on Chinese acquisition are increasingly rich, especially in recent years, the research on Chinese learning strategies shows an obvious upward trend.

Despite extensive research on CLLS, attempts towards systematic study reviews remained scarce. As such, this study aimed to bridge the gap in comprehending and determining CLLS attributes and propensities. In this vein, a systematic review proved essential in identifying, choosing, meticulously evaluating pertinent studies and gathering and assessing the study data from the review. The systematic literature review enables the researcher to understanding the main issues and development trends of CLLS research in the past decade, which serves a more specific guiding significance for the said research area. It also helps to identify research gaps in the current understanding of CLLS. On the other hand, the systematic literature review can further optimise the training module of CLLS and improve the pedagogical effect of Chinese as a Second Language or Foreign Language. Consequently, the authors' research rigour, gap identification, and study directions required for future studies could be perceived. This research aimed to analyse CLLS study outcomes quantitatively and qualitatively in the past decade with statistical techniques to perform an in-depth analysis of current CLLS concerns and attributes (regionalisation, localisation, or nationalisation).

Generally, regionalisation denoted "societal integration and the often undirected process of social and economic interaction" Hurrell (2007, p. 4). In Wei (2012, p.151), Regionalised Chinese Teaching implied "to study Chinese teaching according to different geographical divisions, with a definite object in sight, to better improve the teaching effect". Additionally, Gan (2004) previously highlighted "nationalisation" incorporating various Chinese teaching and studies across different nations. Meanwhile, "localisation" encompassed the ideas, teaching content, teaching (teachers, textbooks, and teaching methods), and services associated with local settings (Li & Shi, 2017). On another note, Wu (2013) denoted that although "localisation" was occasionally interchangeable with "nationalisation", "nationalisation" distinctly indicated the country as a unit. This CLLS study outlined all three elements that would be duly illustrated.

METHODOLOGY

This paper uses the systematic review method to evaluate Chinese learning strategy research development in the past ten years. Systematic literature reviews use systematic and clear methods to select literature based on clear, stylised questions and critically evaluate relevant research (Petticrew & Roberts, 2006; Higgins & Green, 2008). This paper searches the literature on the study of learning strategies of Chinese as a second language or foreign language collected in China National Knowledge Infrastructure (CNKI) as well as Google Scholar, Scopus, and Proquest from 2011 to 2020 through keywords. It does a systematic review according to the research aims.

In the first step, the records determined through the database search were chosen if their titles complemented the specific study topic. In the second step, chosen record references and topic meta-analyses and reviews were sought for further records. The third step implied screening the selected record abstracts (to be omitted if the selection criteria were unmet). Lastly, the full-text articles were evaluated for eligibility. Notably, the articles were only incorporated into the review if the following selection criteria were fulfilled: (a) the study subject merely encompassed students who learned Chinese as a second or foreign language in China or abroad; (b) The literature only encompassed empirical CLLS articles (review articles, book series, books, book chapters, and conference proceedings were omitted); (c) The article should be peer-reviewed; (d) only literature between 2011 and 2020 were utilised.

Summarily, 1563 articles appeared to meet the criteria (based on the titles and abstracts). Specifically, 1316 were omitted following title and abstract reading. Lastly, only 155 articles that fulfilled the search criteria were presented (Figure 1). Following the literature analysis, specific CLLS-based attributes and limitations from the past decade and potential CLLS counterparts required due regard.
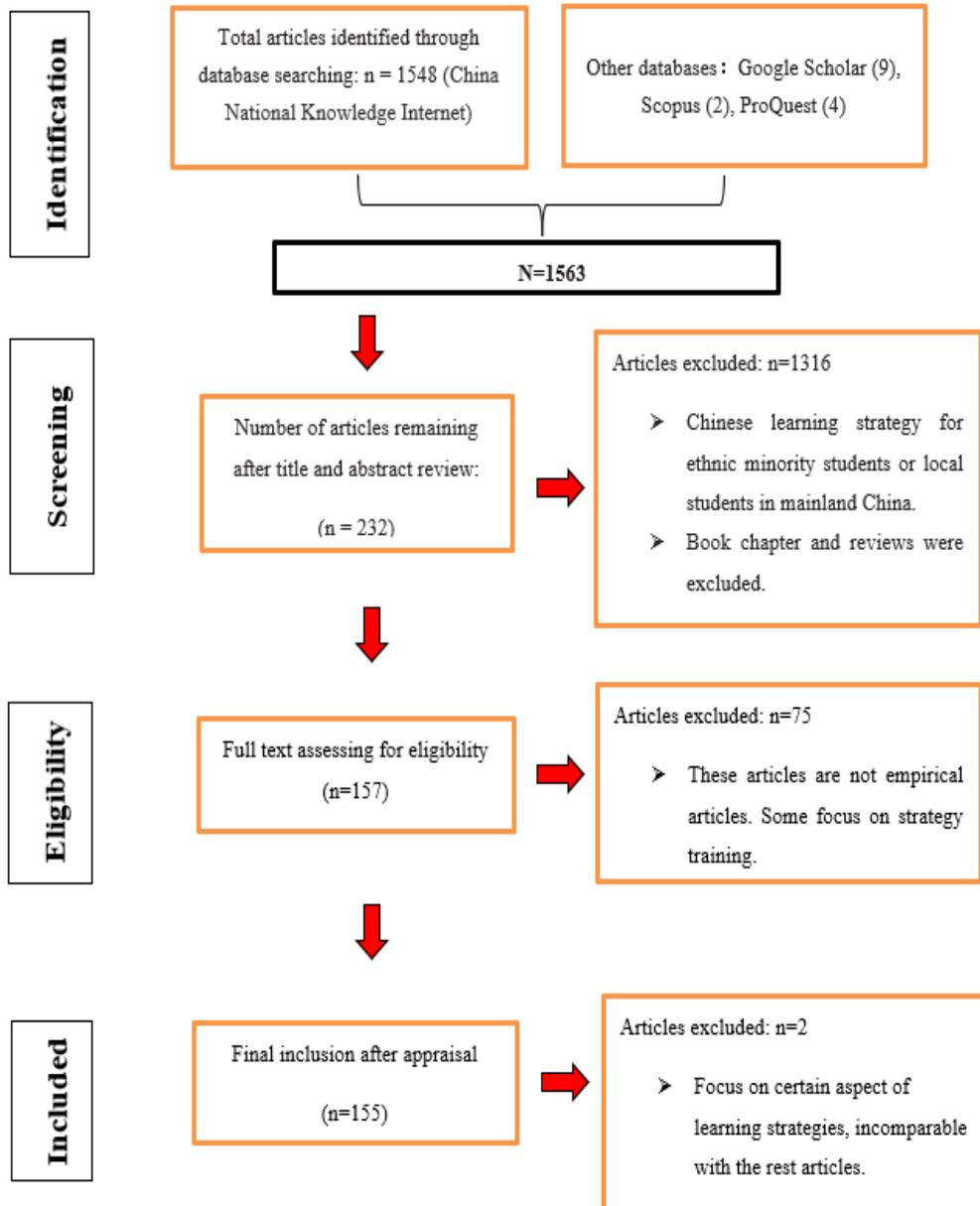
Figure 1. PRISMA flow chart for systematic literature review

RESULTS

For almost three decades, CLLS-related studies were primarily conducted by local Chinese scholars. Past research mainly emphasised the advent of learning strategy theories abroad. Specifically, most of the CLLS study results concerned international students in China. With the improvement of comprehensive national strength in China, the number of people learning Chinese worldwide is gradually escalating. Likewise, Chinese language teaching was also becoming more diversified. The tremendous increase of Chinese learners on a global scale subsequently widened the vision and scope of the research. Figure 2 presents the development trend of CLLS articles over the past decade.
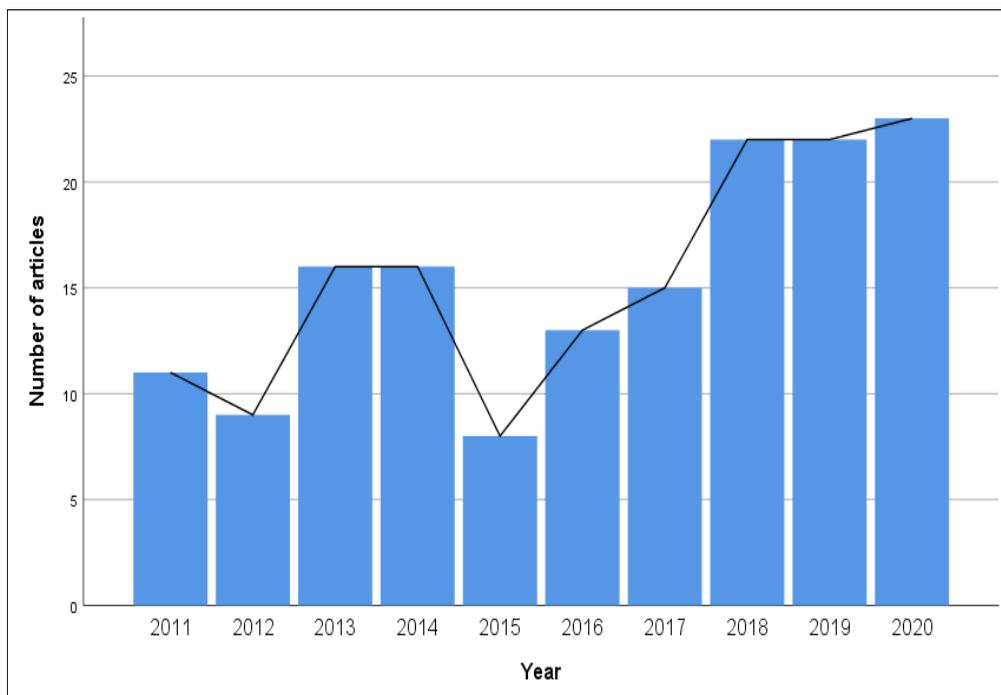
Figure 2. The CLLS articles by year (CNKI)

Figure 2 presents a decade long CLLS research development pattern. Resultantly, CLLS scholars between 2011 and 2020 demonstrated an overall ascending trend (excluding a slight decline in 2012 and 2015). Notably, 22 and 23 CLLS articles were published in 2019 and 2020, respectively (the highest number of CLLS articles published in the past ten years). Thus, the study indicated the rising popularity of CLLS-oriented studies among researchers.

On another note, the study results demonstrated the following characteristics: First of all, at present, the research on Chinese learning strategies is mainly focused on the background of the target language environment. Secondly, with China's comprehensive national strength,

the research on the nationalisation and regionalisation of Chinese learning strategies is gradually increasing, showing an upward trend. Finally, the research on influencing factors of Chinese learning strategies is becoming more extensive and in-depth.

Analysis of CLLS among International students with Diverse Cultural Backgrounds

Regarding CLLS-oriented studies, much research selected international students in China as the study respondents and performed comparative analyses on learner differences strategy implementation. In the study context, 95 out of the 155 CLLS articles belonged to the Chinese setting and accounted for 61.29% of the total number of articles (Figure 3).

Most studies comprehensively depicted the CLLS of foreign learners from a macro perspective (Cai, 2014; Cui & Yan, 2011; Guo, 2019; Liu & Yuan, 2017; Lv, 2013; Wang, 2015; Xu, 2018; Yang, 2012). Based on current articles, most scholars selected study measures (following local conditions) and samples (international students in specific regions) to demonstrate similarities and differences. Chinese students' learning strategies were also compared to various native language contexts as Chinese language teaching in China mainly involved mixed-class teaching with students from diverse cultural backgrounds. A survey on Thai and American students' learning strategies reported that Thai students' choice of learning strategy was more diverse (Chen, 2013). Specifically, both Thai and American Chinese students often implemented social strategies. Additionally, American learners utilised less memory and emotional strategies. Given the variances
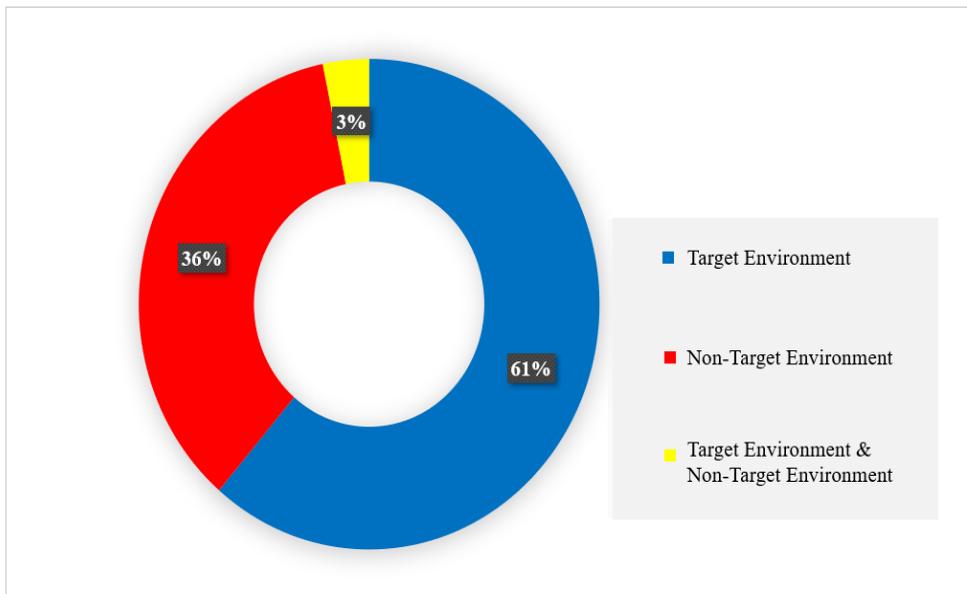
Figure 3. The proportion of articles on different language learning contexts (CNKI)

between Thai and American learners' choice of learning strategy, bias was a consistent factor. Most students preferred social, metacognitive, and compensation strategies, whereas the least preferred strategies were cognitive and memory. Although the finding corresponded to Jiang (2000), the results differed from Wu (2007). Additionally, Yu & Huang (2016) revealed that African students' metacognitive and memory strategies among were remarkably higher than the overall data. In contrast, compensation, emotional, social and cognitive strategies were significantly lower than the overall data. Based on a t-test using 103 independent samples (second language learners of Chinese), it was reported that African students used various learning strategies compared to second language learners of Chinese.

Comparisons on learning strategy utilisation in various cultural backgrounds and linguistic contexts denoted one of the studies attributes in this period. For example, L. Zhou (2013) compared Thai students' learning approaches in the target language environment against non-target language counterparts. Consequently, the overall CLLS frequency incorporated by primary-level Thai students within the target-language context proved higher than non-target language counterparts. Meanwhile, Zhang (2012) compared learners CLLS utilisation in different contexts during and post-class. Additionally, Li (2016) compared Indian and Filipino students as second language Chinese language learners. Furtherly, some other studies on students'

learning methods between Chinese and non-Chinese character circles were compared with writing and reading Chinese characters (Ding, 2018; Liu, 2018; Wang, 2018). For instance, Huang (2018) compared Thai students listening methods within two linguistic settings. The comparative studies encompassing students from distinct cultural backgrounds: showed that different cultural backgrounds and mother tongues impacted learning strategy alternatives. In this vein, Chinese language teaching should duly consider learners' cultural backgrounds and mother tongues.

Meanwhile, the foreign learners in China demonstrating characteristics of expertise selectivity, and the foreign learners who primarily studied traditional Chinese medicine reflected different CLLS. For example, Liu et al. (2019) examined South Asian medical students in China and reported that the overall CLLS usage varied from ordinary students (high to low) and the frequency of metacognitive, social, compensation, cognitive, memory, and effective strategies. Regardless, it was revealed that Western learners were prone to holistically implement compensation and memory strategies (Jiang, 2011). Additionally, B. B. Li (2014) indicated that respondents often utilised metacognitive as opposed to memory strategies. Resultantly, the similarities and variances of learning strategies among medical students in China reflected that most medical learners tended to use metacognitive compared to memory strategies. In this vein, the internal influencing factors could be

identified from the differences. Based on the similarities, the medical students in China inadequately implemented memory strategies, consequently providing ideas for the next level of strategy teaching.

The Country-Specific Distribution of CLLS Research

Given the rapid internationalisation of the Chinese language, more countries were involved in Chinese language teaching. According to CLEC (2020), there have been 541 Confucius Institutes and 1,170 Confucius classrooms set up in 162 countries and regions. Consequently, studies on nationalised Chinese teaching have gained researchers' attention. As each country had various national conditions, languages, and culture, the similarities or differences in students' learning processes were worth studying. Based on the literary analysis, studies on nationalised CLLS currently consisted of 34 countries on five continents. Specifically, Thai, American, and South Korean students were the most studied respondents (Figure 4).

In Thailand, different learning strategy groups were studied with varying conclusions. For example, Zhang (2014) assessed Thai high school students' learning strategies and discovered that learning strategies were commonly implemented in Chinese language learning. Specifically, social and memory strategies were commonly used by learners. Likewise, J. Li (2014) performed a CLLS survey among Thai students in Sarawittaya (Sara) Middle school, Thailand. It was revealed that the

students frequently implemented cognitive and social strategies, whereas the choice of memory strategies was less commonly used and lacked methods, thus contradicting the past results. As additional CLLS research on different learning groups in Thailand (Zhang, 2017; Lin, 2016; Zhou, 2019; Li, 2018; Zheng, 2014; Lu, 2012) varying results in strategy use, further CLLS-based studies proved necessary. Researches on CLLS among American students were also conducted. A survey on American students taking short-term classes at Zhejiang University indicated that compensation and social strategies were often implemented by American students, whereas memory and emotional strategies were infrequently used (J. Wang, 2011). However, the study findings slightly varied from Zhang (2011). For example, Zhang (2011) performed a survey of Chinese learners in two American universities and reported that metacognitive and social strategies were most commonly implemented by American students, whereas compensation and emotional strategies were the least commonly used. Thus, learners with the same language and cultural backgrounds portrayed particular variances in the choice of CLLS various influencing factors, such as learning, target language, and non-target language environments, as discussed below.

Apart from research on learners from a single country, the study of regional learning strategies in a particular region was also emphasised. For example, Yuan et al. (2011) investigated the relationship between the CLLS, attitudes, and motivations of Southeast Asian students. It was reported
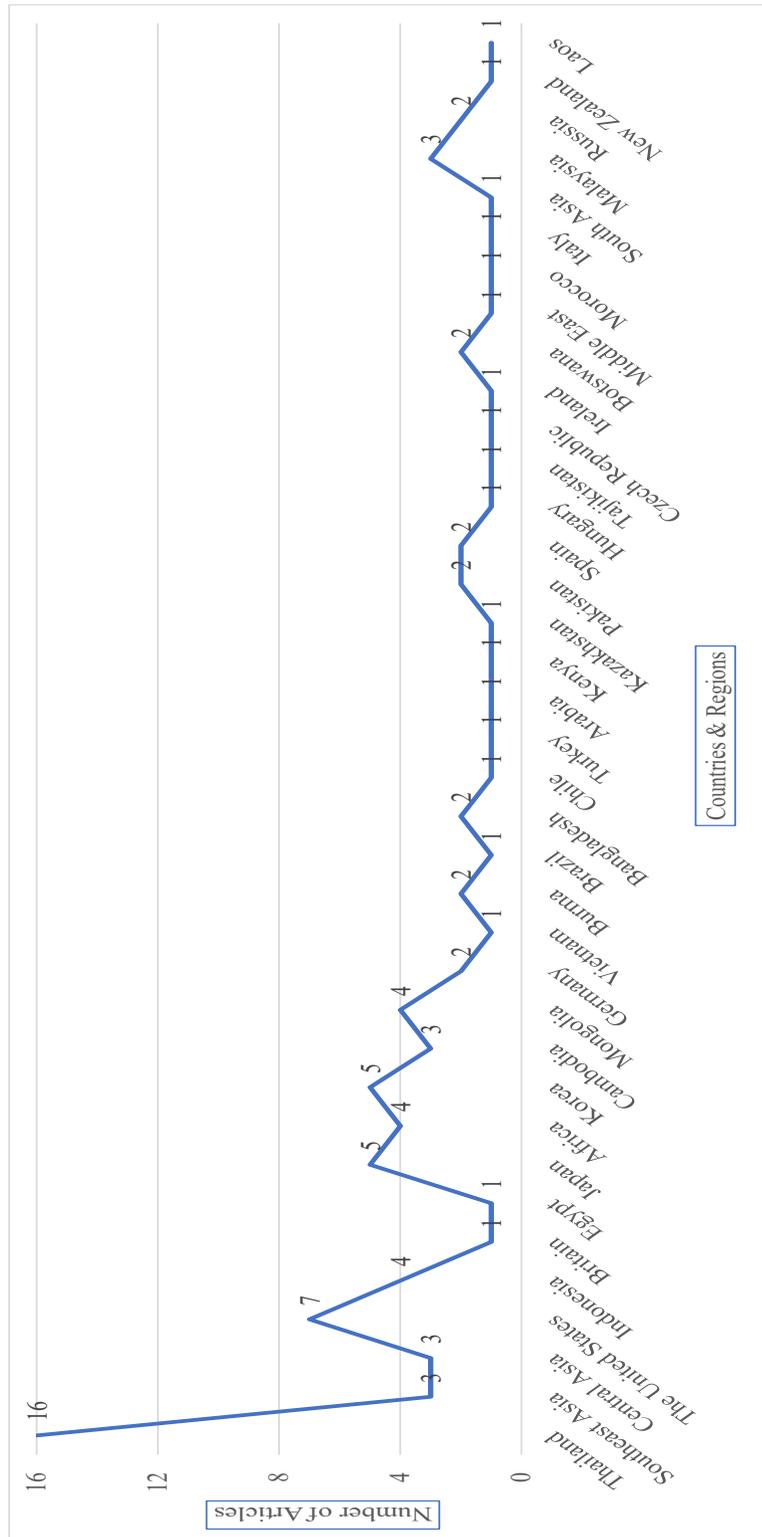
Figure 4. The distribution of CLLS research in single language backgrounds

that the most frequently employed CLLS by Southeast Asian students were emotional strategies with a significant correlation between learning strategies, attitudes, motivations, self-evaluations. Regardless, a study on the CLLS of students from five Southeast Asian countries indicated that social and metacognitive strategies were most commonly employed, whereas memory and emotional strategies were least often implemented (H. Li, 2014). Regional studies on the Chinese language also involved specific areas, such as the Middle East (Ma, 2019), Central Asia (Cui &Yan, 2011; Liu, 2013; Wang, 2013; Wang & Li, 2013; Zhang & Wan, 2019), Africa (Luo, 2019; Wu, 2018; Yu & Huang, 2016; Zhao, 2012). Given that different area involved different usage of learner strategies, studies on CLLS implementation and the influencing factors of various regions regarding Chinese language teaching and cross-cultural communication had specific practical implications. Overall, country-specific research on learning strategies significantly facilitated TCFL (Teaching Chinese as Foreign Language). The results reported that different countries reflected different learning strategy alternatives. Thus, strategy training for students from different countries was proposed.

Research on the Relevant Factors Affecting CLLS

Due to individual differences and environmental factors, students demonstrated different learning strategy alternatives and implementation, hence

resulting in palpable differences regarding learning effects in the same environment. Concerning the influencing factors of learning strategy classifications, Ellis (1994) stated that students' factors could be categorised into background factors (Chinese proficiency, duration of learning Chinese, Chinese score levels, and the number of students learning a foreign language) and personal factors (age, learning style, learning motivation, learning anxiety, and tolerance of ambiguity). According to studies on domestic and foreign researchers on the factors influencing learning strategies, the external learning context was a key determinant that could be classified into the classroom and social contexts. The classroom context mainly denoted teachers, textbooks and teaching tasks, whereas the social context primarily denoted the mother tongue and target language environments.

Apart from past studies, more research on the relationship between individual factors and learning strategies has been conducted. Although most works of literature emphasised individual factors (gender, age, learning level, and learning period), the influence of gender on learning strategies was deemed controversial. Nevertheless, some studies revealed that gender significantly affected CLLS usage. For example, P. Wang (2016) surveyed on 78 Brazilian students reported that male students often utilised learning strategies than their female counterparts. Contrarily, J. Wang (2011) revealed that women employed learning strategies more often than men in correspondence to Lin (2016),

Ou (2018), T. Xie (2018), Wang, 2015, Zhao (2012), Zhang (2014), and Zhang & Ge (2015). Furthermore, Gu (2014) indicated significant variances between genders regarding selecting memory, cognition and social strategies. Similarly, some studies also concluded that gender influenced strategy selection (Chen, 2017; Li, 2017; Li, 2019; J. Q. Wang, 2011; Zhang, 2017; Zhai, 2019; Zhu, 2018). Regardless, some other study results reported that no significant correlations between gender and learning strategy were identified (Li, 2019; Luo, 2019; Y. Wang, 2017; Zhang & Wan, 2019).

Other factors also impacted learning strategies. For example, it was indicated that learning time remarkably affected learning strategy choices (Zhai, 2019), learning concept was positively correlated to learning strategy implementation (Xie, 2019), and learning motivation was significantly correlated to learning strategies (Zhou et al., 2014). Additionally, Sheng (2019) examined the correlation between ambiguity tolerance and learning strategies among Burmese learners and indicated that ambiguity tolerance and learning strategy implementation frequency was negatively related to gender among primary Chinese language learners. In contrast, the intermediate Chinese learners' ambiguity tolerance and learning strategy use frequency reflected a positive correlation. Yang (2019) also indicated that social identity, other second language learning experiences, and the fundamentals of the Chinese language were significantly correlated to learning strategy usage. Chen (2017) pointed that

gender, age, nationality, learning duration, and other factors influenced the alternatives and implementation of learning strategies for international students. For example, Yang (2015) indicated a positive correlation between students' self-evaluation and CLLS in various grades and years of study.

On another note, Guo (2019) mentioned that the disposition of international students directly affected the choice and implementation of learning strategies. Furthermore, Wang (2015) demonstrated that students' utilisation of learning strategies pertained to teaching methods. In a survey on international students' learning strategies at Hebei University, Y. Wang (2017) revealed that students did not generally have a clear understanding of learning strategies (consciously or unconsciously) as teachers paid little attention to international students' learning strategies. Zhang and Ge (2015) examined the interaction between Mongolian students' factors (majors, learning years, and CLLS). It was demonstrated that Chinese language proficiency and the ability to mould international students influenced learning strategies. Although researchers generally affirmed that several factors influenced learning strategies, studies on the area above remained lacking.

Although sufficient research exists on students' factors, the influences of environmental factors (social, cultural, and teaching environments) and learning strategies remain disregarded. Hence, in-depth explorations on the different factors influencing CLLS usage should

be conducted for theoretical and practical guidance to improve the effects of teaching and to learn the Chinese language.

Conclusively, CLLS studies are rapidly increasing on a global scale. Furthermore, research on learners from various countries and regions highlighted a novel CLLS pattern in this period. Studies on the key determinants of CLLS is also gaining more depth and cohesion with emphasis on individual and environmental aspects. In this regard, the research facilitated a relatively objective and sound understanding of CLLS with a basic knowledge of CLLS attributes with various cultural backgrounds. Consequently, CLLS training proved relevant as a significant reference to teaching Chinese as a second and foreign language.

DISCUSSION

For almost a decade, CLLS studies reflected an overall upward trend. In this regard, scholars are gradually focusing more on the students' "learning" process. Current dissertations and literature reviews also emphasised studies on a single country, multi-angle studies on learning strategies and relevant influencing factors and presented specific similarities and differences. Regardless, research on CLLS as a second language reflected several deficiencies.

First, an incongruence between nationalisation and regionalisation studies involving CLLS was identified. Studies on international students' CLLS in certain countries are gaining momentum following researchers' focus on CLLS among students from Central and Southeast Asia, Europe and America, East Asia, and Africa. However, studies on students from South Asian countries revealed uneven results. For example, more CLLS studies involving Thailand were identified compared to other countries (Yang, 2015; Zhang, 2017; Zheng, 2014). In contrast, CLLS-based studies in Malaysia only amounted to three relevant articles (Chen & He, 2017; Cheng, 2018; Xu, 2018), whereas only Chen and He's (2017) study consisted of local students. Hence, the regionalisation and nationalisation of learning strategies needed to be established.

Secondly, the study samples were small with single research methods. Although studies on learning strategies in Master's and Doctoral theses and dissertations have been annually increasing, scholars were too constrained in sample selections. Based on the statistics, 34 out of the 134 theses samples were lower than 50 (Figure 5).

The research methods were also too simplified. In current works of literature, the main research methods employed quantitative and descriptive analyses. Moreover, the data mining process was not in-depth, thus failing to resolve the fundamental issue. In contrast, only several articles were qualitatively analysed (He, 2011; Jiang, 2011; Peng, 2017; Zhang, 2015; Li, 2017) or employed a hybrid research method (Du, 2018; Ding, 2018; Kuo, 2015; Liu, 2012; Lin, 2016; Ning, 2019; Shu, 2013). Although research methods on domestic learning strategies employed quantitative interviews, many studies did not disclose the interview content and findings
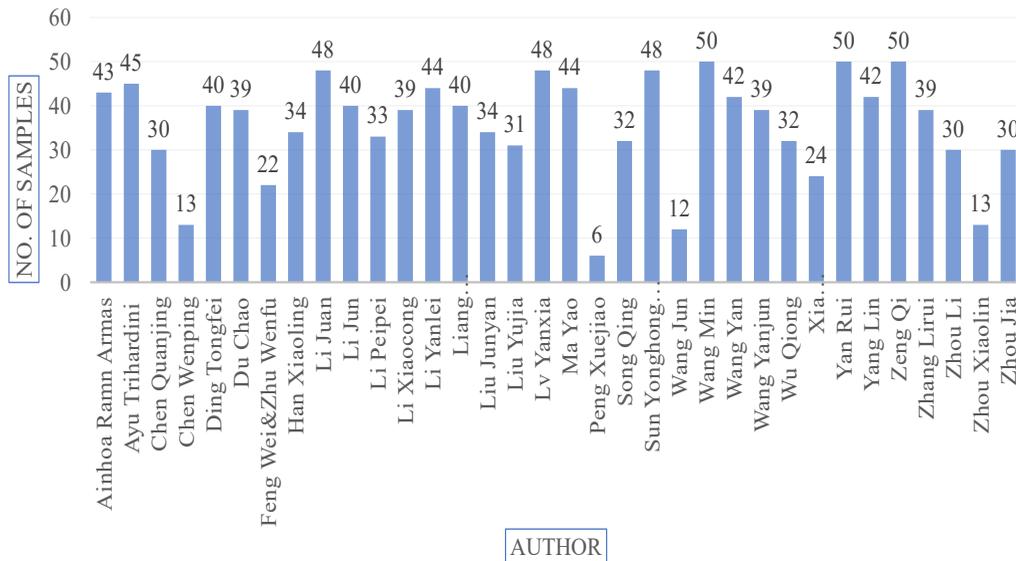
Figure 5. Number of samples

(L. L. Zhang, 2013; Li, 2016), which only utilised a single research method.

Regarding CLLS strategy training, domestic CLLS training was included in the final part of the literature. Specifically, researchers' proposals were derived from personal teaching experiences and methods (Li, 2017; Liu, 2012; Tan, 2018). The lack of a systematic teaching mode that integrated theory with practice insufficient policy-based intervention studies led to the incomprehensiveness of current empirical studies on CLLS training. Hence, it was vital to explore adequate language training modes to guide Chinese language teaching and learning.

CONCLUSION

By categorising literature on CLLS (locally and globally) in the past decade, it was reported that researchers primarily emphasised international students' learning strategies in different cultural backgrounds. The study also reflected that students with different cultural backgrounds selected different CLLS.

Based on the review of the prior research, there is a population gap. Studies on the learning strategies of Chinese learners with a single cultural background was a relatively new trend. Given the improvement of Chinese language internationalisation, many nations have included Chinese as an important foreign language. Consequently, studies on the nationalisation and regionalisation of CLLS was crucial for the effective acquisition of the Chinese language.

Also, a comprehensive and in-depth analysis of the influencing factors of students' CLLS in different countries and regions was carried out to guide Chinese

learners' learning process and methods from different backgrounds through individual learning and environmental factors.

This study has identified an apparent theoretical gap in the prior research concerning CLLS. By using literary analyses from the past decade, it was believed that CLLS research content could be more detailed, and there is still much room for research on Chinese learning strategies outside of China. Meanwhile, the study samples could be more diversified. Mixed-method research designs and the learning strategy training concept should also be elevated to the theory instruction paradigm to enhance the teaching activities between instructors and students.

In short, this study revealed the overall characteristics of the development of CLLS in the past decade, among which the trend of regionalisation and nationalisation has certain guiding significance for future research. The study has also thoroughly identified the population and theoretical gaps, which are worthy of further investigation. Therefore, it is essential to research Chinese learners in different regions and countries. In addition, the diversified research perspectives of CLLS have far-reaching implications for future research on Chinese as a second language or foreign language teaching. Future researchers are highly recommended to explore the influencing factors and teaching modes of CLLS from different perspectives. Therefore, it deserves ample and substantial research space to explore in studying Chinese language learning strategies in the future.

ACKNOWLEDGEMENT

REFERENCES

Cai, Y. (2014). Investigation and research on Chinese learning strategies of overseas students with Master's degree in Chinese international education. *Research on Teaching Chinese as a Foreign Language, 1*(02), 63-69.

Cao, X. Y. (2010). Investigation and analysis of the foreign students' Chinese language learning strategies. *Chinese Journal of Education, Foreign Language Teaching, 6*, 134-136.

Chen, C. Y., & He, F. T. (2017). A survey of Chinese character learning strategies for students of witchcraft origin in engineering department of University of Malaysia. *Journal of Chinese Studies, 2*(15), 16-30.

Chen, Q. J. (2017). Investigation and analysis of Chinese learning strategies of Asian and African foreign students - A case study of international students from Shanghai Institute of Electrical

Engineering. *Anhui Literature (Second Half), 12*(02), 125-127.

Chen, T. X. (2013). Study on Chinese learning strategies of Thai and American students in non-target language environment. *Overseas Chinese Education, 18*(01), 28-34.

Chen, X. F. (2008). *Chinese language learning strategy research of foreign students* [Master's thesis, Xiamen University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2009&filename=2009081758.nh

Cheng, L (2018). *Study on vocabulary learning strategies for Malaysian Chinese teacher project* [Master's dissertation, Beijing Foreign Studies University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018009850.nh&DbName=CMFD2019

CLEC. (2020). *Center for language education and cooperation*. http://www.chinese.cn/page/#/pcpage/mainpage

Cui, J., & Yan, L. P. (2011). Investigation and analysis of Chinese learning strategies of Central Asian students from Xinjiang. *Language and Translation, 27*(04), 67-707.

Ding, Y. (2018). *A Comparative study of writing learning strategies between Chinese and Non-Chinese students* [Master's dissertation, Nanjing Normal University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018262568.nh&DbName=CMFD2019

Du, L. J. (2018). *The study on spoken Chinese learning strategies of foreign students from Russia* [Master dissertation, Jilin University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018214561.nh&DbName=CMFD2019

Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.

Everson, M. E. (1986). *The effect of word-unit spacing upon the reading strategies of native and non-native readers of Chinese: An eye-tracking study* [Doctoral thesis, Ohio State University]. ProQuest. https://www.proquest.com/openview/79bf7043182eb4dca5a66932026e720d/1?pq-origsite=gscholar&cbl=18750&diss=y

Gan, R. Y. (2004). *Research on the formulation of national vocabulary for teaching Chinese as a foreign language: A case study of South Korea* [Doctoral thesis, Beijing Language and Culture University]. CNKI. http://222.161.207.51:8000/rwt/CNKI/http/NNYHGLUDN3WXTLUPMW4A/kns/detail/detail.aspx?FileName=2005077926.nh&DbName=CDFD2005

Gu, Y. Y. (2014). *Study on Chinese learning strategies for overseas students from Guangxi* [Master's thesis, Guangxi University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201501&filename=1014380438.nh

Guo, S. H. (2019). Influence of international students' personality tendency on Chinese learning strategy selection. *Journal of Higher Education, 30*(04), 17-19.

Hayes, E. B. (1988). Encoding strategies used by native and non-native readers of Chinese Mandarin. *The Modern Language Journal, 72*(2), 188-195. https://doi.org/10.2307/328241

He, H. H. (2011). *A multi-case study of Chinese learning strategies for foreign students* [Master's thesis, Nanjing Normal University]. CNKI. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=1011249561.nh&DbName=CMFD2012

Higgins, J. P. T., & Green, S. (Eds.) (2008). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons Ltd. https://doi.org/10.1002/9780470712184

Hong, C., & Jieyan, L. (2016, February 12). Xi's remarks on the belt and road initiative. *People's Daily Online*. http://world.people.com.cn/n1/2016/0212/c1002-28119992.html

Huang, Q. F. (2018). *A comparative study on The Listening Strategies of Thai College students in two language environments* [Master dissertation, Shandong University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019008876.nh&DbName=CMFD2019

Hurrell, A. (2007). One world? Many worlds? The place of regions in the study of international society. *International Affairs, 83*(1), 127-146. https://doi.org/10.1111/j.1468-2346.2007.00606.x

Jiang, A. (2011). Investigation and study on Chinese learning strategies of Western students. *Modern Chinese (Language Research Edition)*, *7*(03), 121-123.

Jiang, X. (2000). A study on learning strategies of Chinese as a second language. *Language Teaching and Research, 22*(01), 61-68.

Jiang, X., & Zhao, G. (2001). Primary stage foreign students Chinese characters learning strategies research. *Journal of Language Teaching and Research*, *23*(4), 4-6.

Kuo, Y. L. (2015). *Reading strategy use by Chinese as a foreign language learners* [Doctoral dissertation, The Indiana University]. ProQuest. https://search.proquest.com/openview/796b149bb65bc86f940a2afdf58e222d/1?pq-origsite=gscholar&cbl=18750

Li, B. B. (2014). Investigation and analysis of Chinese learning strategies for medical students. *Education and Teaching Forum, 6*(32), 172-173.

Li, B. Y. (2018). *Research on holistic - Analytical cognitive style and Chinese learning strategies of Chinese learners in Tajikistan* [Master's thesis, Xinjiang Normal University]. CNKI. http://y.dllibrary.net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.aspx?dbname=CMFD202002&filename=1020732719.nh

Li, H. (2014)*. Research on the correlation between individual factors and Chinese learning strategies of overseas students in five Southeast Asian countries* [Master's thesis, Yunnan Normal University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201501&filename=1015004122.nh

Li, J. (2014). *Investigation on Chinese learning strategies of Thai students and improvement of teachers' teaching method* [Master's thesis, Xi 'an International Studies University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201501&filename=1014073599.nh

Li, J. (2016)*. Investigation and research on Chinese Learning strategies of middle and senior Indo-Filipino students* [Master's thesis, Fujian Normal University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201701&filename=1017029589.nh

Li, J. Q. (2017)*. Study on Chinese learning strategies for Kazakh students studying in China* [Master's thesis, Shandong Normal University]. CNKI. http://y.dl-library.net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017240199.nh

Li, Y. (2017). *Exploring self-regulated learning (SRL) and listening strategy instruction in a Chinese L2 classroom*. [Doctoral dissertation, The University of San Francisco]. ProQuest. http://www.proquest.com/en-US/products/dissertations/individuals.shtml

Li, Y. L. (2019). *Study on Chinese learning strategies of German middle school students in non-target language environment* [Master's thesis, Qingdao University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019624251.nh&DbName=CMFD2020

Li, Y. M., & Shi, C. H. (2017). Thoughts on localization of international Chinese education. *Chinese Language, 66*(02), 245-251.

Lin, K., & Lv, X. (2005). Analysis of Chinese Learning Strategies for Vietnamese Students. *Journal of Chinese Language Institute, Jinan University, 4*, 19-24.

Lin, Y. (2016). *Case study on Chinese learning strategies of primary and middle school students in Singapore International School in Indonesia* [Master's thesis, Guangzhou University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201701&filename=1016192641.nh

Liu, F. Q. (2012). Investigation and analysis of Chinese learning strategies of Japanese college students. *Chinese Learning*, *33*(04), 97-103.

Liu, J. Y. (2019) *Investigation on Chinese learning strategies of Arab overseas students* [Master's thesis, Central China Normal University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019253049.nh&DbName=CMFD2020

Liu, L. (2018). *A survey of Chinese character learning strategies of students with and without Chinese character circle* [Master dissertation, Nanjing Normal University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018262563.nh&DbName=CMFD2019

Liu, R. N., & Yuan, Y. C. (2017). Research on Chinese learning strategies of international school students. *Overseas Chinese Education, 22*(01), 48-54.

Liu, Y. J. (2013). The influence of self-esteem on spoken Chinese learning strategies of central Asian international students. *The Technology Entrepreneur, 07*, Article 182.

Lu, Y. N. (2012). *Research on the oral Chinese learning strategies of Thai students* [Master dissertation, Shandong University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201301&filename=1012465769.nh

Luo, X. (2019) *Investigation and research on Chinese learning strategies of African foreign students in China* [Master's thesis, Nanjing Normal University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019256538.nh&DbName=CMFD2020

Lv, Y. X. (2013). *Investigation and analysis of Chinese learning strategies in second language acquisition theory* [Master dissertation, Hebei University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201401&filename=1013032540.nh

Ma, Y. (2019). *Research on Chinese learning strategies of foreign students in the Middle East* [Master's thesis, Xi 'an International Studies University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019909888.nh&DbName=CMFD2019

Ning, J. (2019). *Self-regulation and Chinese character learning strategies in Irish Higher Education* [Doctoral dissertation, The University of Dublin]. TARA. http://hdl.handle.net/2262/89567

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524490

Ou, F. Y. (2018). *Investigation on the use of Chinese learning strategies among Bangladeshi students in Yunnan University* [Master's thesis, Yunnan University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018251391.nh&DbName=CMFD2020

Oxford, R. (1990). *Language learning strategies: What every teacher should know.* Newbury House Publishers. https://doi.org/10.1111/j.1540-4781.1989.tb06367.x

Oxford, R. L., & Nyikos, M. (1989). Variables affecting choice of language learning strategies by university students. *Modern Language Journal*, *73*(3), 291-300. https://doi.org/10.1111/j.1540-4781.1989.tb06367.x

Peng, X. J. (2017). *A case study on Chinese learning strategies of excellent overseas students in Xinjiang University* [Master's thesis, Xinjiang University]. CNKI. http://y.dl-library.net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017712448.nh

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing. https://doi.org/10.1002/9780470754887

Qian, Y. L. (2006). A survey on Chinese reading learning strategies of Korean Students. *Teaching Chinese in the World, 04*, 80-88.

Rubin, J. (1987). *Learner strategies: Theoretical assumptions, research history and typology.* In A. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 15-31). Prentice/Hall International.

Sheng, D. D. (2019). *Study on the correlation between tolerance of ambiguity and learning strategies of intermediate Chinese learners in Myanmar* [Master dissertation, Yunnan Normal University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019922917.nh&DbName=CMFD2020

Shu, C. H. (2013). *The adoption of learning strategies of multilingual adult learners of Chinese in South Texas* [Doctoral dissertation, Texas A&M University Kingsville]. ProQuest. https://www.proquest.com/docview/1440470588

Stern, H. H. (1983). *Fundamental concepts of language teaching*. Oxford University Press.

Tan, R. C. (2018). *Research on Chinese learning strategies for students of Laos* [Master's thesis, Guangxi University for Nationalities]. CNKI. http://y.dl-library. net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.aspx?dbname=CMFD201901&filename=1018129140.nh

Wang, B. J. (2015). *Study on Chinese learning strategies for foreign students in China* [Master's thesis, Yunnan University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201502&filename=1015610017.nh

Wang, J. (2011). Investigation and analysis on the use of Chinese learning strategies in short-term training courses for American college students in China. *Chinese Journal of External Education, 5*(04), 20-21.

Wang, J. Q. (2011). *Research on the correlation between Chinese learning strategies and other individual factors for middle and senior international students* [Master's thesis, Shandong University]. CNKI. http://y.dl-library.net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017240199.nh

Wang, M. (2018). *A comparative study of reading learning strategies between Chinese and non-Chinese students* [Master's dissertation, Nanjing Normal University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018262610.nh&DbName=CMFD2019

Wang, P. (2016). *Investigation on Chinese learning strategies of Brazilian students and its implications for Chinese teaching in Brazil* [Master's thesis, Hubei University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201901&filename=1016143523.nh

Wang, X. Q., & Li, M. Y. (2013). Learning strategies for improving spoken Chinese skills of central Asian students. *Chinese Journal of Language and Translation, 000*(003), 84-88.

Wang, Y. (2017). *Investigation on Chinese learning strategies of overseas students* [Master's thesis, Hebei University]. CNKI. http://y.dl-library.net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017971779.nh

Wang, Y. Y. (2013). *Investigation and research on Chinese learning strategies of Mongolian junior middle school students* [Master's thesis, Minzu University of China]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201401&filename=1014125630.n

Wei, H. (2012). Regional Chinese education and research of language types. *Journal of Yunnan Nationalities University (Social Sciences), 30*(03), 151-156.

Wu, Q. (2018). Research on oral teaching based on Chinese learning strategies of African foreign students in China. *Chinese Language Teaching Newsletter, Journal D (Academic Journal), 41*(06), 65-67.

Wu, Y. H. (2013). A theoretical discussion on "localization" and "universality" of textbooks for international Chinese teaching. *Application of Language and Character, 21*(3), 117-125.

Wu, Y. Y. (2007). *A study of Chinese learning strategies for foreigners in different environments* [Doctoral thesis, Shanghai Normal University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CDFD9908&filename=2007199849.nh

Xie, T. (2018). *Study on Chinese learning strategies for primary and secondary school students in New Zealand* [Master's thesis, Central University for Nationalities]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018321141.nh&DbName=CMFD2019

Xie, T. T. (2019). *An empirical study on the relationship between Chinese learning beliefs, learning strategies and academic achievements of international students studying in China* [Master dissertation, Shandong University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019081729.nh&DbName=CMFD2019

Xu, T. T. (2018). *Study on learning strategies of intercultural communication competence of international students in Jiangsu Province* [Master dissertation, Jiangnan University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018265654.nh&DbName=CMFD2019

Xu, Y. L. (2018). *Research on the relationship between Chinese writing strategy and writing scores of Malaysian international students* [Master dissertation, Beijing Foreign Studies University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1018009689.nh&DbName=CMFD2019

Yang, J. (2015). *A study on Chinese learning strategies for college students in Northern Thailand* [Master's thesis, Yunnan Normal University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201601&filename=1015904103.nh

Yang, L. (2012). Foreign students Chinese language learning strategy research. *Journal of New Western (Theory), 22*(12), 155-156.

Yang, Y. (2019). *The study on the learning strategies about classroom teaching for beginners based on Confucius Institute as University Hassan* [Master dissertation, Shanghai International Studies University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019114674.nh&DbName=CMFD2019

Yu, X. L., & Huang, C. (2016). Analysis of the current situation of Chinese learning strategies for African foreign students - Based on a comparative study with other second language learners. *Journal of Changchun University, 26*(02), 81-85.

Yuan, Y. C., Rao, Y. P., & Yuan, Y. (2011). Empirical research on Chinese learning strategies, attitudes and motivations of Southeast Asian students. *Academy, 4*(5), 64-69.

Zhai, J. Y. (2019). *Research on spoken Chinese learning strategies of middle and advanced level learners - A case study of Southeast Asian students in Guangxi University* [Master dissertation,

Guangxi University]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019215535.nh&DbName=CMFD2020

Zhang, B. H. (2017). *Research on Chinese learning strategies of Chinese and Non-Chinese high school students in North Thailand* [Master's thesis, Yunnan Normal University]. CNKI. http://y.dl-library.net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017730055.nh

Zhang, H. (2014). *Investigation and research on Chinese learning strategies of senior high school students in Thailand* [Master's thesis, Shaanxi Normal University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201501&filename=1014399911.nh

Zhang, J. (2011). Chinese learning strategies for American Chinese learners. *Journal of Kaifeng University, 25*(03), 64-68.

Zhang, L. L. (2013). *Research on Chinese learning strategies of Central Asian students studying in China* [Master's thesis, Chongqing University]. CNKI. https://kns.cn ki.net/KCMS/detail/detail.aspx?dbname=CMFD201401&filename=101304342 0.nh

Zhang, L. R., & Wan, Y. (2019). Investigation and analysis of Chinese learning strategies for international students from five Central Asian countries. *Chinese Learning, 230*(02), 94-104.

Zhang, M. M. (2012). *Investigation on the influence of language learning environment changes on Chinese learning strategies of Korean students in China* [Master's thesis Yunnan University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2012&filename=1012419163.nh

Zhang, R. F., & Ge, S. R. L. (2015). Study on characteristics and differences of Chinese learning strategies of Mongolian foreign

students. *Journal of Inner Mongolia Normal University, 44*(01), 122-128.

Zhang, X. G. (2015). Methodological issues in the study of Chinese learning strategies for foreign students. *Journal of Jilin Institute of Education (Late Edition), 31*(06), 68-69.

Zhao, J. (2012). *Research on Chinese learning strategies for African foreign students* [Master's thesis, Northeast Normal University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201301&filename=1013142837.nh

Zheng, Y. P. (2014). *Research on Chinese learning strategies of Thai high school students in non-target language environment* [Master dissertation, Jinan University]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201501&filename=1015002860.nh

Zhou, J. (2019). *Research on the correlation between Chinese learning strategies and HSK scores of Thai students* [Master's thesis, Guangxi University for Nationalities]. CNKI. http://kns.cnki.net/kns/detail/detail.aspx?FileName=1019112169.nh&DbName=CMFD2020

Zhou, L (2004). *The relationship between language learning strategies and Chinese listening comprehension for intermediate Korean students* [Master dissertation, Beijing Language and Culture University]. CNKI. http://222.161.207.51:8000/rwt/CNKI/http/NNYHGLUDN3WXTLUPMW4A/kns/detail/detail.aspx?FileName=2004102281.nh&DbName=CMFD2004

Zhou, L. (2013). *A comparative study of Chinese learning strategies for elementary level Thai students in target-language and non-target-language environments* [Master's thesis, Minzu University of China]. CNKI. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201401&filename=1013317105.nh

Zhou, X. T., Li, M. Q., & Sun, F. Y. (2014). A study on the relationship between Chinese learning motivation and learning strategies of international students - A case study of two universities in Hunan Province. *Journal of Guilin University of Aerospace Technology, 19*(02), 204-206.

Zhu, X. (2018). *Study on Chinese learning strategies for Japanese students of Confucius Institute of Sapporo University* [Master's thesis, Shenyang University]. CNKI. http://y.dl-library.net.cn:8000/rwt/CNKI/https/NNYHGLUDN3WXTLUPMW4A/KCMS/detail/detail.
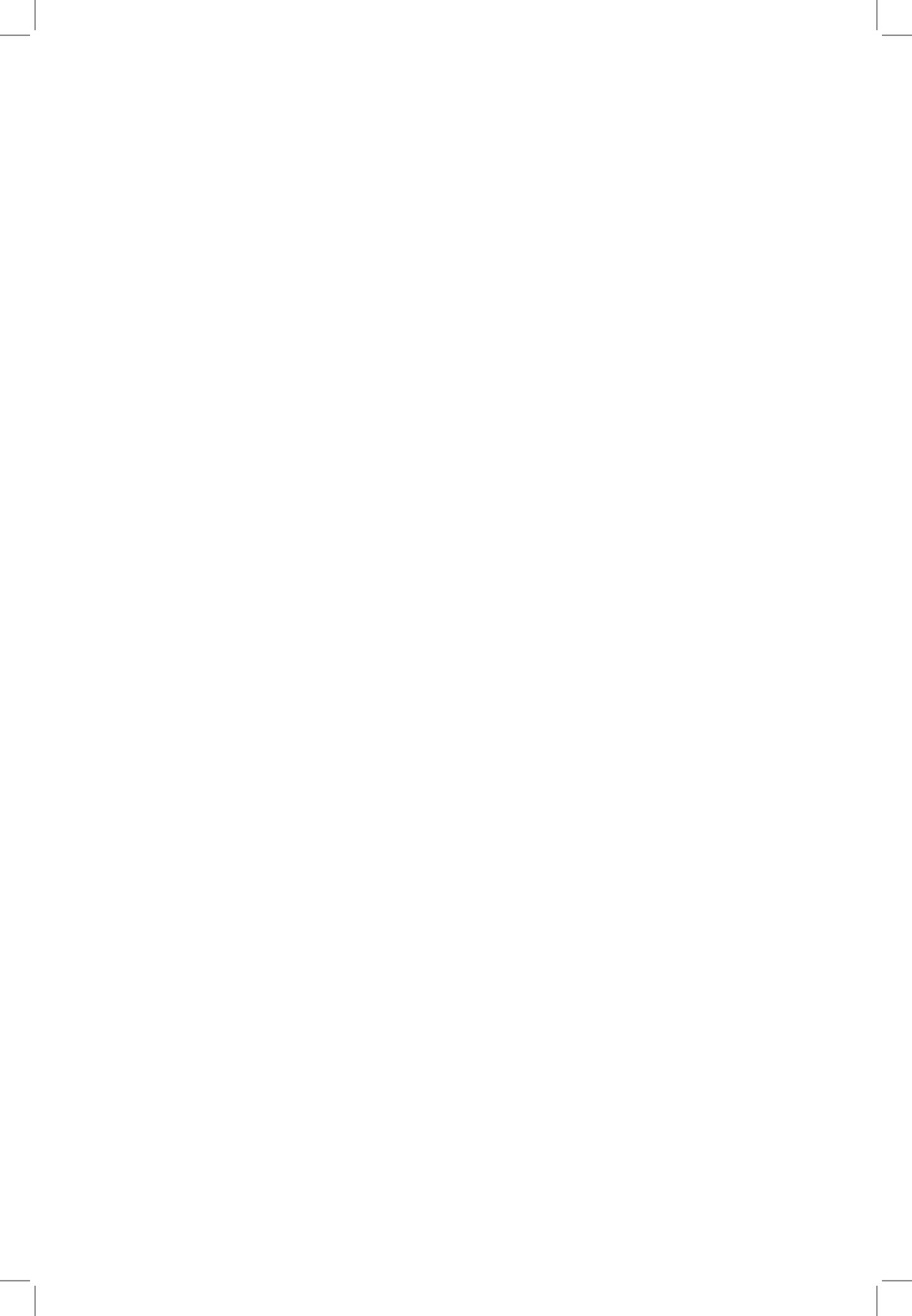
English Language Speaking Anxiety, Self-Confidence and Perceived Ability among Science and Technology Undergraduate Students: A Rasch Analysis

Kamal J I Badrasawi[1*], Noor Lide Abu Kassim[1], Ainol Madziah Zubairi[1], Elia Md Johar[2] and Siti Sakinah Sidik[1]

[1]*Kulliyyah of Education, International Islamic University Malaysia (IIUM), 50728 Kuala Lumpur, Malaysia*
[2]*Akademi Pengajian Bahasa, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia*

ABSTRACT

The purpose of this paper is to analyse English language speaking anxiety, self-confidence, and perceived ability in English oral communication among Science and Technology undergraduate students. It also aims to identify any significant differences in these constructs based on selected students' demographic variables. The study employed the survey method with a 41-item questionnaire administered to a voluntary response sample of three hundred 3rd and 4th-year science and technology undergraduates from three Malaysian public universities. The Polytomous Rasch model was used to analyse the data. The analysis showed that the participants experienced English speaking anxiety, low confidence, and high perceived ability in English oral communication. There were significant mean differences in English speaking anxiety across the type of university as well as in confidence and perceived ability based on academic program. The participants were more confident and could perform better in familiar situations and communicate on familiar topics to familiar audiences. The findings suggest that the participants need more training on English oral communication. More authentic situations are also needed for them to practise and improve their proficiency levels. Other suggestions include providing lecturers with training modules, re-assessing the current language policies, and implementing certain programmes at the tertiary education level. Language programmes could be directed towards more social situations to enable

undergraduates to make English a social practice, lower English speaking anxiety, and boost confidence.

Keywords: Confidence, English oral communication skill, perceived ability, Rasch Model Analysis, science and technology undergraduates, speaking anxiety

INTRODUCTION

The English language is considered a lingua franca due to its dominance in many social, economic, scientific and political activities (Crystal, 2004; Nishanti, 2018; Pandarangga, 2015; Pennycook, 2014; Reddy, 2016). Reports have shown that about 750 million people use English as a second language (ESL), and it is prioritised in around 70 countries in the world (Reddy, 2016). Furthermore, English language learners have reached 1 billion, while roughly 2 billion people have mastered the language, with an estimation that half of the world population might be English proficient in the next few decades (English Cultural Council as in Xue & Zuo, 2013). The English language has also become the medium of instruction in many higher education institutions, and it has also been used as a criterion with which students secure admissions into tertiary education programs (Pandarangga, 2015). As such, English communication competency is considered among the highly significant requirements of university graduates' quality in Asia (UNESCO, 2012).

Globally, many countries are constantly working to improve English language proficiency among their people and learners, as English has become the most important language in the world (Hudson & Hudson, 2003). Companies and institutions hire employees who can communicate in the English language efficiently within a wide range of workplace communicative events due to the substantial roles the language plays in the current worldwide transactions (Pandarangga, 2015; Sheth, 2016). The same idea is mooted in Roshid and Chowdhury's (2013) notion that employers are looking for graduates with high English communication skills, mainly those who can explain ideas, identify issues, and solve problems related to their work constructively. For instance, English communication skill is an imperative employability requirement in India to get a better job (Clement & Murugavel, 2015). Clement and Murugavel further emphasize that engineering graduates can only internationally communicate if they are proficient in English communication skills, mainly those related to their profession. Wijewardene et al. (2014) assert that competency in English—especially the spoken and written—is among the crucial factors determining graduates' employment in the private and public sectors in Sri Lanka. The same trend is seen almost worldwide.

Of the four language skills, speaking skill is considered the most important. Ur (1996) argues that those who know the language speak it, implying that it is important to use it effectively rather than knowing it (Scrivener, 2005). We live in a time where the need to speak English fluently is dire, especially for those who want to advance in certain fields of human endeavours (Al-Sibai, 2004). The literature

on English oral proficiency has shown several factors affecting its improvement. These factors include speaking anxiety (Ahmed et al., 2017; Bux et al., 2015; Dordinejad & Ahmadabad, 2014; Kumar, 2018; Ramamuruthy, 2019; Salem & Al Dyiar, 2014; Zhang & Zhong, 2012); low self-confidence (Gürler, 2015; Kalanzadeh et al., 2013; Mandokhail et al., 2018; Tridinanti, 2018) and perceived ability in English oral communication (Alawiyah, 2018; Linnenbrink & Pintrich, 2003; Pajares, 1996; Sunyi, 2017; Zahiri et al., 2017). This study primarily aims to analyze English language speaking anxiety, self-confidence, and perceived ability in English oral communication among science and technology undergraduate students using the Rasch Model in Malaysia.

English Language Anxiety and its Effects on Spoken/Oral Interaction

Horwitz et al. (1986) describe language anxiety as a multiplex phenomenon of "self-perceptions, beliefs, feelings, and behaviours related to classroom language learning arising from the uniqueness of the language learning process" (p.128). Spielberger (1983) defines language anxiety as "a subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the autonomic nervous system" (p.1). It could be classified into two categories, namely 'trait anxiety' and 'situational anxiety' (i.e., state anxiety) (Akkakoson, 2016; Spielberger, 1983). Trait anxiety is the anxiety that language learners experience in every situation (Pappamihiel,

2002). If individuals fail to minimise this negative feeling, it becomes permanent in their temperament (Riasati, 2011). On the other hand, situational anxiety is anxiety expressed by a second language learner in specific situations (MacIntyre, 1999). This kind of anxiety is usually felt because of the lack of familiarity of language learners with the particular situation in which they are to use the language, and once they become familiar with it, their anxiety diminishes or even vanishes (Riasati, 2011).

Interestingly, speaking skills in a second or foreign language is the most anxiety-inducing skill among individuals (Zhang & Zhong, 2012). Furthermore, research has revealed a connection between English speaking anxiety and students' speaking fluency; meaning that speaking anxiety has an unfavourable effect on learners' English-speaking fluency (Salem & Al Dyiar, 2014) and, in particular, students' oral presentation as well as their conversation with English native speakers (Bux et al., 2015). Therefore, a number of research (quantitative and qualitative) has been conducted to identify the causes of speaking anxiety, its destructive effects on speaking as well as how the problem can be addressed (Ahmad et al., 2017; Bux et al., 2015; Kumar, 2018; Ramamuruthy, 2019; Salem & Al Dyiar, 2014). For instance, Ahmad et al. (2017) found the inter-language meaning system as the cause of postgraduate students' English language speaking anxiety, while Kamaruddin et al. (2019) found low self-esteem and social anxiety as the factors contributing to the listening and speaking

anxieties among Malaysian university students majoring in non-English programs. Moreover, Mulyono et al. (2019) identified negative attitudes, language barriers and intercultural communication apprehension as the main factors provoking English-speaking anxiety among non-English native speakers studying in the Indonesian universities.

Moreover, Ramamuruthy (2019) found fear of being negatively evaluated as the main factor for English speaking anxiety among diploma students studying at an international college in Malaysia. Amiri and Puteh (2018) found that several factors such as insufficient linguistic competency, inadequate knowledge of the presentation's content, students' negative perception towards the examiners, and examiners' linguistic deficiency in understanding presentations lead to speaking anxiety among international students studying doctoral programs in different Malaysian universities. Sadighi and Dastpak (2017) found that fear of making mistakes, being negatively evaluated, and inadequate vocabulary knowledge were the major causes of speaking anxiety among ESL Iranian students. It is important to maintain that speaking anxiety level might differ due to demographic variables (Badrasawi et al., 2020). In conclusion, it could be inferred that language anxiety affects the speaking performance of ESL/EFL learners, and the higher the level of language anxiety, the worse their performance in speaking will be, and vice versa.

Perceived Ability and its Effects on Spoken/Oral Interaction

Perceived ability refers to individuals' self-perception regarding their ability to perform effectively in a specific situation based on their skills and capabilities. A good example is the Common European Framework of References (CEFR), where perceived ability is directly linked to learners' ability to achieve the stipulated 'Can do' statements (Alderson, 2017). The CEFR describes what learners can do across five language skills: spoken production, spoken interaction, reading, listening and writing (Alderson, 2017). For all five skills at each level, there are sets of detailed 'Can Do' statements. For example, the spoken interaction focuses on the learners' production and participation in conversations and discussions. Perceived ability is important for students to participate in maintaining, starting, taking turns and ending conversations. Without adequate perceived ability, learners will stumble thus fail to achieve the target performance stipulated in the 'can do' statements.

Literature has shown that perceived ability is among the factors that affect students' English oral communication skills. This is not because of their deficiency in mastering the linguistic aspect of communication, but it is as a result of the negative feeling they usually have about themselves and the audience such as communication apprehension, fear of making mistakes and fear of negative evaluation as affirmed by Horwitz et al. (1986). Speaking anxiety tends to be higher when the ESL/EFL speakers perceive their

speaking ability to be low, and vice versa. Cognitive component, as emphasised by Shrauger and Schohn (1995), is an integral part of individuals' perceived ability. It refers to the self-evaluation of performance, meeting of own expectations and continuous excellence compared to others. Several studies have contended that English language learners who perceive their ability as high are confident to successfully engage themselves in English conversations with others expressing their ideas overtly, and the reverse is also true (Alawiyah, 2018).

When people are optimistic in performing excellently in a task, they become strongly motivated and driven by their enthusiasm and interest to achieve the desired goals (Linnenbrink & Pintrich, 2003; Pajares, 1996; Sunyi, 2017). Zahiri et al. (2017) studied the effects of speaking anxiety and perceived ability on monologue speaking skills of students at a state senior Islamic high school in Medan, Indonesia. They found that both speaking anxiety and perceived ability affected students' monologue speaking skills. While the former adversely affected students' speaking skills, the latter had a positive effect on it. The more anxious students are in speaking, the worse the speaking skills will be. Similarly, the higher their perception of their ability, the better they will be in speaking skills. Desmaliza and Septiani (2017) showed a significant relationship between students' perceived ability and their speaking skills, positively influencing the performance of students in oral communication activities.

Confidence and its Effects on Spoken/ Oral Communication

Koriat et al. (1980) describe confidence as the belief in oneself to perform tasks successfully. It could also refer to one's realistic sense of capacity and possessing sufficient knowledge. Brown (2004) stresses that successful activities require a high level of self-confidence, including second language acquisition (Kalanzadeh et al., 2013). Hart (1989) purports that confident learners would most likely get the task done. Confidence in spoken interaction is usually associated with the speakers' certainty about using the language. Thus, confidence plays a crucial role in motivating learners to communicate (Tanveer, 2007). The higher the confidence, the most likely the learners will be involved in communication activities. Tsou (2005) reported that high self-confidence was positively correlated with oral performance and concluded that self-confidence is crucial in learners' inclination to communicate. It is supported by Stenstrom (2014), who asserts that confidence is important in spoken interaction as it is a two-way process. Other researchers found a positive, strong relationship between self-confidence or self–esteem of FL or SL learners and their oral proficiency of speaking skills (Gürler, 2015; Mandokhail et al., 2018; Tridinanti, 2018).

Study Setting

In Malaysia, English has its status as an important second language, and it is used as the medium of instruction for science and technology in higher institutions. In

addition, it is extensively used in various settings, for instance, social, commercial and national and international transactions. The Malaysian educational system values English language acquisition among students at all school levels, and the Malaysia Education Blueprint (2013-2025) has stressed improving the students' English language proficiency at all stages. Considerable efforts have been devoted to improving graduates' English language proficiency to work in a globalised economy whereby the English language is the international language of communication, as clearly mooted in the Blueprint 2013-2025 (Ministry of Education, 2012). More importantly, this focus encourages the graduates to participate in the workforce and contribute to the country's development in the future. Graduates proficient in English and who have leadership and technical skills get more opportunities to find a job in their respective fields (Ismail, 2011).

In order to improve the level of English language proficiency from preschool to tertiary education, Malaysians have adopted CEFR with the establishment of the English Language Standards and Quality Council (ELSQC) in 2013. Malaysian Science and Technology undergraduates went through lessons for ESL and CEFR-aligned tests such as Malaysian University English Test (MUET), Cambridge Placement Test (CPT) and an English assessment test administered by the British Council (Aptis) before graduating. The purpose of taking at least one of these tests is primarily to measure their proficiency in the language. Since CEFR provides a globally accepted

measurement framework, educational institutions and employers can easily compare qualifications to other exams in their countries. The minimum expectation for university graduates is B2 which ensures they can understand more complex texts, handle more abstract topics and technical discussions, and communicate and interact comfortably with native speakers. C2 is the highest level CEFR scale of achievement, required only for those entering certain professions, such as English language teachers.

However, undergraduate Science and Technology students must be concerned because a recent report has indicated an urgent need for more qualified and skilled graduates in science, technology, engineering and mathematics (Chin, 2016). Overall, extensive research has shown that undergraduate students have low proficiency (i.e. not up to the desired levels) in the English language (Ismail, 2011; Musa et al., 2012; Nair et al., 2012, Rusli et al., 2018). The language competence of Malaysian undergraduates is still a long way from satisfactory level though they have learned English for 11 to 13 years in schools (David et al., 2015). As a result, communication skills amongst Malaysian graduates have deteriorated (Shakir, 2009). The Malaysia Education Blueprint states that "poor English proficiency among graduates has been consistently ranked as one of the top five issues facing Malaysian employers since 2006" (Ministry of Education, 2012, p.12). In addition, recent reports have shown that the number of unemployed Malaysian graduates are increasing due to their lack

of the required levels of English speaking skill to get or secure a job (Free Malaysia Today, 2017; The Sun Daily, 2018). Besides, the Salary Surveys 2016 by the Malaysian Employers Federation (MEF) found that over 90% of respondents were required to improve their English capability to get a job (Malaysian Employers Federation, 2016).

Evidently, in such pressing situations where further investigation is needed, this study aims to analyse English language speaking anxiety, self-confidence, and perceived ability in English oral communication among undergraduate science and technology students using the Rasch Model. Also, it aims to find the significant differences in mean scores of the factors about selected demographic variables (i.e. gender, academic year, university type/category and faculty).

METHODOLOGY

This study used the survey method to determine English language speaking anxiety, self-confidence and perceived ability in English oral communication among science and technology undergraduate students. Based on the related literature, items measuring *Foreign Language Anxiety Scale* (FLAS) were pooled from previous studies (Ali, 2017, McCroskey, 1970; Pappamihiel, 2002; Yim & Yu, 2011) while items measuring *Confidence* and *Task Difficulty* were based on a questionnaire that was developed to measure confidence and task difficulty in oral proficiency testing, a study by Kassim and Zubairi (2003). The Common European Framework of Reference (CEFR) Can-Do Statement rubrics was applied to measure the perceived ability. Altogether, there were 41 items on a five-point Likert scale, categorised into three sections: *English speaking anxiety* (ANX) (12 items); *Confidence in oral communication/interaction (CON)* (19 items); *Can-Do statements (perceived ability in oral communication) (CAN)* (10 items). These items were piloted, and the Cronbach's Alpha values for the three constructs were 0.87, 0.74 and 0.87, respectively. Two Malaysian research universities and a comprehensive university offering Science, ICT and Engineering programmes were identified as the study population. Third- and fourth-year undergraduate students from those academic programmes were invited to participate in the study, from which three hundred students volunteered. The questionnaires were administered in person to those who agreed to participate— However, some who were unable to join answered the questionnaire via Google Forms. The breakdown of the sample in terms of the programme of study and institution is presented in Table 1.

The collected data were analysed based on the Polytomous Rasch model using *Winsteps* version 4.1.0 (Linacre, 2018). Unlike other kinds of analysis, interval data are always used in Rasch analysis. It uses logit units; therefore, it is possible to get the difficulty to measure for each item and for category. All persons and items are placed on the same interval scale to see their distributions. The most difficult items to endorse are positioned toward the upper part of the scale and

Table 1
Study sample's demographic characteristics

Variable	Level	Frequency (n)	Percentage %
University	Comprehensive University	110	36.7
	Research University A	90	30.0
	Research University B	100	33.3
Faculty	Science	100	33.3
	ICT	100	33.3
	Engineering	100	33.3
Year	Third 3rd	140	46.7
	Fourth 4th	160	53.3
Gender	Male	116	38.7
	Female	184	61.3

vice versa. Another important point is that Rasch analysis ensures if the items contribute meaningfully to the construct by investigating the item Fit statistics. The inferential analyses (One-way analysis of variance (ANOVA) and Independent samples t-test) were also conducted to compare mean scores of English-speaking anxiety, Confidence and Perceived ability in English oral communication across selected demographic variables (i.e. university type/category, faculty, academic year and gender). The results are displayed in Tables and Figures.

RESULTS

Psychometric Properties of the English Speaking Anxiety Scale

A 12-item scale was used to measure English language speaking anxiety of the 300 Science and Technology undergraduate students from Malaysia's three selected public universities. Table 2 shows the psychometric properties for all items using the Rasch Model. It is important to note

that the two misfit items (i.e. items 6 and 10) with infit value > 1.5 were not included in the final analyses as recommended in the literature (Bond & Fox, 2015). Though the two items were recoded, they still showed misfit values. Both items were misfits as they shared the same characteristics in that they were worded positively. The item reliability was high (0.95), with separation index (3.04) > 2; and the reliability of a person's ability is also high (0.90) with person separation index (4.29) > 2. For the point-measure correlation coefficients, ten items had positive values, ranging from 0.74 to 0.83. It means that all items were working in the same direction to define the English language speaking anxiety construct. The ten items' infit and outfit Mean-square statistics were within the recommended range (0.5–1.5), indicating that they contributed meaningfully to the measured construct (i.e., English speaking anxiety). Thus, the scale's unidimensionality was met, with the variance explained by the measures being 64%. The largest factor

Table 2
Reliability, separation, item fit statistics and point-Measure correlation coefficients

No	Item	Infit		Outfit		PT-Measure CORR
		MNSQ	ZSTD	MNSQ	ZSTD	
1	I feel very self-conscious about speaking English in front of others.	1.07	0.8	1.08	0.9	.77
2	I am afraid that native speakers will laugh at me when I speak English.	.96	-0.5	.94	-0.7	.80
3	I lack self-confidence when I speak in English to others.	.80	-2.6	.80	-2.6	.83
4	I always feel that others speak English better than I do.	1.14	1.6	1.09	1	.78
5	I doubt my ability to speak English properly.	.83	-2.2	.84	-2	.82
6	I enjoy speaking English with native speakers.	**Misfit item (infit > 1.5) DELETED**				
7	My hands tremble when I am giving a speech.	1.33	3.7	1.38	4.1	.74
8	My heart beats very fast while waiting for my turn to start a speech.	1.09	1	1.11	1.2	.78
9	While preparing for giving a speech, I forget facts I really know due to tension and nervousness.	.89	-1.4	.86	-1.7	.81
10	I feel relaxed while giving a speech.	**Misfit item (infit > 1.5) DELETED**				
11	When I make a mistake while giving a speech, I find it hard to concentrate on the parts that follow.	.92	-1	.94	-0.6	.79
12	While giving a speech, I experience a feeling of helplessness building up inside me.	.87	-1.6	1.03	0.4	.77
Means		.99	-0.2	1.01	0.0	
P.SD		.16	1.9	.16	1.9	
Reliability of item difficulty measures		0.95				
Item separation		3.03				
Reliability of person ability		.90				
Person separation		4.26				
Variance explained by measures		60.4%				
Unexplained variance in 1st contrast		2.3 (< 2 items) Linacre 2019.				

extracted from the residuals was equivalent to 2.3 units, which had a strength of about two items (Linacre, 2019).

English Speaking Anxiety and its Factors

Table 3 shows the order of the item difficulty measures arranged from the highest to the lowest. Figure 1 displays the hierarchy and distribution of items and persons on the same interval scale. Overall, students easily endorsed the scale items as the person ability mean (0.27) was higher than the item difficulty mean (0.00), indicating that the participants had experienced English language speaking anxiety. The least endorsed items were placed towards the upper part of the scale, and highly

Table 3

Item difficulty measures (English speaking anxiety and factors)

No	Item	Difficulty measures	S.E.
12	While giving a speech, I experience a feeling of helplessness building up inside me.	0.57	0.08
2	I am afraid that native speakers will laugh at me when I speak English.	0.29	0.08
7	My hands tremble when I am giving a speech.	0.19	0.08
3	I lack self-confidence when I speak in English to others.	0.17	0.08
11	When I make a mistake while giving a speech, I find it hard to concentrate on the parts that follow.	0.17	0.08
5	I doubt my ability to speak English properly.	0.14	0.08
9	While preparing for giving a speech, I forget facts I really know due to tension and nervousness.	-0.09	0.08
1	I feel very self-conscious about speaking English in front of others.	-0.28	0.08
4	I always feel that others speak English better than I do.	-0.56	0.08
8	My heart beats very fast while waiting for my turn to start a speech.	-0.60	0.08
Means		0.00	0.08
P.SD		0.36	0.00
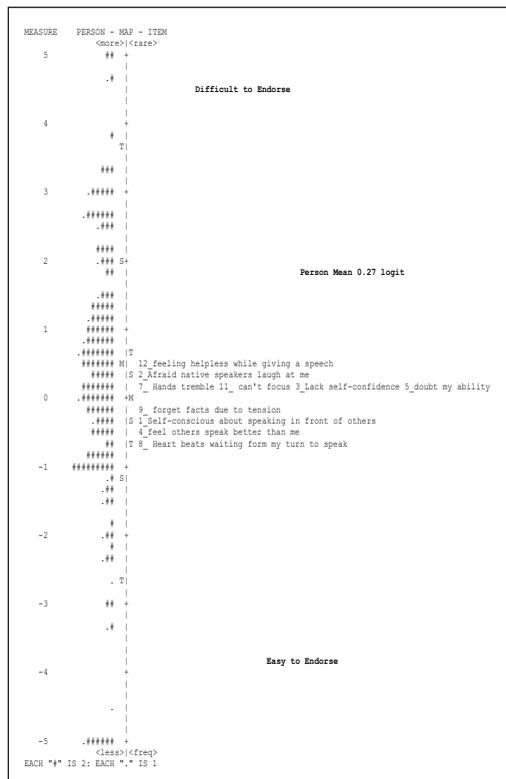
Figure 1. Person-Item map (English Speaking Anxiety)

endorsed items were placed towards the lower part. It indicated that the participants felt worried and most anxious before they spoke. They were thinking about others whom the participants believed were better than themselves.

Psychometric Properties of Self-Confidence in Oral English Communication Scale

A 19-item scale was used to measure the confidence of 300 Science and Technology undergraduates from three Malaysian public universities when they communicated in an Individual Interview (IV1-IV7), in Paired Discussion (PD8-PD12) and Group Discussion (GD13-GD19). The results in Table 4 showed that the item reliability was very high (0.98), with separation index (6.28) > 2; and the person ability reliability was also high (0.92) with person

Table 4

*Item fit statistics, point-measure correlation coefficients, reliability and separation (individual interview **IV**, paired discussion **PD** and group discussion **GD**)*

No	Item	Infit		Outfit		PT-Measure CORR
		MNSQ	ZSTD	MNSQ	ZSTD	
	Individual Interview					
IV1	Tasks that require me to respond (answer) immediately make me nervous.	1.12	1.5	1.12	1.4	0.69
IV22	I find tasks that require my response (answer) without preparation frightening.	1.19	2.2	1.26	2.8	0.68
IV3	I worry when I do not know what the interviewer is going to ask me.	0.85	-1.8	0.84	-1.9	0.75
IV4	I perform well on tasks that give me time to prepare.	1.04	0.5	1.13	1.5	0.71
IV5	Talking about familiar topics make me feel more confident.	1.05	0.6	1.06	0.7	0.74
IV6	Interacting with an interviewer I am familiar with makes me feel comfortable.	0.99	-0.1	1.03	0.3	0.73
IV7	My performance on a task depends on the interviewer that I get.	0.99	-0.1	1.03	0.4	0.70
	Paired Discussion (in Peers)					
PD8	I find tasks that require my immediate response (answer) frightening.	1.10	1.3	1.12	1.3	0.68
PD9	I worry when I do not know what my peer is going to ask me.	1.27	3.2	1.25	2.6	0.64
PD10	I perform well on tasks that give me time to prepare.	0.88	-1.5	0.92	-0.9	0.74
PD11	I perform well on tasks that require me to interact with a peer.	0.88	-1.5	0.91	-1	0.73
PD12	My performance on a task depends on the peer that I get.	1.07	0.9	1.09	1	0.72
	Group Discussion					
GD13	Tasks that give me time to prepare my response (answer) make me feel comfortable and relaxed.	0.72	-3.7	0.77	-2.9	0.78
GD14	I like tasks where I am given time to prepare my response (answer) while the other group members take turns speaking.	0.89	-1.4	0.9	-1.1	0.75
GD15	I worry when I do not know what my group members are going to ask me.	1.09	1.1	1.25	2.6	0.66
GD16	It does not matter to me whether I am given time to prepare my response (answer).	**Misfit item (infit > 1.5) DELETED**				
GD17	I perform well when I interact with group members with whom I am familiar.	0.83	-2.1	0.85	-1.7	0.77
GD18	Talking about familiar topics make me feel more confident.	0.8	-2.5	0.8	-2.4	0.78
GD19	Interacting with group members I am familiar with makes me feel more confident.	0.91	-1.1	0.93	-0.8	0.75

Table 4 *(continue)*

No	Item	Infit		Outfit		PT-Measure CORR
		MNSQ	ZSTD	MNSQ	ZSTD	
Means		0.98	-0.3	1.01	0.1	
P.SD		0.14	1.8	0.15	1.7	
Reliability of Item difficulty measures		0.98				
Item separation		6.28				
Reliability of person ability		0.92				
Person Separation		3.37				
Variance explained by measures		52.7%				
Unexplained variance in 1st contrast		4.07 (No issue & Disattenuated correlations are 1 or closer to 1)				

separation index (3.37) > 2. As for the point-measure correlation coefficients, all items were found to have positive values, ranging from 0.64 to 0.78. It means that all items were working in the same direction to define the construct of confidence in English oral communication. However, one item (i.e. GD 16) was deleted because it was a misfit (infit > 1.5). Other items' infit and outfit mean-square values were within the accepted range (0.5 to 1.5), indicating that they contributed meaningfully to the measured construct. In Table 4, the variance explained by the measures was 52%, and the largest factor extracted from the residuals was equivalent to 4.07, which had a strength of about four items. Therefore, it would not affect the scale as all other indicators have been met. It is supported by the values of Disattenuated correlations, which were one or very close to 1 (Linacre, 2019).

Self-Confidence in English Oral Communication

Overall, Figure 2 shows that the participants lack confidence in English oral interaction
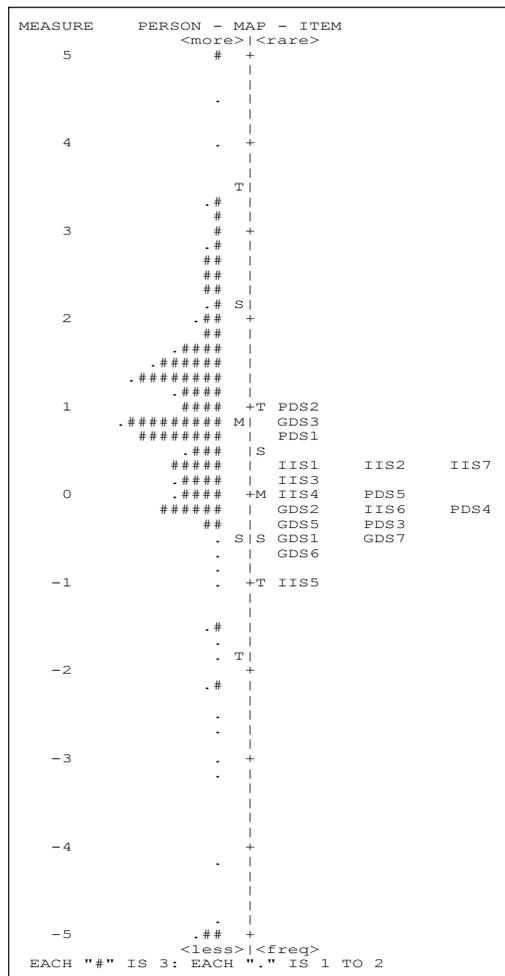
Figure 2. Person-Item Map (interview **IV**, paired discussion **PD** and group discussion **GD**)

or communication in individual interviews, paired or group discussions. They were easy to endorse the items on self-confidence (0.76 logits), which was negative. Table 5 indicated that students did not feel confident when asked to speak about unfamiliar topics, speak to unfamiliar people, or not have enough time to prepare regardless of the context, either an individual interview, paired or group discussions. It is indicated by their high endorsement of the items placed at the bottom part of the interval scale. All

the items interacted on familiar topics or familiar people, and the participants had enough time to prepare in all contexts. On the contrary, the items they stated were not very worried, not frightened, and not nervous or not concentrated when they were asked to interact with other people, talk on unseen topics and interact without preparation received low endorsement.

Though the participants showed a lack of confidence in oral communication on unfamiliar topics, interacting with unfamiliar

Table 5

*Item difficulty measures (confidence in English oral communication: individual interview **IV**, paired discussion **PD** and group discussion **GD**)*

No	Item	Difficulty Measures	S.E
PD 9	I worry when I do not know what my peer is going to ask me.	1.05	0.07
GD 15	I worry when I do not know what my group members are going to ask me.	0.83	0.07
PD 8	I find tasks that require my immediate response (answer) frightening.	0.61	0.07
IV 7	My performance on a task depends on the interviewer that I get.	0.40	0.08
IV2	I find tasks that require my response (answer) without preparation frightening.	0.37	0.08
IV1	Tasks that require me to respond (answer) immediately make me nervous.	0.32	0.08
IV3	I worry when I do not know what the interviewer is going to ask me.	0.16	0.08
PD12	My performance on a task depends on the peer that I get.	0.07	0.08
IV4	I perform well on tasks that give me time to prepare. (Individual)	0.00	0.08
PD11	I perform well on tasks that require me to interact with a peer.	-0.11	0.08
GD14	I like tasks where I am given time to prepare my response (answer) while the other group members take turns speaking.	-0.19	0.08
IV 6	Interacting with an interviewer I am familiar with makes me feel comfortable.	-0.20	0.08
PD10	I perform well on tasks that give me time to prepare. (Paired)	-0.34	0.08
GD17	I perform well when I interact with group members with whom I am familiar.	-0.40	0.08
GD13	Tasks that give me time to prepare my response (answer) make me feel comfortable and relaxed.	-0.46	0.08
GD19	Interacting with group members I am familiar with makes me feel more confident.	-0.50	0.08
GD18	Talking about familiar topics make me feel more confident.	-0.62	0.08
IV 5	Talking about familiar topics make me feel more confident.	-0.98	0.09

people or interacting without enough time or preparation, it is imperative to highlight that their confidence differed based on the context (i.e. individual task, paired or group discussions). Figure 3 shows the means and the hierarchy order of the items under each context. They mostly lacked confidence in interacting in pairs (Mean = 0.26), followed by individual interviews (0.01), and in groups (-.022). Having investigated the items under each category, it is noticeable that the participants were more confident in interacting on familiar topics, with familiar people and having enough time to prepare (Figure 3).

Psychometric Properties of Perceived Ability (Can-Do) Scale in English Oral Communication

A 10-item scale was used to measure the participants' perceived ability (Can-Do Statements) on their oral communication in the English language. The results in Table 5 indicated that the item reliability was high (0.91), with separation index (3.15) > 2; and the person ability reliability was also high (0.94) with person separation index (2.89) > 2. Moreover, all items had positive point-measure correlation coefficients, ranging from 0.77 to 0.88. It means that all items were working in the same direction to define
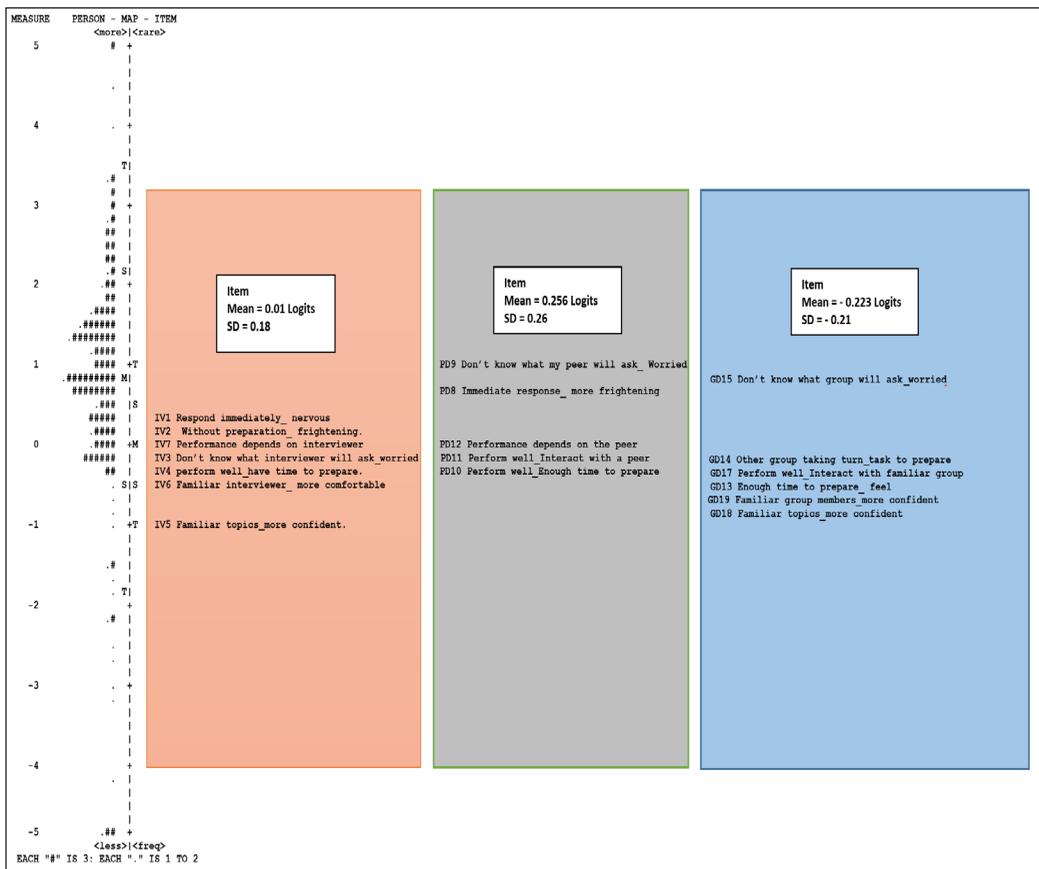
Figure 3. Person-Item Map (interview **IV**, paired discussion **PD** and group discussion **GD**)

the construct (i.e., perceived ability in oral communication in English). All items also had infit and outfit mean-square values within the recommended range (0.5 - 1.5), indicating meaningfully to the measured construct. Table 6 also shows the variance explained by the measures was 66.8%, and the largest factor extracted from the residuals was equivalent to 2.00, which has a strength of about two items. Thus, it shows that the scale's unidimensionality was not violated.

Table 6
Reliability, separation item fit statistics and point-measure correlation coefficients (Can-Do statements)

No	Item	Infit		Outfit		PT-Measure CORR
		MNSQ	ZSTD	MNSQ	ZSTD	
DO1	I can take part effortlessly in any conversation or discussion.	1.28	3.1	1.33	3.6	0.80
DO2	I can have a good familiarity with idiomatic expressions and colloquialisms.	1.40	4.4	1.44	4.7	0.77
DO3	I can express myself fluently and spontaneously without much obvious searching for expressions.	1.06	0.7	1.05	0.6	0.84
DO4	If I do have a problem, I can restructure the conversation without stopping any interaction.	0.8	-2.5	0.82	-2.3	0.87
DO5	I can use language flexibly and effectively for social and professional purposes.	0.88	-1.5	0.86	-1.7	0.87
DO6	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers possible.	0.75	-3.3	0.74	-3.4	0.87
DO7	I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life.	1.12	1.4	1.09	1.1	0.84
DO8	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest.	1.06	0.7	1.04	0.5	0.85
DO9	I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	0.81	-2.4	0.8	-2.5	0.88
DO10	I can present a clear, smoothly-flowing description or argument in style appropriate to the context with an effective logical structure.	0.73	-3.6	0.73	-3.5	0.88
Means		0.99	-0.3	0.99	-0.3	
P.SD		0.22	2.6	0.23	2.7	
Item difficulty measure Reliability				0.91		
Item separation				3.15		
Person ability reliability				0.94		
Person separation				2.89		
Raw variance explained by measures				66.8%		
Unexplained variance in 1st contrast				2.00		

Perceived Ability in English Oral Communication

Overall, the participants were positive (not high) toward their oral interaction in English. The person ability means (0.31 logits) could be deduced, higher than the item difficulty mean (0.00 logits). However, the results in Table 7 and Figure 4 show that they could not interact fluently or spontaneously because they might have not enough vocabulary and expressions that helped them interact mainly in unfamiliar situations. Therefore, it was difficult for them to endorse the items at the upper part of the scale. On the other hand, they felt they could interact in familiar situations and on topics related to their social and professional settings.

```
MEASURE     PERSON - MAP - ITEM
              <more>|<rare>
     7        .## +
              ## |
     6           +
                 |
            . T|
     5           +
              ## |
            #### |
     4        ## +
              ## |
              ## |
     3      . # S+
          .###### |
           ##### |
           .### |
     2      ####### +
        .########### |
        ######### |
          ##### |
     1           +
          .######## |T CDS2
           ###### M|
          ######## |S CDS3   CDS4    CDS6
     0 ############## +M CDS1
           .### |S CDS10  CDS5    CDS8    CDS9
             ## |   CDS7
          .#### |T
    -1      . # +
              |
             . # |
           ## S|
    -2      .## +
            ### |
              # |
    -3      ###### +
             ## |
            .## |
            ### |
    -4      .## +
            T |
             ## |
    -5           +
               # |
               # |
    -6      ###### +
              <less>|<freq>
EACH "#" IS 2: EACH "." IS 1
```
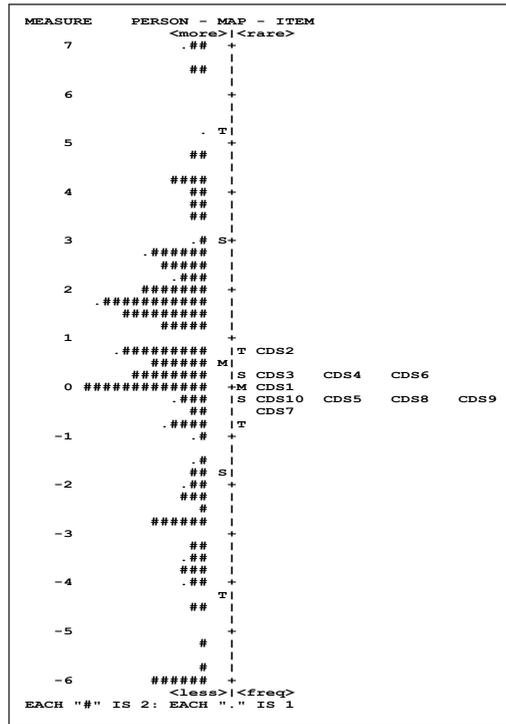
Figure 4. Person-Item Map (Can Do Statements) in English oral communication skills

Table 7
Item difficulty measures (Can Do statements)

No	Item	Difficulty Measures	S.E
DO 2	I can have a good familiarity with idiomatic expressions and colloquialisms.	0.70	0.1
DO4	If I do have a problem, I can restructure the conversation without stopping any interaction.	0.32	0.1
DO 3	I can express myself fluently and spontaneously without much obvious searching for expressions.	0.29	0.1
DO6	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers possible.	0.19	0.1
DO1	I can take part effortlessly in any conversation or discussion.	-0.03	0.1
DO10	I can present a clear, smoothly-flowing description or argument in style appropriate to the context with an effective logical structure.	-0.13	0.1
DO5	I can use language flexibly and effectively for social and professional purposes.	-0.15	0.1
DO8	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest.	-0.3	0.1
DO9	I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	-0.37	0.1
DO7	I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life.	-0.52	0.1

RESULTS OF INFERENTIAL ANALYSIS

The inferential analyses (One-way analysis of variance (ANOVA) and independent samples t-test were conducted to identify the significant differences in mean scores of English-speaking anxiety, confidence, and perceived ability in English language oral communication with selected demographic variables (i.e., gender, academic year, university type/category and faculty. The results in Table 8 show no statistically significant differences in the mean scores of speaking anxiety, confidence, and perceived ability in English language oral communication for both students' academic year and gender, $p>.05$. In contrast, it shows a significant difference in speaking anxiety mean scores for university type/category (comprehensive and other two research universities), ($p<.05$). Post-hoc analysis using the Tukey test indicated that the students who came from research university (A) had the highest level of speaking anxiety (M= .8638 logits), which significantly differed from the other two universities, comprehensive M= .0441 and research university (B) M= -.0116 logits, $p<0.5$. Furthermore, the ANOVA results show a significant difference in students' confidence in speaking English mean scores across faculty (Science, Information and Technology, and Engineering), $p<.05$. Post-hoc analysis using Tukey test indicated that the students in Engineering and Information and Technology Faculties had higher significant difference from those in Science Faculty (1.04 logits, 1.03 logits and .208 logits respectively), $p< .05$. Finally, the same results were found for the perceived ability in English oral communication. The students in Engineering and Information and Technology Faculties had a higher significant difference from those in Science Faculty (1.013 logits, .688 logits and -.760 logits), $p < .05$.

Table 8
Results of *One-way analysis of variance (ANOVA) and Independent samples t-test*

Construct	Variable	N	Mean±SD (Logit)	*p value*
	Gender			
Anxiety	Male	116	-.0381±2.383	.055
	Female	184	.467±1.899	
	Academic Year			
	Third	140	.451±1.560	.156
	Fourth	160	.114±2.489	
	University Category			
	Comprehensive	110	.044±2.033	.006*
	Research A	90	.864±1.787	
	Research B	100	-.012±2.362	
	Faculty			
	Science	100	.411±1.224	
	Information & Technology	100	.503±1.504	.093
	Engineering	100	-.100±3.080	

Table 8 *(continue)*

Construct	Variable	*N*	Mean±SD (Logit)	*p value*
	Gender			
Confidence	Male	116	.537±1.743	.081
	Female	184	.896±1.768	
	Academic Year			
	Third	140	.675±1.618	.460
	Fourth	160	.826±1.885	
	University Category			
	Comprehensive	110	.489±2.006	.128
	Research A	90	.962±1.584	
	Research B	100	.862±1.607	
	Faculty			
	Science	100	.208±2.753	
	Information & Technology	100	1.03±.857	.001*
	Engineering	100	1.04±.788	
	Gender			
Can-Do	Male	116	.313±2.652	1.000
Perceived ability	Female	184	.314±3.186	
	Academic Year			
	Third	140	.144±2.859	0359
	Fourth	160	.462±3.094	
	University Category			
	Comprehensive	110	.527±3.037	.232
	Research A	90	.516±3.162	
	Research B	100	-.104±2.743	
	Faculty			
	Science	100	-.760±4.269	
	Information & Technology	100	.688±1.828	.000*
	Engineering	100	1.013±1.887	

* The mean difference is significant at the 0.05 level

DISCUSSION

This study suggests that, in general, the participants had experienced English language speaking anxiety that could affect their English oral communication that concurs with Salem and Al Dyiar (2014), who found a negative relationship between English speaking anxiety and speaking fluency. In addition, speaking anxiety adversely affects students' oral presentation and conversation with English native speakers (Bux et al., 2015). As a result, the participants in this study mainly felt worried and anxious before they were asked to speak. They often lost their control before they delivered the speech in front of others. They prefer to talk about familiar topics and interact with familiar persons, and they need enough time to prepare for the communication. These findings show that the participants demonstrate situational anxiety since it appears in

specific situations (MacIntyre, 1999), for example, in unfamiliar contexts where the learners are required to use the language. This anxiety ends or reduces when learners become more familiar with the new context (Riasati, 2011). The participants also tend to think about others whom they believed to be better than them. It might be because most participants showed a lack of confidence in English oral communication during individual interviews, paired or group discussions (i.e. regardless of the context). Kamaruddin et al. (2019) found that low self-esteem and social anxiety contributed significantly to the level of both listening and speaking anxieties among Malaysian university students majoring in non-English programs. On the same note, Desmaliza and Septiani (2017) found a significant correlation between students' perceived ability and speaking skills, positively influencing students' performance in oral communication activities. In this study, the participants perceived that they could not interact fluently or spontaneously because they might not have enough vocabulary and expressions to help them interact mainly in unfamiliar situations.

The finding agrees with Amiri and Puteh (2018)'s study, which found that among the factors that caused anxiety to international postgraduate students in different Malaysian universities are inadequate linguistic competency and inadequate knowledge of the presentation's content. On the other hand, they perceived a higher ability to interact about familiar situations and topics related to their social and professional

settings. It is because they had enough vocabulary and expressions with which to interact. Sadighi and Dastpak (2017) found inadequate vocabulary knowledge as one of the main sources of speaking anxiety among students. Ahmed et al. (2017) found that the participants with insufficient linguistic competence led to speaking anxiety and affected their oral communication since they could not express themselves in a wide range of communicative situations. Zahiri et al. (2017) found that speaking anxiety and perceived ability affect students' monologue speaking skills. The former was found to have a negative effect on speaking performance, while a higher perception of their ability has made them better in speaking performance. Besides, lack of vocabulary and expressions result in the participants to have negative or low perception on their self-confidence. Stenstrom (2014) maintained that confidence is an important factor in spoken interaction as a two-way process. This idea was coined in Tsou (2005) who found that high self-confidence is positively correlated with oral performance as it determines the learners' willingness to communicate. Other researchers found the exact relationships between self-confidence and speaking competency (Gürler, 2015; Mandokhail, 2018; Tridinanti, 2018).

Though the surveyed participants showed a lack of confidence in oral communication on unfamiliar topics, interacting with unfamiliar people or interacting without enough time for prior preparation, they lacked confidence in interacting in pair, individual and group interviews. They

might feel more comfortable in groups as they took additional time to prepare, and the same applied when they were in individual interviewing. McDonough (2004) reported that learners improved their speaking skills when put in group and pair tasks. These findings are alarming as university students are expected to be able to communicate at the C1 level of the CEFR, where communication expectations are about the ability to smoothly engage in synchronous discourse involving a wide range of social, academic and professional topics without much searching and assistance (Council of Europe, 2018).

The independent samples t-test shows no significant differences in the mean scores of speaking anxiety, confidence in speaking English and perceived ability for gender and student academic year. However, the female students reported a higher mean score in speaking anxiety, as indicated by the mean differences. The female students might be more concerned about their appearance in front of others in the conversation, which might affect their confidence and ability in oral communication. Literature has reported different findings in terms of gender and ESL or EFL speaking anxiety. For instance, Batiha et al. (2016) found no significant differences in the mean scores of speaking anxiety in research conducted on the factors of speaking anxiety among EFL university learners due to gender. The same findings were reported by Ahmed et al. (2017), who conducted research to identify the factors responsible for ESL oral communication anxiety among postgraduate students in Pakistan. Other studies reported that females scored higher levels of EFL speaking anxiety (Ahmed & Alansari, 2004), whereas Elaldi (2016) reported that male EFL University students had higher speaking anxiety levels than female students. It seems these variations depend on research contexts.

For confidence and perceived ability, the one-way ANOVA analysis indicated that engineering students were more confident and perceived higher ability in oral communication than their counterparts in science or information and technology faculties despite the type of university. This result indicated that undergraduate engineers might have realised the importance of English speaking skills in their future careers. Past studies found that many engineering graduates could not secure a job due to their inability to command good English (Kakepoto, 2013; Sheth, 2016; Ting et al., 2017). Over the years, research has focused on the significance of English for engineers at the workplace (Božić & Pintarić, 2018; Dewi et al., 2015; Hossain, 2013; Rajprasit & Hemchua, 2015; Spence & Liu, 2013). Sheth (2016) proclaims that engineering employers give priority to graduate engineers with competence in English speaking over their counterparts who tend to be highly tech-savvy but with a low level of English language speaking skills. As for Science undergraduate students, there is a need to improve awareness about the importance attached to the English language in their future profession. It would motivate them towards enhancing their competence in spoken English.

CONCLUSIONS AND IMPLICATIONS

This study mainly aimed to analyse speaking anxiety, confidence and perceived ability in English oral communication among Science and Technology undergraduates in comprehensive and research universities in Malaysia. Furthermore, it aimed to find the significant differences in the three sub-constructs based on the selected demographic variables. Overall, the participants experienced speaking anxiety, low confidence, and yet high-perceived ability in English oral communication, with significant differences in mean scores of English speaking anxiety across university category and confidence and perceived ability due to students' specialisation (i.e., science, technology and information and engineering). However, engineering students were more confident and could perform better in oral communication than their science or information and technology faculties counterparts. In addition, the findings show that the participants felt more confident and more able in familiar situations, communicating on familiar topics to familiar audiences, contrary to the expectations for the C1 level in the CEFR.

The students enrolled in science, engineering and technology programmes need more training on English oral communication. Lecturers need to encourage students to practice oral communication in English in front of the class to improve their self-confidence and mitigate their anxiety. Furthermore, the topics should be varied in familiarity and content to prepare them for future careers. In other words, they should be provided with authentic situations to practice and improve their levels in English oral communication. Besides, students should be encouraged to do oral presentations individually, in pairs and groups, reflecting the real discourse they are expected to engage in as they enter the working world. Furthermore, the university should provide students with training modules on enhancing their confidence and perceived ability as these factors play a substantial role in students' level of English oral communication. Contextually, the programs could also be geared towards more social-like situations and the classroom setting so that undergraduates can make English a social practice. As a result, students will be more qualified for future employability.

ACKNOWLEDGEMENTS

REFERENCES

Ahmed, M., & Alansari, B. (2004). Gender differences in anxiety among undergraduates from ten Arab countries. *International Journal of Social Behavior and Personality, 32*(7), 649-655. https://doi.org/10.2224/sbp.2004.32.7.649

Ahmed, N., Pathan, Z. H., & Khan, F. S. (2017). Exploring the causes of English language speaking anxiety among postgraduate students of University of Baluchistan, Pakistan. *International Journal of English Linguistics*, *7*(2), 99-105. https://doi.org/10.5539/ijel.v7n2p99

Akkakoson, S. (2016). Speaking anxiety in English conversation classrooms among Thai students. *Malaysian Journal of Learning and Instruction, 13*, 63-82. https://doi.org/10.32890/mjli2016.13.1.4

Alawiyah, T. (2018). Speaking self-efficacy and EFL student teachers' speaking achievement. *Jurnal Pendidikan dan Pengajaran, 5*(1), 87-96. https://doi.org/10.19109/ejpp.v5i1.2052

Alderson, J. C. (2017). The Common European Framework of Reference for Languages (CEFR) for English language assessment in China (Foreword to special issue). *Language Testing in Asia, 7*, Article 20. https://doi.org/10.1186/s40468-017-0049-9

Ali, M. A. K. (2017). English language anxiety: Development and validation of a brief measure. *International Journal of Psychology and Educational Studies, 4*(2), 42-53. https://doi.org/10.17220/ijpes.2017.02.005

Al-Sibai, D. (2004). *Promoting oral fluency of second language learners: Educational linguistics*. King Saud University.

Amiri, F., & Puteh, M. (2018). Oral communication apprehension among international doctoral students. *English Language Teaching, 11*(2), 164-171. https://doi.org/10.5539/elt.v11n2p164

Badrasawi, K. J., Solihu, A., & Ahmad, T. B. (2020). Second language speaking anxiety among Malaysian postgraduate students at a Faculty of Education. *International Journal of Education and Literacy Studies, 8*(2), 54-61. https://doi.org/10.7575/aiac.ijels.v.8n.2p.54

Batiha, J., Noor, N., & Mustaffa, R. (2016). Speaking anxiety among English as a foreign language learner in Jordan: Quantitative research. *International Journal of Education and Research, 4*(10), 63-82.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Routledge.

Božić, D. L., & Pintarić, L. (2018). Comparison of employers' and students' perceptions regarding communication skills. *The Journal of Teaching English for Specific and Academic Purposes, 6*(1), 63-82. https://doi.org/10.22190/jtesap1801063b

Brown, H. D. (2004). *Teaching by principles: An interactive approach to language pedagogy*. Harper and Row Publisher.

Bux, I. G., Ali, F. B., Umrani, T., & Ali, S. L. (2015). English speaking anxiety: A study of undergraduates at Lasbela University, Baluchistan. *International Researchers, 4*(1), 54-61.

Chin, C. (2016, February 28). Nation builders wanted. *The Star Online*. https://www.thestar.com.my/news/nation/2016/02/28/nation-builders-wanted-malaysias-dream-of-becoming-a-highincome-nation-by-2020-could-crumble-if-we-d/

Clement, A., & Murugavel, T. (2015). English for employability: A case study of the English language training need analysis for engineering students in India. *English language teaching, 8*(2), 116-125. https://doi.org/10.5539/elt.v8n2p116

Council of Europe (2018). *Common European Framework of Reference for languages: Learning, teaching, assessment - Companion volume.* Council of Europe Publishing.

Crystal, D. (2004). *The Cambridge encyclopedia of the English language.* Ernst Klett Sprachen Verlag.

David, A., Thang, S., & Azman, H. (2015). Accommodating low proficiency ESL students' language learning needs through an online writing support system. *Journal of Social Sciences and Humanities, Special Issue1*(2015), 118-271.

Desmaliza, D., & Septiani, T. (2017). Students' self- efficacy and their speaking skill at lower secondary school. In *International Conference on Education in Muslim Society (ICEMS 2017)* (pp. 122-127). Atlantis Press. https://doi. org/10.2991/icems-17.2018.24

Dewi, Z. Z., Azlin, M. Z., Bakar, S. B. A., & Sumaiyyah, O. (2015). English language communication needs at workplace as perceived by students. *Journal of Science & Engineering Technology, 3*(2), 27-31.

Dordinejad, F. G., & Ahmadabad, R. M. (2014). Examination of the relationship between foreign language classroom anxiety and English achievement among male and female Iranian high school students. *International Journal of Language Learning and Applied Linguistics World*, *6*(4), 446-460.

Elaldi, S. (2016). Foreign language anxiety of students studying English Language and Literature: A Sample from Turkey. *Journal of Educational Research and Reviews, 11*(6), 219-228. https:// doi.org/10.5897/err2015.2507

Free Malaysia Today. (2017, November 13). Economist: Poor English, lack of experience costing grads jobs. *Free Malaysia Today.* https:// bit.ly/2XFBWXk

Gürler, İ. (2015). Correlation between self-confidence and speaking skill of English language teaching and English language and literature preparatory students. *Current Research in Social Sciences*, *1*(2), 14-19.

Hart, L. E. (1989). Classroom processes, sex of student, and confidence in learning mathematics. *Journal for Research in Mathematics Education, 20*(3), 242-260. https://doi.org/10.2307/749514

Horwitz, E. K., Horwitz, M. B., & Cope, J. A. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*, 125-132. https://doi. org/10.1111/j.1540-4781.1986.tb05256.x

Hossain, J. (2013). ESP needs analysis for engineering students: A learner-centered approach. *Journal of PU, Part B, 2*(2), 16-26.

Hudson, P., & Hudson, S. (2003). An evaluation of an English language immersion programme for ESL pre-service teachers. *GEMA Online Journal of Language Studies, 3*(1), 42-65.

Ismail, N. A. (2011). Graduates' characteristics and unemployment: A study among Malaysian graduates. *International Journal of Business and Social Science, 2*(16), 94-102.

Kakepoto, I. (2013). *Oral presentation and communication skills in workplace environments among engineers and engineering students in Pakistan* [Doctoral thesis, Universiti Teknologi Malaysia]. Universiti Teknologi Malaysia Institutional Repository. http://eprints.utm.my/ id/eprint/44038/

Kalanzadeh, G., Mahnegar, F., Hassannejad, E., & Bakhtiarvand, M. (2013). The influence of EFL students' self-esteem on their speaking skills. *The International Journal of Language Learning and Applied Linguistics World, 2*(2), 77-84. https://doi.org/10.5430/wjel.v2n4p24

Kamaruddin, S. F., Daneil, I. L., Rijeng, J. S., Bohari, M. M., & Eng, T. H. (2019). The impact of listening and speaking anxieties on the Fourth

Industrial Revolution: What can educators do? *Language & Communication*, *6*(1), 409-423.

Kassim, N. L. A., & Zubairi, A. M. (2003, July 22-23). *Interaction between test-taker characteristics, task facets and L2 oral proficiency test performance [Paper presentation]. Language Testing Research Colloquium, Reading, United Kingdom.*

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107-118. https://doi.org/10.1037/0278-7393.6.2.107

Kumar, U. H. (2018). English language anxiety and academic achievement. *North Asian International Research Journal of Social Science & Humanities, 4*(3), 138-147.

Linacre, J. M. (2018). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs (Program Manual No. 4.1.0)*. www.winsteps.com

Linacre, J.M. (2019). A user's guide to Winsteps & Ministeps (Rasch-Model Computer Programs). (Program Manual 4.4.5). www.winsteps.com

Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading &Writing Quarterly, 19*(2), 119-137. https://doi.org/10.1080/10573560308223

MacIntyre, P. D. (1999). Language anxiety: A review of literature for language teachers. In D. J. Young (Ed.), *Effect in Foreign Language and Second language learning* (pp.24-43). McGraw Hill Companies.

Malaysian Employers Federation. (2016). *English proficiency critical to enhance employability of graduates*. https://www.mef.org.my/Attachments/PR161124b.pdf

Mandokhail, S., Khan, F. R., & Malghani, M. (2018). Impact of ESL learners' self-esteem on their oral proficiency. *International Journal of English Linguistics*, *8*(3), 210-222. https://doi.org/10.5539/ijel.v8n3p210

McCroskey, J. (1970). Measures of communication-bound anxiety. *Speech Monographs, 37*(4), 269-277. https://doi.org/10.1080/03637757009375677

McDonough, K. (2004). Learner-learner interaction during pair and small group activities in a Thai EFL context. *System, 32*(2), 207-224. https://doi.org/10.1016/j.system.2004.01.003

Ministry of Education. (2012). *Malaysia education blueprint 2013-2025*. Kementerian Pendidikan Malaysia.

Mulyono, H., Sari, R. N., & Ningsih, S. K. (2019). Factors contributing to EFL students' speaking anxiety. *Register Journal, 12*(1), 13-27. https://doi.org/10.18326/rgt.v12i1.13-27

Musa, N. C., Koo, Y. L., & Azman, H. (2012). Exploring English language learning and teaching in Malaysia. *GEMA: Online Journal of Language Studies, 12*(1), 35-51.

Nair, G. K. S., Rahim, R. A., Setia, R., Husin, N., Sabapathy, E., Jalil, N. A. A., & Yusoff, N. I. M. (2012). Malaysian graduates English adequacy in the job sector. *Asian Social Science, 8*(4), 143-147. https://doi.org/10.5539/ass.v8n4p143

Nishanti, R. (2018). The importance of learning English in today world. *International Journal of Trend in Scientific Research and Development, 3*(1), 871-874. https://doi.org/10.31142/ijtsrd19061

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-578. https://doi.org/10.3102/00346543066004543

Pandarangga, S. (2015). The transformation of English as a global language in the world. *LiNGUA*, *10*(2), 90-96. https://doi.org/10.18860/ling.v10i2.3132

Pappamihiel, N. E. (2002). English as a second language students and English language anxiety: Issues in the mainstream classroom. *Research in the Teaching of English, 36*(3), 327-355.

Pennycook, A. (2014). *The cultural politics of English as an international language*. Routledge. https://doi.org/10.4324/9781315843605

Rajprasit, K., & Hemchua, S. (2015). The English language & communication in the international workplace: An examination of Thai computer engineering professionals. *The Southeast Asian Journal of English Language Studies, 21*(3), 109-124.

Ramamuruthy, V. (2019). The effects of task-based approach on speaking anxiety among ESL low proficiency diploma students. *Universal Journal of Educational Research, 7*(6), 1363-1375. https://doi.org/10.13189/ujer.2019.070604

Reddy, R. (2016). Development of communication skills in English for engineering graduates. *IOSR Journal of Humanities and Social Science (IOSR-JHSS), 21*(5), 102-104.

Riasati, M. J. (2011). Language learning anxiety from EFL learners' perspective. *Middle-East Journal of Scientific Research, 7*(6), 907-914.

Roshid, M. M., & Chowdhury, R. (2013). English language proficiency and employment: A case study of Bangladeshi graduates in Australian employment market. *Online Submission, 3*(1), 68-81. https://doi.org/10.13054/mije.13.06.3.1

Rusli, R., Yunus, M. M., & Hashim, H. (2018, April 23-24). Low speaking proficiency among the Malaysian undergraduates: Why and how? In e-*Persidangan Antarabangsa Sains Sosial dan Kemanusiaan* (pp. 678-689). Kolej Universiti Islam Antarabangsa Selangor.

Sadighi, F., & Dastpak, M. (2017). The sources of foreign language speaking anxiety of Iranian English language learners. *International Journal of Education and Literacy Studies, 5*(4), 111-115. https://doi.org/10.7575/aiac.ijels.v.5n.4p.111

Salem, M. S., & Al Dyiar, M. A. (2014). The relationship between speaking anxiety and oral fluency of special education Arab learners of English. *Asian Social Science, 10*(12), 170-176. https://doi.org/10.5539/ass.v10n12p170

Scrivener, J. (2005). *Learning teaching: A guidebook for English teachers*. Macmillan Education.

Shakir, R. (2009). Soft skills at the Malaysian institutes of higher learning. *Asia Pacific Education Review, 10*(3), 309-315. https://doi.org/10.1007/s12564-009-9038-8

Sheth, T. D. (2016). Communication skill: A prerequisite for engineers. *International Journal on Studies in English Language and Literature (IJSELL), 3*(7), 51-54.

Shrauger, J. S., & Schohn, M. (1995). Self-confidence in college students: Conceptualization, measurement, and behavioral implications. *Assessment, 2*(3), 255-278. https://doi.org/10.1177/1073191195002003006

Spence, P., & Liu, G. (2013). Engineering English and the high-tech industry: A case study of an English needs analysis of process integration engineers at a semiconductor manufacturing company in Taiwan. *English for Specific Purposes, 32*, 97-109. https://doi.org/10.1016/j.esp.2012.11.003

Spielberger, C. D. (1983). *Manual for the state-trait anxiety* (From Y). Consulting Psychologists Press.

Stenstrom, A. B. (2014). *An introduction to spoken interaction*. Routledge. https://doi.org/10.4324/9781315845449

Sunyi, D. P. (2017). Investigating students' self-efficacy of public speaking. *International Journal of Education and Research, 5*(10), 97-108.

Tanveer, M. (2007). *Investigation of the factors that cause language anxiety for ESL/EFL learners in learning speaking skills and the influence it casts on communication in the target language*

[Master's thesis, University of Glasgow]. ResearchGate. https://www.researchgate.net/publication/293263722

Ting, S., Marzuki E., Chuah K., Misieng, J., & Jerome, C. (2017). Employers' views on the importance of English proficiency and communication skill for employability in Malaysia. *Indonesian Journal of Applied Linguistics, 7*(2), 315-327. https://doi.org/10.17509/ijal.v7i2.8132

Tridinanti, G. (2018). The correlation between speaking anxiety, self-confidence, and speaking achievement of undergraduate EFL students of private university in Palembang. *International Journal of Education and Literacy Studies, 6*(4), 35-39. https://doi.org/10.7575/aiac.ijels.v.6n.4p.35

The Sun Daily. (2018, January 22). Language main barrier to grads getting jobs (Updated). *The Sun Daily*. https://www.thesundaily.my/archive/language-main-barrier-grads-getting-jobs-updated-BUARCH520890

Tsou, W. (2005). The effects of cultural instruction on foreign language learning. *RELC Journal, 36*(1), 39-57. https://doi.org/10.1177/0033688205053481

UNESCO. (2012). *Graduate employability in Asia*. UNESCO Bangkok.

Ur, P. (1996). *A course in language teaching: Practice and theory*. Cambridge University Press.

Wijewardene, L., Yong, D., & Chinna, K. (2014). English for Employability-the need of the hour for Sri Lankan graduates. *British Journal or Arts and Social Sciences, 17*(1), 137-145.

Xue, J., & Zuo, W. (2013). English dominance and its influence on international communication. *Theory and Practice in Language Studies, 3*(12), 2262-2266. https://doi.org/10.4304/tpls.3.12.2262-2266

Yim, S. Y., & Yu, Y. L. (2011). Validating the English learning anxiety scale for primary school students in Korea. *English Teaching, 66*(2), 101-123. https://doi.org/10.15858/engtea.66.2.201106.101

Zahiri, A., Sibarani, B., & Sumarsih (2017). To what extent do anxiety and self-efficacy effect the EFL students' English monologue speaking skill? *International Journal of Education and Research, 5*(9), 151-160.

Zhang, R., & Zhong, J. (2012). The hindrance of doubt: Causes of language anxiety. *International Journal of English Linguistics, 2*(3), 27-33. https://doi.org/10.5539/ijel.v2n3p27

SOCIAL SCIENCES & HUMANITIES

Speaking Assessments by Japanese English Teachers Pre and Post Implementation of CEFR in the Midst of a Global Pandemic

Emily Ee Ching Choong¹, Pravina Manoharan²* and Souba Rethinasamy³

¹*Niigata City Board of Education, 951-8063 Niigata City, Niigata Prefecture, Japan*
²*School of The Arts, Universiti Sains Malaysia, 11800 Gelugor, Penang, Malaysia*
³*Faculty of Language and Communication, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia*

ABSTRACT

Amid a global pandemic, while schools in many parts of the world were closed to adhere to quarantine orders, schools in Japan resumed face-to-face classes after only a month of closure with strict adherence to COVID-19 guidelines and standard operating procedures (SOP). This study examined how speaking assessments were administered face-to-face for Grade 5 and 6 elementary school students prior to and after introducing the Common European Framework of Reference (CEFR) and amid a global pandemic between April to October 2020. The paper also reports the challenges and strategies employed in carrying out the speaking assessments following the CEFR while adhering to the SOP. The study employed a qualitative research method that utilised semi-structured interviews to elicit information from four teachers who taught in eight schools within Niigata City, Japan. Findings suggest that prior to the implementation of CEFR, not all teachers carried out speaking assessments. However, the implementation of CEFR emphasised the need to teach speaking and carry out speaking assessments. The CEFR also served as guidance for the teachers in preparing the assessment scoring rubrics. The results also showed that the speaking assessments were implemented individually instead of in groups before the pandemic and the presence of the masks, which increased the student's anxiety and affected their performance. However, the teachers employed various strategies to overcome the challenges by modifying the assessment tasks and utilising web conferencing technology.

Keywords: CEFR, English as a foreign language, Japan, pandemic, speaking assessment

INTRODUCTION

The year 2020 saw a complete shift in how teaching and learning were viewed, particularly in classrooms where the face-to-face mode of delivery was either the only or preferred method of instruction. The COVID-19 pandemic induced a drastic shift in learning systems as schools, colleges and institutions of higher learning adjusted their mode of delivery. They tried to implement and adapt to entire online teaching. While schools worldwide were forced to close and shift all face-to-face classes to the virtual realm (Ghazi-Saidi et al., 2020; Gross & Opalka, 2020; Zhang, 2020), schools in Japan faced a slightly different predicament. All schools in Japan were only closed for one month (from March to April 2020) and after that were ordered to reopen. COVID-19 guidelines were implemented in all schools nationwide to ensure the safety of students and teachers. It included wearing masks at all times and avoiding the 3C's—close contact, closed places and crowded places. These regulations were in line with the guidelines issued by the WHO (2019). Therefore, in April 2020, teachers and school administrators resumed face-to-face classes and continued administering assessments while adhering to WHO's guidelines.

To add to the whirlwind of uncertainties, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan decided to follow through with an English Education Reform plan that was announced in late 2019. This new plan which took effect at the start of the new school year in April 2020, included a curriculum designed based on the Common European Framework of Reference for Languages (CEFR) which is often used as the point of reference for language policy and language education across the globe (Byram & Parmenter, 2012; Little, 2007). In addition, this plan was to make English a formal graded subject for elementary school students in Grade 5 and 6 (age 11 to 12 years old) nationwide and ensure standardised assessments across the board. Prior to the 2019 plan, English was taught in classrooms as a foreign subject but was not formally graded for Grade 5 and 6 students (Carreira, 2006).

In April 2020, the new directive based on CEFR standards required English to be taught for 70 hours, which is approximately two hours per week of 'English as a formally assessed subject' for years 5 and 6 (Nemoto, 2018). The plan comprised new methodologies of delivering and assessing English lessons for elementary Grade 5 and 6 students based on CEFR. The directive from the Board of Education is for teachers to achieve a higher tier of A1 by the end of the school year (March 2021). A1 is the basic user level and refers to the ability "to understand and use familiar everyday expressions and fundamental phrases aimed at the satisfaction of needs of a concrete type. At this level, students should be able to introduce themselves and others, ask and answer questions about personal details such as where they live, people they know and things they have. Students should also be able to interact in a simple way, provided the other person talks slowly and clearly and is prepared to help (Council of Europe, 2020).

While this new directive by the ministry was a commendable effort, its implementation in the middle of a global pandemic posed a problem on how the new curriculum and assessment would be executed, as the conventional classroom setting had now changed. Although CEFR provided a framework for assessing language (listening, speaking, reading and writing), many schools in Japan were left with no concrete outline on assessing speaking tests, particularly amid a global pandemic. While textbooks and manuals were provided, the teachers interviewed for this research felt that there were no proper directives on effectively conducting speaking assessments for their students based on CEFR while simultaneously ensuring they abide by the new COVID-19 guidelines. Therefore, because previous assessment methods could not be administered due to new COVID guidelines, teachers were compelled to develop innovative speaking assessment strategies to ensure students were assessed based on CEFR standards.

Over one year, there have been numerous articles, blog posts and YouTube videos on how teachers worldwide have adopted and adapted to conduct effective online assessments for students. However, literature on face-to-face assessments in schools during the global pandemic is scarce, simply because educational institutions from kindergarten to colleges and universities converted their conventional mode of delivery to online lessons during the pandemic. Hence, a study on how teachers who continued to conduct face-to-face assessments and developed alternative assessments strategies is vital. Furthermore, it presents a crucial understanding of how speaking tests were administered successfully despite COVID-19 SOP restrictions and how such strategies can be continued within the new normal post-pandemic.

This paper, therefore, aims to examine how Japanese English teachers administered speaking assessments for Grade 5 and 6 elementary school students in eight schools within the Niigata Prefecture, Japan prior to, and after the implementation of CEFR amid a global pandemic. The paper further discusses the challenges teachers faced in conducting speaking assessments and their strategies to overcome the challenges of conducting speaking tests while following the CEFR framework and adhering to COVID-19 guidelines.

Objectives

This paper aims to address the following objectives;

1. to discover how Japanese English teachers administered speaking assessments before the CEFR framework was introduced.
2. to discover how Japanese English teachers administered speaking assessments after the CEFR framework was introduced during the pandemic.
3. to identify the challenges faced by Japanese English teachers in assessing speaking during the pandemic and the strategies they developed to overcome the challenges.

LITERATURE REVIEW

The Common European Framework of Reference

The Common European Framework of Reference (CEFR) for Languages comprises learning, teaching and assessment. It is often referred to as the globalisation of language education policy (Behforouz, 2020; Byram & Parmenter, 2012). It was developed by the Council of Europe and first published in 2001. It promotes transparency and coherence in language education. The framework can be applied to the teaching and learning of any language. Thus, it is no surprise that it is an exclusive neutral reference in all educational sectors. According to Little (2006), CEFR has been translated into 37 languages, including Japanese. In some countries, the CEFR has helped "to develop both strategic language policy documents and practical teaching materials. In others, it is becoming the most reliable reference for curriculum planning" (Martyniuk & Noijons, 2007, p. 7). CEFR is a descriptive scheme that is particularly useful in analysing the second language (L2) learners' needs, specifying their learning goals, guiding the development of learning materials and activities, and providing orientation for assessing L2 learning outcomes (Little, 2006). CEFR includes six reference levels, and they are A1 *(Beginner)*, A2 *(Elementary),* B1 *(Intermediate),* B2 *(Upper Intermediate),* C1*(Advanced)* and C2 *(Proficiency)*. Within these levels, A1 and A2 are regarded as basic users, B1 and B2 are independent users, while C1 and C2 are referred to as proficient users.

English Education in Japan

Japan is one of the countries with limited opportunities to practise speaking English in a real-life context due to the lack of people who use the language daily. Besides social circles, independent studying and extra English classes at an *eikaiwa* (英会話 or English conversation school), students are presented with little opportunity to acquire the language outside the classroom (Nemoto, 2018). Moreover, there is no need to use English to communicate when the native language of Japanese is used daily (Tsuboya-Newell, 2017).

English is regarded as a Foreign Language (EFL) in Japan, while Japanese is the first language (L1) and the main medium of instruction for all subjects in schools. However, literature has shown that there has always been little exposure for Japanese students to engage with the English language outside the classroom (Mahoney & Inoi, 2015; Negishi et al., 2013; Nemoto, 2018). As a solution, in 2002, Japan introduced English activities as a part of the government's integrated studies initiative for elementary school students (Nemoto, 2018) to have more practice with the language. Then in 2011, a new subject called Foreign Language Activities was introduced in primary schools across Japan to encourage more engagement with the English language (Negishi et al., 2013).

However, research has shown that the efforts did not yield very promising results, as Japanese students are still not competent in the language. It poses an issue especially when Japan aims to have a bigger global

presence and ensure the Japanese people can communicate more effectively in English (Nemoto, 2018). In a study on the challenges in increasing the teaching hours of English in Japanese schools, Nemoto (2018) noted that there was an inconsistency with how lessons were delivered across the nation since English activities were introduced in 2002. He revealed that different teachers adopted different instruction and content delivery methods, tweaking lesson plans and developing rubrics for assessments.

As a result, it created various learning experiences in the classroom and inconsistent assessments that did not accurately measure learning outcomes. To address this discrepancy, in 2011, Grade 5 and 6 students underwent 35 hours of English classes per year (approximately one hour per week) with lesson plan guidelines provided for teachers to ensure some consistency in the teaching of the language (Mahoney & Inoi, 2015). In addition, it allowed all students to receive an equal number of contact hours with the language across the country. Teachers were also given a clearer idea of how to conduct the lessons from the guidelines given. However, because English was not a formal subject within the curriculum, there was still no standardised testing and grading of the students, despite the increased hours and guided lessons (Mahoney & Inoi, 2015). In addition, Mahoney and Inoi (2015) noted that some teachers had trouble assessing learning outcomes in the classroom as there were reports of teachers conducting their tests. However, because the students were

not formally graded, these tests were again not standardised.

MATERIALS AND METHODS

This research employed a qualitative case study research design using semi-structured interviews for data collection. The qualitative approach was most suited as it allowed the researcher to gather detailed information on how assessments were carried out prior to and after the implementation of CEFR.

The following were the main questions asked during the interview;

1. How did you conduct speaking tests prior to the implementation of CEFR?
2. How were speaking tests conducted after the implementation of CEFR?
3. How did the COVID-19 guidelines affect the way speaking tests are done?
4. What were the challenges you faced in conducting face-to-face speaking assessments while adhering to COVID-19 guidelines?
5. What strategies did you employ or develop to overcome the challenges you faced?

Context of the Study

It must be noted here that the implementation of CEFR and the introduction of formal testing were all part of the government's plans to reform the English education system in Japan even before the pandemic hit. However, instead of putting the plans on hold, the Japanese government decided

to proceed and directed all schools to ensure CEFR standards were met as schools resumed face-to-face teaching after a one-month closure. Therefore, the introduction of CEFR coincided with a period when the entire world was affected by a global pandemic. Hence, when discussions in this paper refer to testing after the implementation of CEFR, it also refers to a period where testing was conducted within a classroom with COVID-19 SOPs in place.

The introduction of CEFR provides a more comprehensive approach to how language is learnt and taught, as the equal focus is placed on four skills of English. Figure 1 shows the expected improvements in English language proficiency for all school levels. For Elementary students, the new CEFR standard requires students to master between 600 to 700 new words during their elementary grade, which lasts for four years. It is a challenging feat

particularly when students were not required to remember new words or be tested on them in the past.

Participants

From April 2018, to improve the way English lessons are conducted, the Niigata City Board of Education hired teachers who have an additional licence for teaching English only [*gaikoku-go senka kyō* or 外国語専科教] (Niigata City, 2020). These teachers are referred to as "MEXT (Ministry of Education, Culture, Sports, Science and Technology) teachers". It was done to ensure that teachers conducted English lessons with a specific qualification for the subject. From April 2020 to March 2021, there were 24 MEXT teachers in Niigata (Niigata City, 2020).

Data were collected from four Japanese English language teachers in Niigata City, Japan. All four teachers have a bachelor's
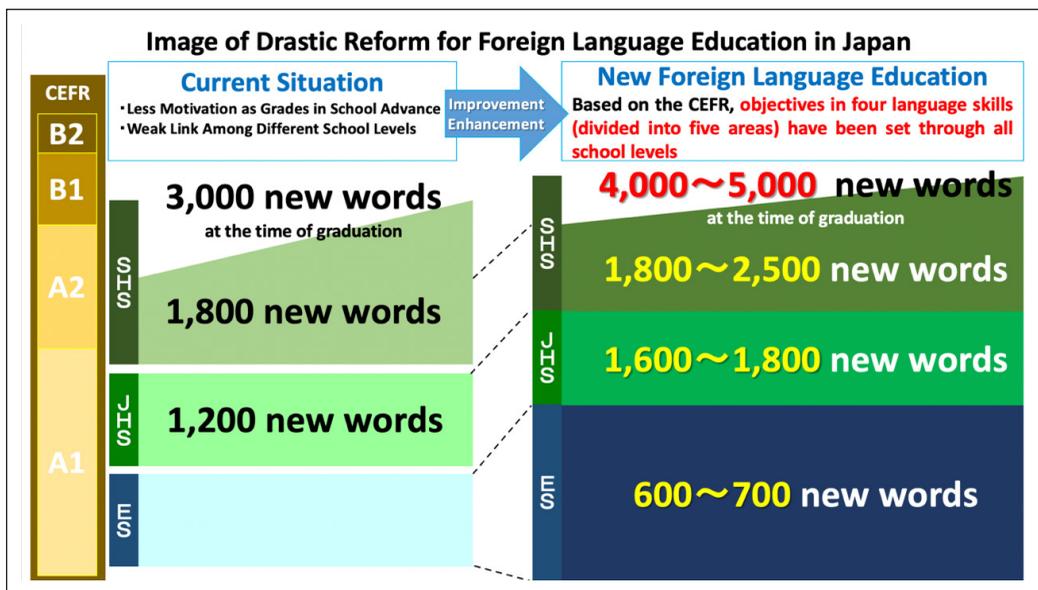
Figure 1. Improvements expected from the reformation of English education in Japan (Niigata City, 2020)

degree, have vast experience teaching English at elementary schools and a teaching licence for elementary schools in Niigata City. Their names have been omitted from this paper to maintain the teachers' privacy, and they have been allocated numeric numbers and referred to as Teacher 1 to Teacher 4. Teachers 1, 2 and 3 are MEXT teachers who work at two or three schools in Niigata City. Teacher 1 has 34 years of teaching experience in elementary schools and two years as a MEXT teacher. Teacher 2 was an elementary school teacher before becoming a MEXT teacher for the first time in April 2020. Teacher 3 has junior high school teaching experience and became a MEXT teacher in April 2019. Teacher 4 has an elementary school teaching licence and has been appointed as the teacher in charge of English at her school. The teachers have conducted speaking tests for over 400 students from April to October 2020 based on the CEFR framework.

Instrument

This research employed semi-structured interviews, and the teachers were asked five open-ended questions to elicit information on how speaking assessments were carried out before the pandemic and how COVID guidelines affected how speaking tests were carried out during the pandemic. These questions were supported by follow up questions that were aimed to gather additional responses where necessary.

The researcher had previously worked with these four teachers, so this qualitative method was the most appropriate.

Furthermore, it presented a comfortable environment and allowed the researcher to have an open conversation with the teachers as they shared their information freely. The interviews with the teachers were conducted individually at their respective schools. The data were then analysed based on emerging themes from the research objectives and expounded in this paper's findings and discussion section.

FINDINGS

This section of the paper will present the findings based on the responses given by the teachers for the questions posed to them. The findings are presented within subheadings based on the objectives of this paper.

Speaking Assessments Prior to the Implementation of CEFR

Prior to the implementation of CEFR, only two of the four teachers interviewed for this research carried out speaking assessments for their students. Teachers 1 and 3 acknowledged that although speaking was not a priority among the four language skills before the new school year (April 2020), they still tried to conduct speaking tests to gauge their students' competency level. Teacher 1 for example noted that she conducted her speaking tests in groups to help students motivate one another. She focused on collaborative work where students were asked to answer as a class or in small groups. She also used her own rubric to mark students' verbal ability. Although not aware of CEFR at that time,

she noted that a rubric for assessment guided teachers and students as they knew what they were being tested on. *"The rubric was very helpful as it helped me gauge my students' speaking ability and areas that they needed extra help with"*.

Teacher 3 also had a similar rubric when assessing speaking. He said, *"I designed a 5 point Likert scale to assess their competency level. Most of the students were between 1 and 2"* (1 being very weak and 5 competent). Teacher 3 was always more concerned with students using the language confidently rather than grading them on accuracy. When explaining the importance of understanding the context of the language, he noted that *"there is no point in them memorising the sentences for the test if they don't know what they mean"*. Therefore, his assessments before the implementation of CEFR was not based on language accuracy but rather on the ability of the student to speak in context. "*I want them to enjoy speaking English and not be afraid of the language"*. For example, he said, *"when I ask the student…how are you today? A simple answer of OK, tells me that they understood my question…and that is more important"*. Teacher 3 also asked students to design their posters or notes and present them to the class. These presentations were mainly done individually, but students had many opportunities to work in groups prior to the presentations. Teacher 3 found this helpful technique for students to speak English using the target grammar or vocabulary depending on the lesson's topic.

Speaking Assessment After the Implementation of CEFR During the Pandemic

After CEFR was introduced, all four teachers noted a guide for what to look out for in assessing their students. For example, Teacher 3 noted that *"with CEFR I knew the kind of level the students had to meet… with CEFR I am able to design lessons that will give my students enough practice in A1 level so when they are assessed, they are assessed fairly"* Teacher 2 who had not conducted speaking tests before the implementation of CEFR found it rather tricky to develop assessments that would meet CEFR standards. She did, however, acknowledge that "*CEFR presents teachers with a good framework for assessment"*. When explaining how she conducted her speaking tests, she explained that she struggled a little with developing a rubric that would test the level of all her students in her class as they have varying levels of competency. Therefore, she had to provide enough materials to help them understand the target language before assessing them.

When asked about how students reacted to speaking assessments now being a compulsory graded English language component, all four teachers noted that students were naturally more anxious during the assessment. "*In the past, my students were not particularly concerned with their pronunciation,"* said Teacher 1. However, she noted that when grading became compulsory, students were hesitant to speak, and they would stop and correct themselves. To her, this was a good mechanism, as

"self-correction is an important element of language acquisition". She also noted that regular assessments were an effective way to prepare students for tests. First, however, she explained that *"they need to familiarise themselves with the process".* This point was also expressed by the other three teachers when asked about students' reactions to the mandatory testing put in place by the ministry.

Challenges Faced and Strategies Developed

All four teachers noted that although CEFR presented them with a clear guideline to assess speaking, they felt that the COVID-19 SOPs made it very challenging for speaking tests to be carried out effectively. In adherence to the COVID-19 guidelines, there was no physical contact between teacher and student or even between themselves. The lack of physical contact in the classroom posed a challenge as it was difficult for teachers to conduct group assessments. To overcome this, Teacher 3 used Zoom to stimulate a video call environment for the speaking test. It was a new experience for his students. It piqued their interest, and *"they actually enjoyed their assessments because they were eager to see me on a computer screen…and because we conducted the assessment on Zoom, I was able to have group assessments where the students were asked to pose simple questions to their friends and they were graded based on CEFR A1 level of competency".* It was an effective mode of assessment as the teacher was adhering to COVID-19 guidelines.

Teacher 3 believes that being correct with grammar usage is not necessary. He stressed that *"assessments are of little benefit if the students merely memorise and do not understand the subject matter".* To stress this point further, he gave an example of an assessment strategy he used on one of his zoom sessions, where he asked his students to say how they all felt about being at home during the one-month lockdown. He highlighted how the *"students were happy to share their experiences and I was grading them on the side but because it seemed like a sharing session, they were freer with the use of the language and were not afraid of how they presented themselves".*

These thoughts were also shared by Teacher 2. She believes that it is important to provide students with an environment that encourages them to use the language and make mistakes. Therefore, the importance of being able to convey meaning is prioritised in her classrooms. She said, "*I follow the guidelines on CEFR but I have to adjust it to my students' level and allow them to gradually progress*". She noted that although the intended level was for the student to reach A1, she presented students with an opportunity to practise the same target language a few times. She noted that with the mask on, it was difficult for students to see her mouth movement. Therefore, she deliberately slowed down her speech and enunciated every word. This method proved effective as it helped train the student's listening along with their spoken ability. "*When I conducted speaking tests before the implementation of CEFR I would make*

students watch my lip movement and also made them place their hand in front of their lips so they could feel the difference in air pressure when certain words are pronounced. However, during the pandemic this was a little difficult to execute so I made them sharpen their listening skills".

Three out of the four teachers interviewed noted that they found it difficult to accurately link assessments to CEFR as many teachers claimed that students still required much work with their spoken skills before they could be accessed. These teachers, therefore, administered more vocal exercises and varied lessons before they began to test their students. For example, in explaining her challenge in adhering to CEFR standards, Teacher 4 said, *"I cannot test my students when they are not ready"*. She, therefore, noted that her students were given ample practice, and she even conducted mock assessments to prepare her students for the actual speaking test.

Another challenge that all teachers faced was the presence of the mask. All teachers noted that the mask posed a hindrance in identifying what students were saying accurately. For example, Teacher 2 noted that *"it is difficult to understand them under the mask as pronunciation is muffled"*. While Teacher 1 said, *"before the pandemic, it was easier to understand what the students were saying during the speaking assessment as we could see their facial expressions… during the pandemic, the masks partially covered the students' faces, and this posed a problem for us teachers"*. However, the presence of the mask indirectly compelled

teachers to focus on speaking elements of the assessment, which are more in line with the CEFR descriptors that do not include facial expressions but instead focus on the tone production of the students, such as pronunciation.

Nevertheless, to overcome the issues caused by the mask, some teachers made their students temporarily remove their face masks and wear a face shield, so their oral region was not blocked. It enabled the teachers to hear the students' responses clearly and view their non-verbal expressions. Another strategy utilised was to speak slowly to the students. It was to enable the students to understand what the teachers were saying and respond accordingly. This method was in line with CEFR's A1 level, where the student should interact with the other person provided the person speaks slowly and clearly. It was reiterated by Teacher 2: "*I had to speak extra slowly to make sure students understood me…. I also made sure students enunciated their words underneath the masks"*.

All the teachers interviewed for this paper brought up the issue of anxiety among students during speaking assessments. Two issues caused speaking anxiety. Firstly, in adhering to COVID -19 guidelines, teachers were only allowed to administer speaking tests individually (in smaller classrooms). Teachers 2 and 3 both noted that many of their students were afraid to speak alone as speaking tests in the past, although not graded, were conducted in groups where students were encouraged to converse with one another. They concurred that

students were less motivated because group assessments were no longer permitted for classrooms with small spaces. According to the teachers, this affected the grades of the students. In addition, Teacher 2 pointed out that COVID-19 guidelines limited the type of assessments that could be carried out. The distance between the teacher and student also made it uncomfortable for shy students to speak, which increased their anxiety levels. Moreover, the teacher had to make the students speak louder, which they were uncomfortable with.

To ease anxiety levels among students, Teacher 2 converted her assessment session into a role-play session. Students were placed at a safe distance and were required to ask the teacher simple questions of A1 level while pretending to be a journalist. This exercise *"broke down their anxiety a little as the focus shifted from the speaking test to questioning the teacher…and they loved it".* In addition, they were not aware that they were being assessed for the session, which made it a lot more relaxed for them. It indirectly presented a less intrusive form of assessment, a new method developed by Teacher 2 to ensure her students were comfortable during the speaking assessment and not fair badly.

In classes that allowed for more than one student for speaking assessments, Teacher 1 found that continuing pair practice at a safe distance helped students prepare better prior to the speaking tests. In addition, it allowed students to gain confidence before being graded. To put students at ease even further, she designed the speaking test in a way that was similar to how the pair practices were conducted prior to the pandemic. The familiar environment of speaking to their friends helped students develop fluency through repetition and ease their nerves. Given that every learner learns differently, Teacher 4 noted that *"some students were more confident with individual tests and were happy not speaking in front of the entire class".* The students mainly relied on notes written to help them during the speaking test and were less nervous during the assessment process.

DISCUSSION

While CEFR provides a clear framework for language teaching and assessment, responses from the teachers showed that on several occasions, the teachers resorted to their methods on how to conduct the speaking tests in line with CEFR standards. However, teachers also fell back on their primary needs and goals for their students within the English classroom. Similar to Nemoto's (2018) findings on the inconsistency of how the lessons are delivered in the classroom, the pandemic has kept the board and teachers from resolving this issue immediately. However, despite these inconsistencies, due to the introduction of the CEFR framework, teachers developed new strategies to ensure speaking tests are carried out effectively based on a globally accepted framework.

The different strategies employed by the teachers were innovative ways to ensure they continued to assess their students' speaking ability based on CEFR's A1

level while at the same time adhering to COVID-19 guidelines and SOPs. Although two out of the four teachers interviewed did not conduct assessments prior to the introduction of CEFR, they did however acknowledge that CEFR provided them with a clear framework. The role of the teacher as the interlocutor in the role-play sessions meant that the student was not placed at a disadvantage. This is because, the student is assessed on a more neutral ground as the teacher was able to adjust the conversation accordingly and this would not affect the student's performance. Ensuring that the assessments were carried out in a fun and safe manner was another strategy that worked to the advantage of the student as it helped calm their nerves before an assessment. The views of Teacher 3 on wanting to keep assessments fun and not stressful validate claims in previous research on reducing examination-oriented learning (Esther, 2012; Leong & Rethinasamy, 2020; Van Lier, 2004; William, 2011) and emphasise the need to focus on the learning experience and provide a more systematic way of assessing, recording and reporting students' learning.

With the case of the masks, while the apparatus might have hindered speech quality, it does not limit all the functions of communication entirely. The mask addresses extra-linguistic strategies that educators can use to their advantage in the classroom. It further emphasises the importance of non-verbal communication features in the language, often overlooked by students learning English, let alone speaking. It

is an area that has received significant attention, particularly in scholarly work discussing second language learners and the importance of non-verbal communication (Carreira, 2006; Richards & Schmidt, 2010; Van Lier, 2004). Therefore, despite the COVID-19 guidelines, teachers could still conduct these speaking tests even with the mask hindering the view of the organ we use to communicate. Having their students pay more attention to other aspects of the spoken language like sound production rather than lip movement was a good way to bring more awareness to the spoken aspect of English and to the different phonetic sounds that may not be evident in the student's mother tongue. In addition, the continuous repetition of such words indirectly provided more opportunities to improve language acquisition and build confidence in speaking. This method by the teachers was also in line with CEFR's A1 level, where the student develops a repertoire of words at a basic level.

Using technology to conduct assessments was a fitting example of how speaking tests can be carried out while adhering to COVID-19 guidelines. Using Zoom as a medium was a good way to test students from a safe distance. At the same time, it helped keep anxiety levels low as students felt comfortable behind the computer screen and were more comfortable speaking. Such innovative assessments are needed, particularly for elementary school students whose very thought of assessments can be quite daunting. This assessment strategy can also be employed

for future speaking assessments. It is the way forward for many institutions of learning where assessments at the initial stages of schooling could be conducted with the aid of technology before moving on with face-to-face assessments. In the years to come, it is anticipated that online learning will continue to be developed as the education technology industry is thriving during the pandemic. Creating more opportunities to connect students online for communication is ideal for making speaking tests more fruitful.

CONCLUSION

Although the data in this study is limited to 4 teachers, it does provide a basic understanding of how speaking tests were conducted prior to, and after the Ministry of Education, Culture Sports, Science and Technology (MEXT) of Japan introduced CEFR. More importantly, the findings are relevant to teachers today as the new strategies can be developed into viable means of assessments when social distancing has become the new norm.

In a global pandemic when the entire world scrambled to adhere to guidelines on social distancing and quarantine orders, Japan was one of the very few countries that decided to continue face to face mode of educational instruction. Although COVID numbers were on the rise, the government only decided to close schools for one month and, resume the face-to-face mode of instruction after that. This move by the government was particularly challenging for teachers since they were faced with two

major concerns. On the one hand, they were now required to formally grade students on their speaking ability based on the CEFR. Nevertheless, on the other hand, they had to administer these tests while adhering to strict COVID-19 SOPs as the pandemic coincided with the government's directives.

This paper has reported findings from four teachers on how speaking tests were conducted for Grade 5 and 6 elementary school students in eight schools within Niigata City. Hence it is premature to make any firm conclusions. However, the findings present important preliminary data on how speaking assessments can be carried out within the new normal even as literature in this area is still very scarce. The findings also present clear evidence that the introduction of CEFR into the elementary 5 and 6 English curricula has offered teachers a better roadmap to manoeuvre through lesson planning and assessments which is vital to monitor students' learning process continuously. It is proposed that future research could look at a larger scale involving a bigger sample of teachers from different prefectures in Japan. In addition, considering the increment in sample size, future studies could use findings from the present study and include an extra instrument of questionnaire and use a survey method to gather information from a larger sample of teachers and students that would elicit more data. With the introduction of speaking tests based on CEFR, we can see that testing provides a more solid ground on how to measure students' learning outcomes despite the use of several methods to achieve

the same goal. The findings show that this English education reform can streamline how the four language skills in English especially speaking, are taught and tested in Japanese elementary classrooms to ensure more consistency in how learning outcomes are assessed.

There are numerous other variables within the four walls of a classroom that dictate how the assessment session can unfold. For example, students' motivation for that day, the presence of face masks and social distancing that hampers auditory functions are some of the issues that can hinder the execution of a successful assessment. Therefore, based on the interviews with the teachers, while CEFR provides a clear framework for assessing the English language, teachers must always be prepared for the worst-case scenario and learn to adapt, modify and restructure assessments accordingly, and in the case of 2020, it was the global pandemic that has altered the course of education for many years to come.

REFERENCES

Behforouz, B. (2020). Common European Framework implementation and English language learning challenges. *Journal of Applied Linguistics and Language Research, 7*(4),130-146.

Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language education policy.* Multilingual matters. https://doi.org/10.21832/9781847697318

Carreira, J. M. (2006). Motivation for learning English as a foreign language in Japanese elementary schools. *JALT Journal, 28*(2), 135-158. https://doi.org/10.37546/JALTJJ28.2-2

Council of Europe (2020). *Global scale – Table 1 (CEFR 3.3): Common reference levels.* https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press.

Esther, S. H. (2012). *Student learning assessment, Asia Pacific education system review series*. http://unesdoc.unesco.org/images/0021/002178/217816E.pdf

Ghazi-Saidi, L., Criffield, A., Kracl, C. L., McKelvey, M., Obasi, S. N., & Vu, P. (2020). Moving from face-to-face to remote instruction in a higher education institution during a pandemic: Multiple case studies. *International Journal of Technology in Education and Science, 4*(4), 370-383. https://doi.org/10.46328/ijtes.v4i4.169

Gross, B., & Opalka, A. (2020). *Too many schools leave learning to chance during the pandemic*. Center on Reinventing Public Education.

Leong, H. J., & Rethinasamy, S. (2020). Narratives of English language teachers on the implementation of school-based assessment (SBA) in Sarawak, Malaysia. *Pertanika Journal of Social Sciences & Humanities, 28*(S2), 35-52.

Little, D. (2006). *The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact*. Cambridge University Press. https://doi.org/10.1017/S0261444806003557

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *91*(4), 645-655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

Mahoney, S., & Inoi, O. (2015). Homeroom teachers' perspectives on goal achievement in Japan's foreign language activity classes. *JES Journal*, *15*(1), 52-67.

Martyniuk, W., & Noijons, J. (2007). *Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States*. Council of Europe.

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In *Exploring language frameworks: Proceedings of the ALTE Kraków Conference* (pp. 135-163). Cambridge University Press.

Nemoto, A. K. (2018) Getting ready for 2020: Changes and challenges for English education in public primary schools in Japan. *The Language Teacher JALT, 42*(4), 33-35.

Niigata City. (2020). *Board of education*. http://www.city.niigata.lg.jp/shisei/soshiki/soshikiinfo/kyouikuiinkai_top/

Richards, J. C., & Schmidt, R. (Eds). (2010). *Longman dictionary of language teaching and applied linguistics* (4th ed.). Pearson Education Limited.

Tsuboya-Newell, I. (October 29, 2017). Why do Japanese have trouble learning English. *The Japan Times*. https://www.japantimes.co.jp/opinion/2017/10/29/commentary/japan-commentary/japanese-trouble-learning-english/

Van Lier, L. (2004). The ecology and semiotics of language learning: *A sociocultural perspective*. Kluwer Academics Publishers. https://doi.org/10.1007/1-4020-7912-5

WHO. (2019). *Coronavirus disease 2019 (COVID-19): Situation report, 94*. World Health Organization.

William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

Zhang, C. (2020). From face-to-face to screen-to-screen: CFL teachers' beliefs about digital teaching competence during the pandemic. *International Journal of Chinese Language Teaching, 1*(1), 35-52.

Teaching CEFR-aligned Writing to Young Learners: Practices and Voices of Teachers

Mohd Dzaquan Imran Mohd Alias[1], Abdul Halim Abdul Raof[2]* and Tina Abdullah[2]

[1]*English Language Department, Centre for Foundation Studies, International Islamic University Malaysia, Gambang, 26300 Kuantan, Pahang, Malaysia*
[2]*Language Academy, Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia*

ABSTRACT

The introduction of the Common European Framework of Reference (CEFR)-aligned English syllabus into the Malaysian education system is a welcome initiative taken by the government as a part of *"memartabatkan Bahasa Melayu, memperkukuhkan Bahasa Inggeris"* to empower Malaysian citizens to compete in this era of globalization (Ministry of Education, 2015). The new initiative has also brought forth shifts in teaching approach and technique, and assessment method. Teachers must incorporate in their teaching, among others, formative assessment and differentiation techniques. It prompted a study to be carried out to obtain insights into the practice of teachers teaching CEFR-aligned syllabus. Five primary school teachers teaching five different writing classes to nine-year-olds participated in the study. Classroom observation protocol and guided reflective interview were used to obtain data for the study. The findings revealed that the classes were mainly teacher-centered, and teachers hardly used differentiation techniques. Several formative assessment elements were exhibited but not comprehensively. Teachers' lack of training in CEFR and ineffective training system may be the cause of this observation.

E-mail addresses:
dzaquanimran@iium.edu.my (Mohd Dzaquan Imran Mohd Alias)
m-halim@utm.my (Abdul Halim Abdul Raof)
tinaabdullah@utm.my (Tina Abdullah)
* Corresponding author

INTRODUCTION

English language education for a primary school in Malaysia aims to provide students with a strong foundation in English to make them proficient, articulate, and

confident users of the language through the implementation of the Standards-Based Curriculum for Primary School or the KSSR syllabus (Ministry of Education, 2015). However, a Cambridge baseline study in 2013, two years after the implementation of KSSR, showed that slightly more than half of the students achieved A1 or A2 (Basic User level) after they completed Year 6. Another one-third did not even get as far as A1 (Cambridge, 2013). It is indicative of the outcome of the previous KBSR English Language curriculum as it did not produce exemplary results (Gill, 2013). It shows that a significant amount of effort towards reform is needed to help these students move beyond the basic user level. The Cambridge baseline study (2013) revealed that students' performance is alarming and has raised concern among educators and policymakers. As a part of educational reform that started in 2015, the adoption of the Common European Framework of Reference (CEFR) in English language education was done with the hope to produce students who possess English language proficiency that will enable them to communicate effectively in both professional and social contexts (Ministry of Education, 2015).

Perhaps the most familiar feature of the CEFR is its six reference levels or scales, ranging from A1 and A2 for basic users, B1 and B2 for independent users, and C1 and C2 for proficient users. This scale acts as an identifier for language users, a basis for an internationally recognized scale for language users, and most importantly, a map to chart language learning (Council of Europe, 2001). Apart from recording

students' progress, the CEFR is distinct from other scales; it helps frame the syllabus and curriculum and related teaching techniques to map students' progress in language learning (University of Cambridge, 2011). For this purpose, formative assessment was adopted into teaching practice. Teachers are expected to exhibit the nine elements of formative assessment in their teaching and utilize differentiation techniques for their students. So, it is with the hope to produce autonomous learners who can take charge of their learning. As argued by Fullan and Stigelbauer (2016), teachers are the agents of innovation and reformation, and as such, the success of reform depends on them (Bantwini, 2009; Wang, 2013). There is thus a need to look into its implementation in schools, particularly on teachers' teaching practices and their concerns, as they are the main determiner to the success of our education reform.

Among the transformations listed in the Wave 2 of education transformation is to enhance teacher coaching and support to improve teachers' delivery of knowledge and effectiveness (Ministry of Education Malaysia, 2012). As 2020 marked the end of this wave, the need to check how much has been achieved is of the utmost importance. There is then a need to listen to teachers to gain insights on what is happening to gauge the implementation of the CEFR-aligned syllabus in Malaysian classrooms and highlight problems found, if any. Therefore, the main aim of this study was to examine the practice of teaching English writing to young learners. The writing class was chosen as writing is

often considered a difficult skill for second language learners to master (Richards & Renandya, 2002). Therefore, shedding light into the practice in the classroom in regarding this skill can be beneficial. This study also probed into the challenges faced by teachers in implementing the syllabus in their classrooms.

Research Questions

This paper reports on a study that investigated the following research questions:

1) What is the practice of teachers teaching CEFR-aligned writing syllabus in Year 3 classes in terms of:
   a) the teaching techniques used.
   b) the differentiation techniques used.

2) What are the challenges that teachers face in teaching CEFR-aligned writing syllabus to Year 3 students?

LITERATURE REVIEW

Formative Assessment in Classrooms

Little (2013) highlighted that the CEFR-aligned syllabus for primary school emphasizes developing learner autonomy via "democratization" of second language learning. An important step to allow for the democratizing of a classroom is through the application of formative assessment. William (2018) defines formative assessment as an assessment that helps teachers and learners be informed of learning progress by looking at the evidence obtained before moving on to the next step. In other words, an assessment is formative if used to identify the learners' needs and chart the next course of action (Ministry of Education Malaysia, 2018). Grades are of least importance as what teachers plan to do with the evidence collected is much more important. The evidence collected is interpreted to identify learners' needs and the progress of their learning. To do so, William and Thompson (2007) proposed three central processes to build a comprehensive formative framework. They are:

1. establishing learners' position in their learning,
2. establishing where they need to go, and
3. establishing ways to go there.

From these processes, William (2018) proposed five key strategies to form formative assessment, namely:

1. clarifying, sharing, and understanding goals and success criteria with learners
2. creating effective classroom discussions, questions, activities, and tasks that could give evidence of students' learning
3. providing feedback that pushes learning forward
4. activating learners as owners of their learning, and
5. activating students to be resources for one another.

Adopting these principles, the Malaysian Ministry of Education (2018) formed nine building blocks for formative assessment that must be present in teachers' teaching and lesson. The building blocks are shown in Table 1.

Table 1
Formative assessment building blocks

Building block	Explanation
Sharing learning objectives and success criteria	In line with making learners aware of their current standing, teachers need to explain what their students expect to achieve and do in class.
Exemplars	Examples can be used to give students ideas on where they are going with their learning. Modeling skills show students what is expected of them.
Starters and plenaries	Starters show learners where they are going with their learning by activating students' schemata through set induction, while plenaries allow teachers to check students' progress.
Questioning	Different types of questions help teachers determine their students' understanding level and be aware of their performance.
Discussion	Teachers can assess students by making students discuss with each other. Through discussions, students also actively assess themselves and their peers by comparing responses.
Quick scans	Teachers gather the majority of responses by asking questions to the whole class or doing activities that require a response. The responses received can be used by teachers to plan the next steps for their students further.
Self-assessment and peer assessment	Involving students in assessment enables them to help each other with their learning. Students' active participation in assessing themselves is a vital piece in formative assessment.
Feedback	Feedback provides the teacher a view on students' progress through planned formative assessment activities and lessons. It recognizes what students did well and what challenges they faced and leads to the next step for teachers to take.

These formative assessment tools then served as the guide for teacher's teaching, acting as the elements that must exist in the writing process stages.

The CEFR-aligned Classroom

Differentiation techniques have been used widely in the field of language teaching. Morgan (2014) pointed out that this technique is partly based on Gardner's multiple intelligences theory (1983) and Vygotsky's Zone of Proximal Development (1978). Instead of focusing on a rigid set of plans, this technique puts students and their different needs at the center of the lesson. Tomlinson (2000) argued that all students are unique, each with different backgrounds, impacting their learning experience. In turn, it serves as the basis of differentiation techniques in the classroom. Tomlinson and Imbeau (2010) argued that the core of differentiation practice is to amend the four elements for curriculum: content, process, product, and affect. Modifying these four elements allows teachers to produce personalized lessons that will maximize students' academic growth. Maximizing the growth through personalization also requires assessment (Tomlinson & Imbeau, 2010). In the case of the CEFR syllabus, it focuses on assessment for learning that emphasizes peer and self-assessment, a shift from the traditional assessment of learning (Sidhu et al., 2018). Peer and self-assessment

are necessary components in producing autonomous language learners and teachers. The differentiation techniques also need peer and self-assessment in their lessons to achieve the CEFR-aligned curriculum.

The reality, however, is far from ideal. Cambridge's (2013) baseline study of the CEFR revealed that teachers observed focused strongly on examination, causing them to revert to the assessment of learning, with some even going as far as not focusing on listening and speaking skills as they are not the focus of examinations. In terms of differentiation techniques used in the classrooms, observers in the Cambridge baseline study found that most of the lessons were too easy for the students as the teachers observed were not sure how to differentiate tasks according to learners' level in the classroom (Cambridge, 2013).

Teaching Writing to Young Learners

Chitravelu et al. (2005) pointed out that one of the points emphasized the writing program in Malaysian schools in the various stages of the writing process. There are five general guidelines in teaching writing to Year 3 pupils, who are at the early writing stage (Chitravelu et al., 2005):

1. Showing meaningful context to teach the mechanics of writing
2. Using students' oral compositions to provide them insights into writing
3. Presenting the benefits of learning to write before them
4. Practicing reading to students various media and forms of writings
5. Developing students' thinking skills

The guideline stated above aims to develop students' autonomy towards learning. If used alongside the formative assessment building blocks, students will have a greater say in shaping their learning, an aim stated in the Ministry of Education Malaysia (2012). For example, showing students a real-life language context and sharing learning objectives and success criteria, would gauge their learning. Furthermore, it implies that the formative assessment and the general guidelines of teaching writing are in accord with each other, prompting the need to investigate how teaching writing is conducted in classes.

Benigno and de Jong (2016) described the standardization of English levels for young learners as often chaotic. It may be because the various English programs have different standards and support (McKay, 2006). The standards are often used to refer to the CEFR as the base for their learning objectives and assessment (Benigno & de Jong, 2016). The need for creating a scale for young learners arose as they argued that CEFR is adult-centric and did not consider their primary and lower-secondary education as there were almost no descriptors for below A1 level despite the majority of a communicative act done by young learners are between the levels below A1 and A1. They first turn into Pearson's (2010) PTE Academic scale used to report the Global Scale of English (GSE) progress. GSE is different from the CEFR scale in which it is more granular in which it has a wider numerical scale ranging from 10 to 90, covering CEFR levels from below A1 to low

C2. In assessing young learners, GSE scales can be used alongside the CEFR scale and not replacing it as GSE's continuous scale allows teachers to record students' progress in series of smaller gains (Benigno & de Jong, 2016). Hasselgreen (2013) attempted to expand the scale for young learners' writing through the Assessment of Young Learner Literacy program (AYLLIT) and proposed that feedback can contribute to writing assessment validity. She then further argued that feedback enables students to see their progress and standing in learning.

Studies on Teaching CEFR-aligned Syllabus in Malaysia

Past studies on the implementation of the CEFR in Malaysia include teachers' concerns on the CEFR at various levels of education, implementation, and challenges. Fatima (2019) identified the view of teachers towards *Pentaksiran Tingkatan Tiga* (PT3) (Form 3 Assessment) for the English language. The study elicited information using the semi-structured interview with five teachers involved in the study utilizing a case study approach with a qualitative research method. Teachers were generally positive and highlighted the need to tailor the CEFR-aligned language according to students' proficiency levels. The study also uncovered the challenges faced by teachers, namely the lack of teaching materials, technological resources, and students' proficiency levels.

The challenges found by Fatima (2019) was echoed in a study by Uri and Aziz (2018), citing inadequate teachers' training,

knowledge and awareness on the CEFR, their resistance and negative sentiment on incorporating the CEFR in their teaching as the challenges faced in implementing the CEFR-aligned syllabus. This study instead used a mixed method approach by interviewing two senior officials in the Malaysian Ministry of Education and employing a questionnaire to 331 English language secondary school teachers from Kuala Lumpur, Putrajaya and Selangor. Another research that examined the implementation of the CEFR-aligned syllabus and the challenges that teachers faced was conducted by Sidhu et al. (2018). A three-pronged procedure that utilized surveys, interviews and document analysis was used in this research. The major findings revealed that teachers had positive opinions on the CEFR though they lacked full understanding of assessment used for the CEFR and CEFR-aligned curriculum altogether. Document analysis in this study also highlighted the lack of teachers' encouragement to self-reflect with little to no constructive feedback in their work. The interview sessions, time constraints, class enrollment, heavy workload, and lack of training were challenges to effectively implementing the CEFR-aligned syllabus in classroom.

Another study by Alih et al. (2021) also discovered that though teachers were found to have positive opinions on the implementation of the CEFR, teachers' low motivation, insufficient materials, facilities, and time, as well as students' low proficiency level, were found to be

the main challenges that teachers faced in implementing the CEFR-aligned syllabus in schools. The mixed-approach study employed a questionnaire to gauge teachers' readiness and a semi-structured interview to elicit the challenges from purposively sampled teachers.

It is important to look at it on the ground where it all happens to gauge the success of reform and look at the preparations that lead to its implementation. The CEFR-aligned syllabus training for teachers in Malaysia was done through the Cascade Training Model, which started with a small group of trainees to progressively larger groups (Ministry of Education Malaysia, 2015). The initial group of 200 National Master Trainers trained by Cambridge English experts, was responsible for training 6000 District Trainers. These district trainers then continued to train other teachers in their district (Aziz et al., 2018). For cascade training to be successful, Hayes (2000) pointed out five criteria that need to be present:

1. The training method should be experiential and reflective, not transmissive
2. The training must allow for reinterpretation as rigid ways of doing things should not be expected
3. The knowledge and expertise shall not be concentrated at the top, and instead, it should be spread as widely as possible
4. Multiple stakeholders of different levels must be involved in the preparation of training materials

5. If possible, responsibilities within the cascade structure should be decentralized.

However, as argued by Aziz et al. (2018), the cascade training model did not meet all the criteria outlined by Hayes (2000) as it was littered with issues such as one-way communication training (lecture-style) and watered-down information being passed down.

As the Cambridge baseline report came out in 2013, naturally, there is a need for a more recent study that investigates the teaching techniques (differentiation) and assessment in a CEFR-aligned class to gauge the current syllabus's implementation. Therefore, the current work implemented a case study with a mixed approach method using previous studies as precedent to evaluate the implementation of the CEFR-aligned syllabus and the challenges teachers faced in implementing it.

METHODS

An exploratory case study with a mixed-method approach was selected to address the aim of this study. It was done to obtain information and insights on the subject matter discussed. This study involves cases in an actual context and setting, so this approach is deemed suitable (Yin, 2009). The use of the case study method is also apt as it was intended to explore the scope of study through in-depth data collection from multiple sources of information (Cresswell, 2013).

The scope of coverage of the study was teachers' teaching technique, focusing on

the nine blocks for formative assessment and the differentiation techniques that they employed in teaching writing to young learners. There were three main instruments in eliciting information required to answer the research questions: classroom observation, guided semi-structured reflective interview for collecting qualitative data, and a questionnaire to collect quantitative data.

The classroom observation protocol was adapted from Kotula and Aguilar's (2014) Writing Instruction Observation Protocol as the observation form embodies the criteria of an effective writing classroom in the Year 3 Scheme of Work from the Malaysian Ministry of Education. Moreover, the items in this checklist reflect the nine blocks of formative assessment and the differentiation technique that teachers need to integrate into their lessons as outlined in the scheme of work. It is divided into four sections: introduction stage, skills instructions and practice, composing, and miscellaneous. Data collected were in the form of the nine blocks of formative assessment teachers used and the differentiation techniques teachers used in teaching the class. After each observation session, teachers

participated in a guided reflective interview, using the items taken from the Teacher's Guide to Common European Framework of Reference, and were aligned with the Scheme of Work from the Malaysian Ministry of Education. It served to detect the challenges teachers face in writing lessons as well as the differentiation techniques used. After all, observations were completed, a questionnaire adapted from Uri and Aziz (2018) was then prescribed to teachers to elicit information on the challenges they faced in teaching writing. The reflective interview was also used to triangulate the findings obtained from the classroom observation and questionnaire.

The participants were five English language-trained teachers teaching the CEFR-aligned writing to Year 3 students. There was five Year 3 classes in the school. Thus, all teachers teaching English to Year 3 were involved in this study. The number is sufficient as Creswell (2013) argued; qualitative research aims to provide an in-depth analysis of the subject matter. Details of the participants are presented in Table 2.

While most of the teachers had similar teaching experience in terms of years, except Teacher E, they all had the same

Table 2
Background of participants of the study

Participant	Gender	Teaching experience (years)	Teaching CEFR-aligned syllabus experience (years)	Undergone training in the CEFR
Teacher A	Male	6	2	Yes
Teacher B	Female	8	2	No
Teacher C	Female	7	2	No
Teacher D	Female	8	2	No
Teacher E	Female	15	2	No

years of experience teaching the CEFR-aligned syllabus. Regarding the CEFR-related training, only Teacher A had some experience of attending one while the others had no such training. The training Teacher A had was conducted by the state education department.

Since the focus of the study was on teaching writing, observations were all on writing lessons carried out in each of the five English language classes. The topics covered in the lessons observed are shown in Table 3.

Each participant was observed three times, with each session lasting approximately 50 minutes. Though there was no discerning pattern, observation sessions were spaced not too far apart between one another. On average, it was five days; this would allow the teachers flexibility if they could not teach their class and the researcher ample time to compile notes on the previous observation session. At the end of each session, a guided reflective interview session was conducted with each teacher. Each interview session lasted approximately 30 minutes. The questions covered aspects of the nine building blocks of formative assessment and how the teachers were applying them in the class, together with the problems they may face in teaching writing in their classes. Figure 1 shows the procedure of data collection.

As shown in Figure 1, each teacher was observed and interviewed three times. The rationale for having three observations and three interview sessions is how the syllabus was arranged. Each chapter in the book contains three different writing lessons: the first lesson deals with writing at the word level before moving to sentence and paragraph level in lessons two and three, respectively.
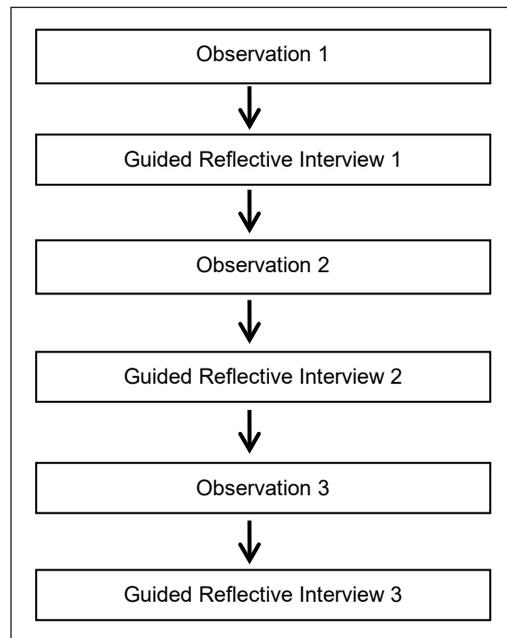
Figure 1. Data collection procedure

Table 3
Topics of writing lessons observed

Lesson	Focused item
Writing lesson 1	'There's' and 'There are some'
Writing lesson 2	'There isn't any' and 'There aren't any'
Writing lesson 3	Freewriting - Describing the toppings of pizzas or the content of salads using 'There's and There are' and 'There isn't any' and 'There aren't any'

The data obtained from all three instruments were then processed. From the observation checklist, formative assessment building blocks from each section exhibited by all teachers were grouped and compared. The data were first grouped according to the stages listed in the observation checklist. After that, teachers' actions and teaching methodology in each stage recorded on the checklist were then compared to one another and matched to the formative assessment building blocks and the differentiation techniques as listed in the Year 3 scheme of work. Concerning the reflective interview, the sessions were recorded and later transcribed. From the transcription data, two patterns emerged, namely teachers' concerns and challenges and their needs. The questions were first grouped into teachers' understanding of the CEFR and the challenges teachers faced in implementing the CEFR-aligned syllabus to analyze the data from the questionnaires. From these two groups, the mean average for the items was then processed using SPSS version 22.

RESULTS AND DISCUSSION

In general, teachers observed exhibited acceptable teaching techniques in applying formative assessment but did poorly on the differentiation techniques in their classrooms. Table 4 summarizes the teaching techniques used by the teachers concerning the formative assessment building blocks.

There were several formative assessment building blocks that all teachers exhibited in their lessons. For a start, all five teachers shared the learning objectives and success criteria by writing them on the whiteboard at the beginning of a lesson. Most, however, neglected to explain their expectations of the students in the lesson taught.

Table 4
Teaching techniques observed in relations to the formative assessment building blocks

Building blocks for formative assessment	Summary of findings
Sharing learning objectives and success criteria	All teachers shared the learning objectives. However, the importance and significance of learning objectives and success criteria were not fully explored and explained to students.
Exemplars	All teachers gave examples for students to model.
Starters and plenaries	All teachers used starters in all of their lessons observed, but only Teacher A used plenaries.
Deliberate practice	All teachers gave practice on the lesson that they taught.
Questioning	Two forms of questions were detected from all teachers: open-ended and closed questions.
Discussion	There was no discussion among students initiated by the teachers except for Teacher A.
Quick scans	Only Teacher A exhibited this.
Self and peer assessment	None of the teachers employed this in their lessons. Teacher A attempted incorporating peer assessment.
Feedback	All teachers provided feedback to students.

Teachers also observed the applied exemplars in the lessons. For instance, when Teacher B wanted the students to select the correct be-verb between 'is' and 'are,' she referred students to her slides by asking them to look at the given sentence structure and pictures before making them connect between them the examples and questions posed. On the other hand, Teacher E gave examples with correct sentence structures before giving sentences with errors in the use of be-verb for students to detect and correct.

In terms of starters and plenaries, all teachers used interesting and colorful starters, from making students count the food item shown (Teacher A), playing hangman (Teacher B), to guessing the pictures shown (Teacher D).

Teachers observed also worked closely with the textbook and workbook prescribed as all of them used the practice exercises in the books for students to work on as deliberate practice. Teacher E, in particular, tried to spice things up by making copies of worksheets from other sources (CEFR websites) and giving them to her students. However, most of the time, the teachers used books as the primary source of resources for practice.

In implementing questioning, the teachers observed had only used open-ended and closed questions. Open-ended questions given were usually simple but appropriate to the students' level like "What fruit do you like?" (Teacher B & C), "Why do you like it?" (Teacher B & C), similar to the closed questions "Do you eat fruits?" (Teacher A).

Teachers gave feedback mostly on the use of be-verb in sentences that the students filled in the blanks with or on the correctness of sentences that they wrote. Teacher E, for instance, made students come up with reasons why the answers they selected were wrong by leading them to look at the noun in the sentences before making them aware of the grammar rules by themselves.

A closer inspection of the data revealed formative assessment criteria only present in the CEFR-trained teacher's lessons. Quick scan, for instance, is a method used to help teachers quickly gauge students' understanding, and this was only practiced by Teacher A. To illustrate, Teacher A asked students to raise their hands if they knew how to pick the correct be-verb in any situation given and counted the number of hands before quickly jotting down the number in his lesson plan to be used as his teaching reflection.

The use of plenaries is next, and it was not prominent as only Teacher A used it to 're-energize' his classroom by asking students to clap and spell words chosen to gain back their focus. The other teachers just asked them to calm down.

For discussion, only teacher A tried to initiate students' discussion by making them work in pairs to identify the sentence that used the correct be-verbs. Other teachers mostly focused on the discussion between teachers and individual students. Peer assessment is the other element that was only present in Teacher A's lesson as he made sure that students would be paired in every class. Students were asked to check

each other's work and give 'stars' on their friend's work through pair work. He also encouraged them to explain wrong answers, if any, to their partner.

It was noticed that none of the teachers incorporated the element of self-assessment in their lessons. Students were not guided to assess the errors themselves, while students were made aware of their errors, either by their partners (Teacher A) or the teachers (Teacher B, C, D, E). Students were often told of their errors but never on making them aware of their progress.

As stated before, only two differentiation techniques could be observed in all teachers' lessons (Table 5). This trend applied to all teachers in all three lessons observed. Teachers started the lesson by first working with the whole class in the introduction stage and skills introduction. In later stages of the lesson (practice and composing) it is observed that teachers gave more attention to lower proficiency students, usually by catering their individual questions at their place. During the initial stages of lessons (introduction stage and skill instructions), teachers also used different types of questions depending on the students' proficiency, with close-ended questions asked to both students with higher and lower proficiency and open-ended only for students with higher proficiency.

The second set of data came from the last session of the guided reflective interview. The last session is highlighted as it was the session where the teachers had completed all three lessons, allowing their reflection to be more comprehensive. It yielded the voices of the teachers mainly on their concerns, hopes, and the challenges they faced.

An emerging pattern can be observed from data obtained from the guided reflective interviews. All teachers voiced out their lack of confidence in teaching, and the uncertainties faced. The other concern that all teachers had stated is students' proficiency level about the difficulty of the syllabus. One of the participants aptly stated, *"...I think the new syllabus is too hard for some students"* (Teacher A). Teacher B also shared the same concern *"...They don't respond much since it's hard. And I'm afraid that I teach them wrongly because it is hard".*

Teachers' lack of confidence may be attributed to the lack of training related to the CEFR-aligned syllabus implementation. As stated by Teacher D, for instance, *"I have been teaching for quite some time, but this new syllabus makes me lack confidence."* Similarly, Teacher C shared the same sentiment: *"So many times I would think that my teaching method is not*

Table 5
Differentiation techniques observed

Differentiation technique	Explanation
Type and amount of support	Teachers mostly worked with the whole class and catered to weaker students individually.
Types of questions asked	Teachers used open-ended questions for the more proficient students and close-ended for the weaker ones.

correct….." Despite having undergone the CEFR training, Teacher A had this to say on incorporating the CEFR in his lesson: "*I am not sure whether I'm teaching [it] correctly.*"

Based on the teaching and differentiation techniques observed, the best statement to summarize the teachers' teaching technique is that the lessons conducted were heavily teacher-centered. Teachers gave feedback on their students' writing, and teachers did not attempt to allow students to help each other. The concerns and challenges in Table 6 could be due to the teachers who were not confident with what they were doing in the classroom about the CEFR-aligned syllabus. Consequently, the discussion among students did not happen, nor was it encouraged, as teachers had complete control of the lessons. The lessons were conducted in one-way communication, with students contributing minimally towards their learning. Developing students' autonomy, which is one of the goals for adapting the CEFR-aligned syllabus, was not seen in the lessons observed. Students were passive receivers; they were not guided to evaluate themselves to become active participants in their learning. Teachers did not emphasize the importance of success criteria and quick scans.

Another finding describing the teachers' teaching technique is that they lacked variety in differentiation techniques used. Differentiation techniques are what teachers need to infuse in their teaching to provide a personalized learning experience as different students have different needs when it comes to learning. In this study, the teachers observed utilized only two kinds of

Table 6

Concerns and challenges, and needs of teachers

Teacher	Concerns and challenges	Needs
A	□ Books used in Year 1 & 2, and Year 3 not from the same series □ The new syllabus □ Students' low proficiency level □ Teacher's lack of confidence in teaching □ Teaching writing is difficult	□ Books for students should be taken from the same series.
B	□ The difficulty of the syllabus □ Students' low proficiency level □ Teacher's lack of confidence in teaching	□ Able to teach appropriately and for students to be able to respond
C	□ Grammar items covered in the book □ Teacher is not familiar with the CEFR □ Students' understanding □ Lack of confidence in own proficiency	□ Students can understand the lessons
D	□ The new syllabus □ Teacher's lack of confidence in teaching	□ To match the teaching techniques to the new syllabus
E	□ The assessment for Year 6 students according to the new syllabus □ Not sure how to prepare students for examinations	□ To get information on public examination for Year 6

differentiation techniques: questions posed and the amount of support given to students. While this can be helpful, especially to weaker students, differentiating the task and outcome expected from students, for instance, can help teachers to control their classroom better and help to enrich the more advanced students. Teacher A, for instance, stated, *"The good students will often finish their work quickly. So, for them, I challenge them to write more sentences or use two grammar items in a sentence. Or else they will go around the class and play."* Each student has unique experiences and background knowledge, and it is only suitable that their learning experience reflects that (Tomlinson, 2000). As Morgan (2014) pointed out, it is significantly more important nowadays as the world is getting even more diverse with students having varying needs.

Table 6 summarizes the findings on concerns and challenges and the needs of the teachers involved in the study.

In discussing the findings in the light of current literature on the CEFR implementation in Malaysia, we consider the voices and practices of teachers gathered within a similar setting. Data obtained from the guided reflective interview of this study suggest that teachers found the new syllabus too difficult both for them and their students. This finding concurs with Alih et al. (2021), in which the study pointed out the mismatch between the syllabus and students' proficiency levels. In addition, teachers were also found to be lacking in confidence to teach the syllabus. Teachers' lack of confidence is echoed in the study

by Sidhu et al. (2018), in which they found that teachers cited insufficient training as one of the major obstacles in implementing the CEFR-aligned syllabus. Similarly, the study by Uri and Aziz (2018) found teachers lacking in training and understanding of the CEFR in general, inhibiting their ability to incorporate the CEFR in their teaching successfully.

Resonating similar findings was the study by Kok and Aziz (2019), which revealed teachers needed guidance in formulating lessons, citing a lack of training and resources such as ICT and teaching and learning supplementary materials. They further found that teachers had issues with using the recommended textbooks, which were not local products. It was also a concern raised by the teachers in Mohammed's (2020) and Alih et al. (2021) studies which highlighted that the contents of the textbooks are international, making it more difficult for less proficient students to comprehend. In the rural areas, the situation is worse, as Nawai and Said (2020) pointed out that not only teachers were found to be lacking in confidence in teaching the new syllabus, they were also reluctant to implement the CEFR in the classroom, which possibly stemmed from inadequate training and unavailability of suitable resources, including textbooks.

The issue of the teacher-centeredness approach is another main finding of this study. Coupled with the lack of differentiation techniques, it mirrors the findings of the Cambridge baseline study (2013), where teachers' presence was found to be too dominant with almost no

practice towards using the language in the real context. Moreover, just like in the baseline study, teachers were unsure how to use differentiation techniques in class (Cambridge, 2013). It is alarming as this may suggest that despite many years having passed since the initial baseline study, only minimal changes could be observed.

CONCLUSION

Teachers' lack of knowledge on the CEFR and, ultimately, the proper teaching techniques for the CEFR-aligned syllabus was found to be the main hurdle of its implementation. The findings of this study show that there is more to be done before the impact of the implementation of the CEFR in the education system can be seen. Though this was a small-scale study, from the findings, we know a little bit more about the practices of teachers, their concerns, and the challenges they faced in trying to facilitate the reform in teaching and learning.

This study has shed some light for education stakeholders in Malaysia to investigate the issues surrounding the implementation of the CEFR-aligned syllabus about the lack of training provided. Cascade training model was chosen for the CEFR-aligned syllabus in Malaysia (Ministry of Education Malaysia, 2015). A selected small group of teachers was trained, and they later trained other teachers who themselves became trainers to more teachers in their school or district. However, the cascade training model, despite being considered to be the best choice when it comes to implementing training for teachers

in large numbers (Karalis, 2016), has been criticized for failing to deliver effective training (Bett, 2016; Dichaba & Mokhele, 2012; Robinson, 2002).

The cascade training model used to train teachers also needs to be revisited to check for its effectiveness. For example, Aziz et al. (2018) argued that it was hard for trainers to share the information with other teachers as the trainers themselves were also teachers, with their main business being teaching. At the same time, it is understandable that the cascade training model was chosen to alleviate budget constraints, the fact that even specialist teachers feel insufficient. However, the superficial nature of the training suggests that an online portal should be made available to allow teachers from all over Malaysia to support each other. In addition, more funding and materials should be given to help teachers resolve problems by themselves.

Other initiatives to assist teachers include a mentor-mentee system that pairs up teachers can be created provided that the mentor has received enough support in the form of materials and training. This system can also share, give, and receive feedback on lesson planning at the school level.

In addition, support to teachers in terms of specific training such as on how best to incorporate innovative, learner-centered techniques in their lessons should be given. Likewise, exposure to students on this new change is to be provided to be more receptive and eventually benefit from it. After all, it is their performance that would reflect the success of the reform.

In preparing our students to be more able to compete in this globalized world, implementing the CEFR-aligned syllabus is considered a step in the right direction. However, by focusing on communication competence, our students must be developed in this aspect. In order to envisage this, the implementation of the CEFR in Malaysia needs improvement. All stakeholders involved must be alert on this and are up-to-date with the latest information to implement the CEFR, a promising reform towards English education in Malaysia.

ACKNOWLEDGEMENT

REFERENCES

Alih, N. A. C., Raof, A. H. A., & Yusof, M. A. M. (2021). Policy change implementation: The case of the CEFR in Malaysian ESL classrooms. *Journal of Nusantara Studies, 6*(2), 296-317. doi.org/10.24200/jonus.vol6iss2pp296-317

Aziz, A. H. A., Rashid, R. A., & Zainudin, W. Z. W. (2018). The enactment of the Malaysian Common European Framework of Reference (CEFR): National master trainer's reflection. *Indonesia Journal of Applied Linguistics, 8*(2), 409-417. https://doi.org/10.17509/ijal.v8i2.13307

Bantwini, D. B. (2009). District professional development models as a way to introduce primary school teachers to natural science curriculum reforms in one district in South Africa. *Journal of Education for Teaching, 35*(2), 169-182. https://doi.org/10.1080/02607470902771094

Benigno, V., & de Jong, J. H. (2016). The global scale of English learning objectives for young learners: Basic inventory of descriptors. *Assessing Young Learners of English: Global and Local Perspectives, 25*, 43-64. https://doi.org/10.1007/978-3-319-22422-0_3

Bett, H. K. (2016). The cascade model of teachers' continuing professional development in Kenya: A time for change? *Cogent Education, 3*(1), 1-9. https://doi.org/10.1080/2331186X.2016.1139439

Cambridge. (2013). *Cambridge baseline 2013: English language in Malaysian schools.* Cambridge English Language Assessment.

Chitravelu, N., Sithamparam, S., & Teh, S. C. (2005). *ELT methodology: Principles and practice.* Oxford University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press.

Cresswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications, Inc.

Dichaba, M. M., & Mokhele, M. L. (2012). Does the cascade model work for teacher training? Analysis of teachers' experiences. *International Journal of Educational Sciences, 4*, 249-254. https://doi.org/10.1080/09751122.2012.11890049

Fatima, S. (2019). Perceived view of teachers towards Pentaksiran Tingkatan Tiga (PT3) [Form Three Assessment] English language: A case study. *Asian Journal of University Education, 15*(3), 34-44. https://doi.org/10.24191/ajue.v15i3.7819

Fullan, M., & Stiegelbauer, S. M. (2016). *The new meaning of educational change* (5th ed.). Teachers College Press.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Gill, R. S. (2013). *Enhancing English teacher language proficiency: The Malaysian experience*. British Council.

Hasselgreen, A. (2013). Adapting the CEFR for the classroom assessment of young learners' writing. *The Canadian Modern Language Review, 69*(4), 415-435. https://doi.org/10.3138/cmlr.1705.415

Hayes, D. (2000). Cascade training and teachers' professional development. *ELT Journal, 54*, 135-145. https://doi.org/10.1093/elt/54.2.135

Karalis, T. (2016). Cascade approach to training: Theoretical issues and practical applications in non - formal education. *Journal of Education and Social Policy, 3*, 104-108.

Kok, N. M., & Aziz, A. A. (2019, March 27). English language teachers' perceptions on the implementation of CEFR-aligned curriculum among primary schools in Malaysia. In *Seminar Wacana Pendidikan 2019 (SWAPEN 2.0)* (pp. 212-222). Kedah, Malaysia.

Kotula, A. W., & Aguilar, C. M. (2014). *Developing a writing instruction observation protocol: Implications for ractice, research, and policy*. Education Development Center Inc.

Little, D. (2013, October, 29-30). *The Common European Framework of References for Languages: Purpose, origin, ethos and implications* [Paper presented]. CEFR Conference. Putrajaya, Malaysia.

McKay, P. (2006). *Assessing young language learners*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733093

Ministry of Education Malaysia. (2012). *Preliminary report - Malaysia education blueprint (2013-2025)*. Ministry of Education Malaysia.

Ministry of Education Malaysia. (2015). *English language education reform in Malaysia: The Roadmap 2015-2025*. Ministry of Education Malaysia.

Ministry of Education Malaysia. (2018). *Teacher handbook: Formative assessment principles and practices*. Cambridge Assessment English.

Mohammed, N. A. (2020). Teachers' views and challenges in using CEFR to teach

English language in primary school [Unpublished Master's thesis]. Universiti Teknologi Malaysia.

Morgan, H. (2014). Maximizing student success with differentiated learning. *The Clearing House, 87*, 34-38. https://doi.org/10.1080/00098655.2013.832130

Nawai, R., & Said, N. E. M. (2020). Implementation challenges of Common European Framework Reference (CEFR) in a Malaysian Setting: Insights on English Teachers' Attitude. *International Journal of Academic Research in Business and Social Sciences, 10*(7), 28-41. https://doi.org/10.6007/IJARBSS/v10-i7/7394

Pearson. (2010). *Aligning PTE academic test scores to Common European Framework of Reference for languages*. http://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf

Richards, J. C., & Renandya, W. A. (Eds.). (2002). *Methodology in language teaching: An anthology of current practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667190

Robinson, M. (2002). Teacher reforms in South Africa: Challenges, strategies and debates. *Prospects, 11*, 289-299.

Sidhu, G. K., Kaur, S., & Chi, L. J. (2018). CEFR-aligned school-based assessment in the Malaysian primary ESL classroom. *Indonesian Journal of Applied Linguistics, 8*(2), 452-463. http://doi.org/10.17509/ijal.v8i2.13311

Tomlinson, C. A. (2000). *The differentiated classroom: Responding to the needs of all learners*. ASCD.

Tomlinson, C. A., & Imbeau, M. B. (2010). *Leading and managing a differentiated classroom.* ASCD.

University of Cambridge. (2011). *Using CEFR: Principles of good practices at University of Cambridge.* Cambridge ESOL Examinations.

Uri, N. F. M., & Aziz, M. S. A. (2018). Implementation of CEFR in Malaysia: Teachers' awareness and the challenges. *3L: The Southeast Asian Journal of English Language Studies, 24*(3), 168-183. http://doi.org/10.17576/3L-2018-2403-13

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Wang, W. (2013). Teachers' stages of concern and levels of use of a curriculum innovation in China: A case study. *International Journal of English Language Teaching, 1*(1), 22-31. https://doi.org/10.5430/ijelt.v1n1p22

William, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551-57. https://doi.org/10.1080/0969594X.2018.1441807

William, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C.A. Dwyer (Ed.), *The Future of Assessment: Shaping Teaching and Learning* (pp.53-82). Routledge & CRC Press. https://doi.org/10.4324/9781315086545-3

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Sage.

SOCIAL SCIENCES & HUMANITIES

Determining English Language Lecturers' Quality of Marking in Continuous Assessment through Rasch Analysis

Mardiana Idris

Education Department, Institut Pendidikan Guru Kampus Temenggong Ibrahim, Jalan Datin Halimah, 80350, Johor Bahru, Johor, Malaysia

ABSTRACT

English language lecturers at matriculation colleges are generally equipped with assessment criteria for marking students' written assessment. However, these criteria are normally susceptible to lecturers' interpretation and understanding, which threatens quality marking. Therefore, this study aimed to determine the severity and consistency of English language lecturers' marking of English academic writing (EAW) in continuous assessment. The participants were five English language lecturers and 50 matriculation students. Each lecturer selected ten EAWs randomly from 318 matriculation students. The five-part EAW was marked first by the class's lecturer and later, it was marked by pre-assigned peer moderators who did not teach the students. The total data set collected was 250 (5 lecturers x 10 EAWs x 5 parts of EAW). The data were analyzed with Many-Facets Rasch Measurement (MFRM) application. Semi-structured interviews were conducted with both lecturers and students for triangulation purposes. Findings revealed that four out of five lecturers were lenient in marking but the marking was found to be internally consistent with infit and outfit mean squares for each lecturer ranged between 0.5 and 1.5. From interview responses analyzed, students perceived their lecturers as fair but strict in awarding marks. These responses were consistent with most lecturers' responses on their strict adherence to assessment criteria. Discussion of findings is centered on the issue of severity and consistency of the assessors. This study could offer a practical solution in providing evidence for quality marking of written assessment and, consequently, aid in developing remedial measures for misfit assessors in educational institutions.

Keywords: Consistency, continuous assessment, Rasch analysis, severity, written assessment

INTRODUCTION

Generally, continuous assessment is perceived as a measurement mechanism to gauge the learners' learning progress and gain based on specified and fixed criteria, which normally translate learners' achievement into numerical digits (Carrillo-de-la-Pena & Perez, 2012; Mikre, 2010; Walde, 2016). These digits are then converted into grades, bands, categories, or levels that portray learners' ability to master skills, topics, or subjects. However, how accurate is this portrayal, particularly when it involves subjective marking whereby the assessors solely awarded marks? Despite each assessor's every intention to remain objective, to compound the conundrum further, their marking may be 'affected by classroom relationships and interactions' (Tierney, 2016) in the teaching and learning environment. It leads to the issue of ensuring quality in marking. Quality marking is essential, particularly in continuous assessment, because it affects students' learning. Tierney (2016) and Jiminez (2015) reported that learners generally exhibited their actual performance in learning if they perceived the teachers or lecturers as being fair in assessing their assessments. Therefore, this study attempted to determine lecturers' severity and consistency of marking matriculation English academic writing (EAW) in a continuous assessment.

In this paper, the objective and research questions are first outlined. Then, theoretical underpinnings of assessment and studies related to severity and consistency in marking are discussed in the literature review. Subsequently, the methodology used is elaborated, and this is followed by describing the findings based on the research questions. Finally, discussion, implications, and conclusions are presented.

OBJECTIVE OF THE STUDY

The study's primary objective was to determine English language lecturers' severity and consistency in marking matriculation students' five-part English academic writing (EAW) paper.

RESEARCH QUESTIONS

Three research questions were formulated to guide the study to achieve the primary objective

1. to what extent were the assessors severe in marking matriculation students' EAW in continuous assessment?
2. to what extent were the assessors consistent in marking matriculation students' EAW in continuous assessment?
3. how did lecturers and students perceive the severity and consistency of EAW marking in continuous assessment?

LITERATURE REVIEW

Severity and leniency in marking written assessments have always been dilemmas faced by many lecturers or assessors. Questions that linger include "Did I mark according to the rubric provided?", "Did I award an 'accurate' score that reflects the student's performance?" and "Did my

assumptions of the students' knowledge or behavior cloud my fair judgement?" These lingering quality control indicators may have resulted in learners questioning the scores or marks they have received, particularly if they perceived that they had been assessed severely or unfairly by their assessors. Assessor or rater severity consistently provides scores or 'ratings that are lower or higher than is warranted' (Engelhard, 1994) by learners' performances. In fact, there are many studies on severity of assessors (Han & Huang, 2017; He, 2019; McNamara et al., 2019; Park, 2011) in assessing written task and its impact on quality assessment. Levey (2020) observed that any performance assessment typically judged by human raters will introduce subjectivity. Consequently, this could lead to unreliable scoring. Studies by Fahim and Bijani (2011) as well as Erguvan and Dunyait (2020) reported that assessors' severity and leniency in marking could cause dissatisfaction among test takers, and both studies recommended for rater training to be given to assessors in order to reduce rater variability. Most studies reported that rater training did reduce rater variability but did not eliminate it.

Another imperative criterion for quality marking is consistency, which is often linked to reliability. This study obtained assessors' consistency by providing training for assessors and using multiple assessors (Lang & Wilkerson, 2008; Willey & Gardner, 2010). Many studies have reported the importance of training the assessors before marking to achieve a higher inter-rater or consistency value (Erguvan & Dunyait, 2020; Kayapinar, 2014; Park,

2011; Sundqvist et al., 2020). For example, Hack (2019), in her doctoral thesis on marking processes used in the assessment of extended written responses, quoted a study by Morin et al. (2018) which reported that 'the probability that candidates receive the correct grade (the 'definitive' grade awarded by the team of senior examiners) on a combined English literature and language qualification was only 52%.' (p. 10). Thus, this indicates that the reliability of marking written assessment invites contention if not conducted properly.

Emphasis on the severity and consistency of marking is due to its feedback role in the formative assessment framework. Black and William (2009) conceptualized five key strategies in the assessment process. The first strategy was to clarify and share learning intentions and criteria. The second strategy involved engineering learning tasks that elicit evidence of student learning. Finally, the third strategy pertained to providing feedback that moves learners forward. It was achieved through written feedback given by fair and consistent markers, which guided their subsequent performance. The fourth strategy concerned activating learners as instructional resources, while the fifth focused on activating learner, as the owners of their learning. The framework for assessment strategies is illustrated in Table 1.

The conceptualized framework by Black and William (2009) in Table 1 shows that assessment contributes to quality learning. A direct consequence for learners' improvement in writing skills is through column 3, "Providing feedback that moves

Table 1
Assessment strategies framework suggested by Black and William (2009)

	Where the learner is going	Where the learner is right now	How to get there
Teacher	1. Clarifying learning intentions and criteria for success	2. Engineering effective classroom discussions and other learning tasks that elicit evidence of student learning	3. Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	4. Activating learners as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	5. Activating learners as the owners of their learning	

learner forward." Hypothetically, suppose students received unfair and inconsistent marks or scores as feedback for their written assessment. In that case, it could indirectly affect their learning because feedback or scores given does not truly reflect their ability. As such, learners 'may be moved' in the wrong direction in improving their writing skills.

The severity and consistency of assessors could always be gauged through classical test theory, whereby average scores and reliability of assessors are analyzed. However, this theory alone is not enough to describe the linear relationship between students, items, and subjective marking of assessors. Hence, Many-Facets Rasch Measurement (MFRM) was used in this study. MFRM is a psychometric analysis that can identify assessors' severity and consistency in marking subjective assessment (Prieto & Nieto, 2014; Eckes, 2005). Meadows and Billington (2005) outlined the advantages of MFRM, which include:

> "Using a many-facets analysis, each question paper item or behavior that

was rated can be directly compared. In addition, the difficulty of each item, as well as the severity of all judges who rated the items, can also be directly compared. Person abilities can be evaluated whilst controlling for differences in item difficulty and judge severity." (Meadows & Billington, 2005; p. 6)

Based on these advantages, the MFRM has been used in many large-scale assessments and certifications, including developing the Common European Framework of Reference (CEFR) (Council of Europe, 2009).

METHODOLOGY

This methodology section describes the participants involved in the study and the instruments used to collect the data. The nine phases of the procedures are also described.

Participants

The lecturers (labeled as assessors henceforth) were five English language

lecturers who taught matriculation English 1, English 2 and Malaysian University English Test (MUET) to matriculation students. The assessors had ten to fourteen years of teaching experience. Four out of five assessors had experience in marking the MUET Writing paper. In addition, all assessors were well versed with the rubrics and scoring guide provided by the Matriculation Division as they had been given training prior to marking the assessment. Based on the appointment letters by the college, each lecturer was appointed as an assessor for their own students' assessment and a moderator for their peers. One of the lecturers was appointed as a chief moderator.

As for students, they were 50 engineering matriculation students. On average, they were 18 years old. Most students were categorized as having intermediate to advanced levels of English language proficiency based on their *Sijil Pelajaran Malaysia* (SPM) English results.

Instruments

Two types of instruments were used in this study–students' EAW and a semi-structured interview. Fifty EAWs were randomly selected from 318 matriculation students. The 50 scripts were selected due to the procedures outlined by the Matriculation Division, whereby English language lecturers must moderate ten EAW scripts from their classes. For the EAW, the students were required to write a personal statement to a university for placement purposes. Students had to write their statements in

five parts. Part 1 was an introduction to the personal statement. Part 2 was a content paragraph in which students were required to describe their past experiences using the past tense. Part 3 was another content paragraph that required students to describe their current undertakings, while Part 4 was the last content paragraph which required students to write in the future tense. Finally, Part 5 was the conclusion to the personal statement. For a complete sample of the paper, please refer to Appendix A.

In terms of scoring criteria, Part 1 and 5 used five scoring levels, with Level 1 (Limited user) as the lowest and Level 5 (Excellent user) as the highest. Generally, Parts 1 and 5 employed holistic assessment criteria (Appendix B). As for the content paragraph, it also used five scoring levels. The levels were: Level 1 (very weak), Level 2 (weak), Level 3 (Fair), Level 4 (Good), and Level 5 (Very Good). However, Parts 2, 3, and 4 used an analytic assessment criterion that focused on three components: focus, organization, and language (Appendix C)

Semi-structured interviews with lecturers and students were also conducted to corroborate the quantitative findings.

Procedures

The study was conducted in nine phases. Phase 1 focused on training the assessors and the moderators to mark the EAWs. Chief Moderator gave the training, and during training, assessors were encouraged to ask questions to have the same understanding of the criteria. After all, assessors were clear with the rubrics and scoring guide,

and they conducted the same briefing to their students prior to assessment. Next, students attempted EAW in Phase 2. Every assessor marked their scripts for two weeks in Phase 3. Then, in Phase 4, scripts were moderated by peer moderators. For Phase 5, none of the scripts had to be moderated by the Chief Moderator since the difference in raw scores was not more than five marks. Phases 6, 7, and 8 involved MFRM analysis, interviews, and transcription. Finally,

Phase 9 concentrated on the findings. The summary of all nine phases involved is presented in Table 2.

FINDINGS

Descriptive statistics, Rasch variable map (Wright map), assessor measurement report, and interview responses are used to report the findings based on the research questions initially presented.

Table 2
Summary of nine phases of the study

Phase	Description	Analyses involved
Phase 1	• Lecturers were appointed as assessors for the continuous assessment. Assessors received training on scoring guides and criteria from the Chief Moderator. Assessors asked questions to the Chief Moderator when doubts arose. • All matriculation students were given the scoring guide and criteria. Lecturers explained the scoring guide and criteria to the students.	Not applicable
Phase 2	• 318 students attempted all five parts of the EAW.	Not applicable
Phase 3	• Each assessor randomly selected 10 EAW to be marked using the scoring guide and criteria. Assessors were given two weeks for marking.	Raw scores
Phase 4	• Each assessor submitted their ten (10) marked EAW scripts to their peer moderator. Moderators were given a week to mark. Rating/judging designs for both assessors and moderators were preplanned to ensure a smooth analysis in the MFRM software (Facets)	Raw scores
Phase 5	• Moderators returned the marked scripts to the first assessors. Since the difference of marks was not more than five in each EAW, the scripts were not submitted to the Chief Moderator.	Raw scores
Phase 6	• The researcher analyzed the data in Facets software: 3 facets rating scale—assessors, students' EAW, and items with rating 1 to 5.	Facets analysis
Phase 7	• A semi-structured interview was conducted with lecturers. • A semi-structured interview was conducted with students.	Not applicable
Phase 8	• Transcription of interview	Thematic analysis
Phase 9	• Analysis of findings	• Descriptive statistics • Rasch variable map (Wright Map) • Assessor measurement report

Descriptive Statistics

Table 3 shows mean ratings by lecturers for parts 1, 2, 3, 4, and 5 of EAW. Based on Table 3, it shows that Lecturer 5 seemed to be severe with the rating awarded as the mean for each part was categorized as a competent user (3) and fair (3) while the rest of the lecturers were awarded good standing (4) for most parts of the EAW. At first glance, it could indicate that Lecturer 5 was severe in marking, but this did not entirely explain the severity of the assessor since it was based on means. Therefore, MFRM analysis was used.

Severity of Assessors in Marking EAW

Figure 1 illustrates a graphical description of three facets analyzed in the MFRM – student ability, part (or item) difficulty, and assessor severity- along a logit scale of a Rasch ruler. Logit is the unit used in reporting the MFRM analyses. The first column is a measure column (Measr) which ranges between -2 logits and +8 logits, with 0 as the mean. The second column (Students) displays students' ability based on the ratings awarded. Higher ability students are closer to the top, while less able students are closer to the bottom. The third column displays the five parts of the EAW. The parts are ordered according to the level of severity imposed by assessors. The harsher a part is assessed,
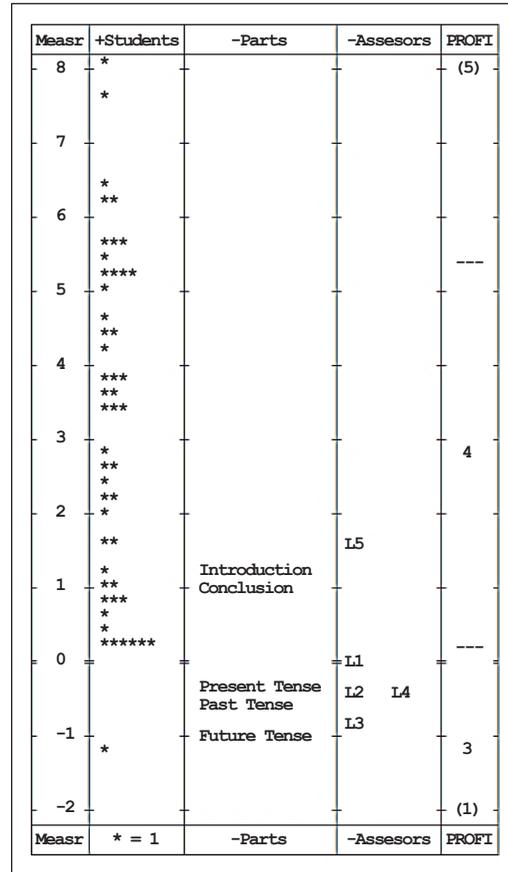
Figure 1. The Wright map for students' ability, level of EAW difficulty, and assessors severity

Table 3
Mean rating by lecturers for parts 1, 2, 3, 4, and 5 of EAW

Lecturer/ Part	Lecturer 1	Lecturer 2	Lecturer 3	Lecturer 4	Lecturer 5	Mean
1 (Introduction)	3.5	3.9	4.2	4.3	2.9	3.7
2 (Past Tense)	4.1	4.4	4.6	4.3	3.6	4.2
3 (Present Tense)	4.1	4.3	4.5	4.4	3.6	4.2
4 (Future Tense)	4.0	4.6	4.8	4.4	3.6	4.3
5 (Conclusion)	3.6	3.7	4.0	4.3	3.5	3.8
Mean	3.9	4.2	4.4	4.3	3.4	4.04

the higher is the position of the part on the map. Conversely, the lower the position of a part on the map, the less harsh the part is assessed. The fourth column displays five assessors coded as L1 to L5. Severe assessors are located closer to the top, while lenient assessors are located closer to the bottom. The fifth column displays the rating scale used (1– 5).

Based on Figure 1, the student ability ruler indicates that the students scored highly on the EAW as 49 out of 50 students were above mean 0 while only one student was rated below mean 0. In addition, student ability was clustered within scale 4 (good) as indicated from 0 logits to +8 logits. This distribution pattern implied that students could be highly proficient despite being randomly selected by the lecturers.

Next to the student ability column is the part ruler. The parts are ordered with an introduction as the harshest part rated by assessors while future tense as least harshly rated. There seems to be a clear pattern distinction as the introduction and conclusion (holistic criteria) are closer together. In contrast, present, past, and future tense (analytic criteria) are clustered together. Despite this distinction, the parts do not differ much within -1 logit and 1 logit. It suggests that both analytic and holistic criteria received approximately similar attention from the assessors since they are clustered together. However, holistic criteria (Introduction and Conclusion) seem to receive more attention than analytic criteria since they significantly differ from the rest.

Besides parts, assessors are also modeled with the most severe ones at the top and the most lenient ones at the bottom of the Rasch ruler. The ruler shows that L5 is the most severe assessor while L3 is the most lenient. The map also indicates more lenient raters than severe ones as four assessors are positioned below mean 0.

The final ruler displays the five rating scales. The range of the rating ruler for all five categories starts from 1 until 5. Although the rating scale has five levels, levels 1 and 2 are absent from the ruler. It implies that these levels were not awarded to students.

Consistency of Assessors in Marking EAW

The Wright map described earlier was only a brief representation of all the facets investigated for quality control. Therefore, to address the second research question, an assessor measurement report is needed.

Table 4 shows the assessor measurement report, ordered from the most severe to the most lenient raters. Infit and outfit mean-squares for four raters were between 0.5 and 1.5 logits, and these values were the recommended range for productive measurement. Although the infit and outfit mean-squares of L2 (infit: 1.58 and outfit 1.55) slightly exceeded the recommended range, these values, however, did not distort the measurement as they did not exceed 2.0. According to Linacre (2014), separation of more than two and reliability of more than 0.8 were indications of data that fit the measurement model. The values of separation and reliability statistics provided at the bottom of Table 4 indicated that the data fitted the model since the

Table 4
Measurement report on lecturers' severity in marking

Lecturers	Severity Measure	Model S.E	Infit MnSq	Outfit MnSq
L5	1.66	0.18	0.41	0.39
L1	-0.01	0.20	1.14	1.24
L4	-0.37	0.23	1.13	1.18
L2	-0.44	0.22	1.58	1.55
L3	-0.84	0.24	0.88	0.85

Note. S.D: 0.95; Separation: 4.44; Reliability (not inter-rater): 0.95

separation was 4.44, and thus, reliability was high with 0.95. In addition, the standard deviation (S.D) given at the bottom of the Table 4 indicated that the data were clustered towards the mean with less than one standard deviation. It suggests that assessors had a similar rating tendency. As for assessor severity, this was gauged from logit measures reported in the second column of the table. The range of severity measure from the most severe assessor (L5: 1.66 logits) to the most lenient (L3: -0.84 logits) was about 2.5 logits. Table 4 shows that four out of five assessors were lenient in awarding their ratings for the written assessment.

From the severity measures provided, it was found that most assessors tended to rate the essays leniently. However, the severity measures of L1, L2, L3, and L4 did not differ much, and most importantly, they did not exceed -1 logits. In fact, since the severity measures clustered between -0.01 logits and -0.84 logits, it might indicate that they had a similar understanding of the assessment criteria. However, the L5 severity measure exceeded 1 logit (1.66). Therefore, it may indicate a departure from applying the assessment criteria objectively.

Internal consistency was measured through assessors' infit mean-squares. Infit mean-square is less sensitive to outliers, but they are more sensitive towards unexpected ratings (Yan, 2014). Hence, infit mean-square is the benchmark for assessors' internal consistency in awarding scores. Based on Table 4, L5 displays infit mean-squares lower than 0.5 (0.41 logits), which indicated that the value was influenced by rating patterns and thus, posed a greater threat to measurement (Linacre, 2014). Although the L2 infit mean-square was 1.58 logits, this value did not distort the measurement as it did not exceed 2.0 logits. The infit mean-squares of three assessors were between 0.88 logits and 1.14 logits. These values indicated that most assessors were largely internally consistent in marking the EAW.

Perception on Severity and Consistency of Marking EAW by Assessors and Students

Analyses from the semi-structured interviews revealed a stark contrast between what was perceived by the students and the lecturers with the MFRM analysis obtained. Two questions were posed to students:

1. Do you think your lecturer was fair in marking your essays? Please provide your reasons.
2. Do you think your lecturer was strict in awarding you the marks? Please state your reasons.

For the first question, all the students believed their lecturers were fair in awarding the EAW marks. Two themes emerged from their reasoning: 1) marks awarded reflected students' performance or ability, and 2) marks awarded the assessment criteria. More than half of the students mentioned that the marks awarded were based on their performance in writing, and therefore, they perceived it as fair. For example, S2 remarked that "because it depends on my writing task. She knows how to evaluate it," while S23 justified the marks given by stating (verbatim), "I can see which task my weakness and the marks are given is what I deserve." Nearly half of the students also opined that their lecturers assessed their EAW based on the assessment criteria. For example, S1 justified the marks received by stating, "I know my lecturer gave it by following the guidelines." At the same time, S20 observed that "I think everyone is treated fairly according to the rubric."

As for the second question, most students believed their lecturers were strict in awarding them the marks. Only two students (S11 and S21) were not sure whether their lecturer (L3) was strict in awarding them marks, while five students (S1, S4, S18, S20, and S22) thought that their lecturers (L2, L3, and L4) were not

strict in awarding marks. Most students, justified their reasoning positively despite stating that their lecturers were strict in awarding marks. For example, S6 remarked that "I did not get a very high mark but get the marks that equivalent to what I do," and S12 concurred by claiming that "because she gives the marks follow by student's talented (skills)." S9 believed that his lecturer had to be strict because "she needs to do so to make sure all her students were excellent."

When questions on severity and consistency of marking were directed towards the lecturers, most lecturers maintained that they would not be strict unnecessarily as they followed the assessment criteria closely while marking the written assessment. It is evident from their responses:

L1: *"I'm not strict in awarding the marks but at the same time I would follow the assessment criteria closely. I will not penalize the marks unnecessary."*

L5: *"Scripts were assessed on fluency, organization and language. Therefore, being strict is a subjective connotation."*

As for consistency, most lecturers believed that they were consistent in their marking as illustrated by the reasoning given by L1 ("I will compare the marks with other scripts if I have any doubt with the marks that I have awarded") and L3 ("I follow the criteria while marking and it is always in front of me").

DISCUSSIONS AND IMPLICATIONS

Findings from the MFRM analysis indicated that only one lecturer was more severe than others (L5: 1.66 logits on severity measure). In contrast, most students perceived that their lecturers were severe or strict in awarding marks, albeit accompanied by positive reasons for why they deserved the marks. This finding is consistent with studies by Fahim and Bijani (2011), and Erguvan and Dunyait (2020), which found that despite training provided, assessors' severity and harshness could not be eliminated. In addition, both students and lecturers were generally unanimous in their perceptions of assessment fairness. This could be attributed to the fact that both parties were exposed to the scoring guide and criteria at the onset of the study (Phase 1). Their responses mirrored Nisbet and Shaw's (2020) 'felt fairness.' In their book 'Is Assessment Fair?' they argued that a sense of fairness carries 'emotive force' and thus, any perception towards fairness in assessment deserves attention. In fact, they highlighted the challenges in 'harmonizing' other assessment concepts, such as validity and reliability with assessment fairness. Since fairness is subjective, students and lecturers' responses in this study were valuable. They provided a glimpse of how quantitative and qualitative findings could offer an inclusive view of assessment concepts.

Consistency or reliability of marking is important in ensuring quality marking. This study indicated that most lecturers were reliable markers based on their infit

mean-squares— ranged between 0.5 and 1.5. In addition to the training provided, it could be hypothesized that their experience in marking standardized examination papers like the MUET might have helped them internalize the assessment criteria. In this study, only L5 (infit mean square: 0.41 logits) did not have extensive experience in marking compared to the rest of the lecturers. However, L5's lack of internal reliability should not be construed as the failure of training given. Other factors could affect the reliability of markers, such as rater fatigue (Mahshanian & Shahnazari, 2020).

Based on the discussion of findings, this study offers a two-pronged solution to two assessment concerns. The first concern pertains to producing evidence of quality marking of written assessment, and the second is to diagnose misfit assessors for remedial measures. Providing a quality rubric does not necessarily translate to quality marking as its application or interpretation may get lost in translation. Therefore, using statistical analyses such as MFRM may provide evidence of quality marking. Educational institutions could download the free version of MFRM (Minifac), which enables its user to analyze up to 2000 data (Linacre, 2014).

From the MFRM measurement reports, misfit assessors could be identified, and remedial measures could be taken. For example, more training and moderation exercises could be prepared for assessors who exhibit variability in marking. Assessor variability could not be eliminated in any performance assessment. However, by

devising appropriate measures to control the marking quality, students will receive fair and just marks or scores that correspond with their ability.

CONCLUSIONS

Many studies on severity and consistency of raters in marking written assessment reported that rater training was crucial in maintaining quality marking. (Park, 2011; Han & Huang, 2017; He, 2019, McNamara et al., 2019). The findings of this study seemed to corroborate this stance as most lecturers were able to mark after training was provided reliably. In addition, the utilization of the MFRM in gauging severity and consistency measures of assessors' tendency in marking contributed to the burgeoning literature of performance assessment. The availability of psychometric testing software such as MFRM enables educational institutions to portray quality marking accurately. Triangulation between Rasch analyses and students' and lecturers' interview responses produced interesting insight into assessment fairness. Fairness has always been a persistent contention in any performance assessment, and hopefully, this finding could add value to its literature.

There were some limitations identified in this study. Firstly, it was found that despite the EAW being randomly selected, the students' scores revealed that most of them were categorized as proficient. This could affect their perception of fairness since the marks were in their favor. It would be ideal to employ students with varying proficiency levels (beginner, intermediate and advanced) in future studies and then interview them on their perception of fairness. Secondly, there were only five lecturers involved in this study. Despite obtaining sufficient data points for MFRM analysis, using a bigger number of lecturers might yield different results in terms of severity and leniency measures. Thirdly, the training provided in this study was short due to lecturers' work commitment. Thus, future studies may want to include longer training hours in their procedures, particularly for novice assessors.

ACKNOWLEDGEMENT

REFERENCES

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31. https://doi.org/10.1007/s11092-008-9068-5

Carrillo-de-la-Pena, M. T., & Perez, J. (2012). Continuous assessment improved academic achievement and satisfaction of psychology students in Spain. *Teaching of Psychology, 39*(1), 45-47. https://doi.org/10.1177/0098628311430312

Council of Europe. (2009). *Common European Framework of Reference for Languages:*

Learning, teaching, assessment. Cambridge University Press.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Engelhard, G. (1994). Examining rater errors in the assesment of written Ccmposition with a Many-Faceted Rasch Model. *Journal of Educational Measurement, 31*(2), 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Erguvan, I. D., & Dunya, B. A. (2020). Analyzing rater severity in a freshman composition course using Many-Facet Rasch measurement. *Language Testing in Asia, 10*(1), 1-20. https://doi.org/10.1186/s40468-020-0098-3

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing, 1*(1), 1-16.

Hack, S. (2019). *How do examiners mark? An investigation of marking processes used in the assessment of extended written responses* [Unpublished Doctoral dissertation]. University of Surrey.

Han, T., & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand, 53*, 112-147.

He, T. (2019). The impact of computers on marking behaviors and assessment: A many-facet Rasch measurement analysis of essays by EFL college students. *SAGE Open, 9*(2), 1-17. https://doi.org/10.1177/2158244019846692

Jiminez, C. E. (2015). *Middle school students' perceptions of fairness and trust in assessment scenarios* (Doctoral dissertation). University of South Carolina, US.

Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research, 57*, 113-136. https://doi.org/10.14689/ejer.2014.57.2

Lang, W. S., & Wilkerson, J. R. (2008, February 7-10). *Accuracy vs. validity, consistency vs. reliability, and fairness vs. absence of bias: A call for quality.* Paper presented at the Annual Meeting of the American Association of Colleges of Teacher Education (AACTE). New Orleans, LA.

Levey, D. (2020). *Strategies and analyses of language and communication in multilingual and international context*. Cambridge Scholars Publishing.

Linacre, J. M. (2014). *A user guide to Facets, Rasch-model computer programs*. Winsteps.com

Mahshanian, A., & Shahnazari, M. (2020). The effect of raters fatigue on scoring EFL writing tasks. *Indonesian Journal of Applied Linguistics, 10*(1), 1-13. https://doi.org/10.17509/ijal.v10i1.24956

McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, justice & language assessment - Oxford applied linguistics.* Oxford University Press.

Meadows, M., & Billington, L. (2005). *A review of the literature in marking reliability.* National Assessment Agency.

Mikre, F. (2010). The roles of assessment in curriculum practice and enhancement of learning. *Ethiopian Journal of Education and Sciences, 5*(2), 101-114. https://doi.org/10.4314/ejesc.v5i2.65376

Morin, C., Black, B., Howard, E., & Holmes, S. D. (2018) *A study of hard-to-mark responses: Why is there low mark agreement on some responses?* Ofqual Publishing.

Nisbet, I., & Shaw, S. (2020). *Is assessment fair?* SAGE Publications Ltd.

Park, Y. S. (2011). *Rater drift in constructed response scoring via latent class signal detection theory*

and item response theory [Doctoral dissertation]. Columbia University.

Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using Many-Faceted Rasch Measurement. *Psicológica, 35*, 385-397.

Sundqvist, P., Sandlund, E., Skar, G. B., & Tengberg, M. (2020). Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology, 8*(10), 3-29. https://doi.org/10.46364/njmlm. v8i1.605

Tierney, R. D. (2016). Fairness in educational assessment. In M. A. Peters (Ed.), *Encyclopedia of Educational Philosophy and Theory* (pp. 1-6). Springer Science+Business Media. https://doi. org/10.1007/978-981-287-532-7_400-1

Walde, G. S. (2016). Assessment of the implementation of continuous assessment: The case of METTU University. *European Journal of Science and Mathematics Education, 4*(4), 534-544. https:// doi.org/10.30935/scimath/9492

Willey, K., & Gardner, A. (2010, November 18-19). *Improving the standard and consistency of multi-tutor grading in large classes* [Paper presented]. ATN Assessment Conference 2010. University of Technology Sydney, Australia.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501-527. https://doi. org/10.1177/0265532214536171CES

APPENDICES

Appenidx A

A sample essay question on personal statement

PART	ELEMENT	QUESTION
1	Introductory paragraph	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write an introductory paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.
2	Body paragraph 1 (past tense)	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a body paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.
3	Body paragraph 1 (present tense)	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a body paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your own personal experience.
4	Body paragraph 1 (future tense)	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a body paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.
5	Conclusion	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a conclusion paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.

Appendix B

Holistic criteria for Parts 1 (Introduction) and 5 (Conclusion)

Component	5 (Excellent user)	4 (Good user)	3 (Competent user)	2 (Modest user)	1 (Limited user)
PART 1 Introductory paragraph	Excellently developed introductory paragraph. Ideas are excellently linked.	Well-developed introductory paragraph. Ideas are well linked.	Satisfactorily well-developed introductory paragraph. Ideas are satisfactorily linked.	Some development for the introductory paragraph. Ideas are poorly linked.	Limited development for the introductory paragraph. Ideas are not linked.
PART 5 Concluding paragraph	Conclusion summarizes the main topics without repeating previous sentences.	Conclusion summarizes the main topics with minimal repeated sentences.	Conclusion summarizes the main topics with some repeated sentences.	Conclusion summarizes the main topics poorly, repeating previous sentences.	Ends abruptly or no conclusion given.

Appendix C

Analytic criteria for Parts 2,3 and 4 (Content Paragraphs)

Component (PART 1,2,3)	5 (Very good)	4 (Good)	3 (Fair)	2 (Weak)	1 (Very weak)
FOCUS	Takes a clear position and supports it consistently with well-chosen reasons and/or examples; may use strategies to promote oneself.	Takes a clear position and supports it with relevant reasons and/or examples through much of the paragraphs.	Takes a clear position and supports it with some relevant reasons and/or examples; there are some developments in paragraphs.	Takes a position and provides uneven support; may lack development in parts or repetitive OR paragraphs are no more than a well-written beginning.	Attempts to take a position but the position is very unclear OR takes a position but provides minimal or no support, may only paraphrase the prompt.
ORGANISATION	It is focused and well organized with effective use of transitions.	It is well organized but may lack some transitions.	It is generally organized but has few or no transitions in parts of the paragraphs.	It is disorganized in parts of the paragraph; other parts are disjointed and/ lack transitions.	It is disorganized or unfocused in much of the paragraphs OR is clear but too brief.
LANGUAGE	Consistently exhibits variety in sentence structure and word choice. Errors in grammar, spelling, and punctuation are few and do not interfere with understanding.	Exhibits some variety in sentence structure and uses good word choice; occasionally, words may be used inaccurately. Errors in grammar, spelling, and punctuation do not interfere with understanding.	Most sentences are well constructed but have a similar structure; word choice lacks variety or flair. Errors in grammar, spelling, and punctuation but do not interfere with understanding.	Sentence structure may be simple and unvaried; word choice is mostly accurate. Errors in grammar and spelling and punctuation sometimes interfere with understanding.	Sentences lack formal structure, and the word is often inaccurate. Errors in grammar and spelling and punctuation interfere with understanding in much of the paragraphs.

Relating a Sustained Monologue Speaking Production Test to CEFR: Towards Alignment

Hazita Azman[1]*, Zarina Othman[2], Chairozila Mohd. Shamsuddin[2], Wahiza Wahi[2], Mohd Sallehuddin Abd Aziz[1], Wan Nur'ashiqin Wan Mohamad[2], Shazleena Othman[2] and Mohd Hafiszudin Mohd Amin[2]

[1]*Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
[2]*School of Liberal Studies (CITRA UKM), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*

ABSTRACT

This article evaluates a sustained monologue speaking production test to validate its link to the CEFR model. The monologue test is a low-stakes production test that engages the test taker in sustained monologue tasks targeted at B2-C1 of the CEFR levels. The evaluation of the test included determining the extent to which the monologue speaking tasks and the single assessment criterion-related rating scale developed for the test are valid and reliably aligned to CEFR benchmarked descriptors. The socio-cognitive framework for test evaluation was adopted, and an explanatory sequential mixed-methods research design was implemented. The evaluation revealed some contentious points of contrast between the test items and the language demand that each item prompted in production. Consequently, selected items were improved or deleted to ensure the appropriate competency levelled at B2-C1 are correctly prompted. Additionally, the findings underlined the imperative need for test developers to adhere to five inter-related sets of procedures in the justification of a claim that the monologue speaking test is aligned to the CEFR. These include familiarisation, specification, standardisation and benchmarking, standard-setting, and validation. It emerged that thorough familiarity with the CEFR by test item writers and examiners is a fundamental requirement for a test closely related to CEFR construct and levels. Thus, familiarisation training of CEFR and its illustrative

Hazita Azman, Zarina Othman, Chairozila Mohd. Shamsuddin, Wahiza Wahi, Mohd Sallehuddin Abd Aziz,
Wan Nur'ashiqin Wan Mohamad, Shazleena Othman and Mohd Hafiszudin Mohd Amin

descriptors is a mandatory prerequisite for ensuring test items and assessment of the elicited production correspond to the levels and ratings described in the CEFR model.

INTRODUCTION

When the Ministry of Education Malaysia (2015) decided on CEFR as the governing framework of international standards for developing English language proficiency programmes at preschool, school and tertiary levels, the need to align language curriculum, teaching and learning, and assessments to CEFR became obligatory. In doing so, the corresponding content and performance levels descriptors drawn from CEFR were made the target proficiency level for each of the education stages (Ministry of Education Malaysia, 2015): preschool at A1, primary at A2, secondary at B1, post-secondary at B2, tertiary at B2-C1, and teacher education at C1-C2 (Khan et al., 2019; Uri & Aziz, 2020). Furthermore, in line with the Ministry of Education Malaysia's (MoE) aspirations, the Ministry of Higher Education of MoHE (2018) required universities to align their English language assessments to CEFR or adopt CEFR aligned proficiency tests.

Hence, the initiative to develop and implement a sustained monologue speaking production test at a local university was motivated by three major factors. Firstly, the ability to speak and communicate proficiently in English has been commonly identified as a competency sought after by employers when recruiting new graduates. Second, the onset of globalisation has made this requirement increasingly imperative for non-native speakers of the language (Manokaran et al., 2021).

Secondly, the launch of the roadmap for English language education reform by the Ministry of Education Malaysia in 2015 provided direction for the standards of English language competencies that language curriculum from preschool to tertiary levels are expected to reach. These standards, informed by the Common European Framework of Reference for languages or CEFR (Council of Europe, 2011), stipulated students at the tertiary level to graduate with at least a minimum proficiency level equivalent to an independent user at CEFR B2-C1 levels (Ministry of Education Malaysia, 2015). Towards this end, MoHE required universities to employ CEFR aligned tests only to report their students' proficiency levels (MoHE, 2018).

However, subjecting students to CEFR aligned examinations that are readily available in the market raises the issue of affordability, especially for most students at public universities. Thus, developing an internal low stake test became the preferred option for our university. As such, this circumstance is the third impetus for developing the Sustained Monologue Speaking Production Test or SMSPT, henceforth, using the CEFR model as a referred criterion of standards.

SMSPT was designed to elicit long turn speaking samples that can be assessed to gauge the ability to speak directly on a selected topic in a sustained monologic communication style. The test is conducted face-to-face with an interlocutor who prompts the test taker to respond to a selected speaking topic. The topics are thematically linked to social and workplace domains. The candidates are given a few minutes to understand the question before responding. Then, they are allowed to enquire for clarification from the interlocutor if necessary. Finally, they are given a maximum of three minutes to respond. The test performance is recorded and rated remotely by two trained examiners.

The developers of SMSPT were informed by several CEFR resources, which included the *Manual for relating language examinations to the CEFR* (Council of Europe, 2009), the *Structured overview of all CEFR scales* (Council of Europe, 2011), the *CEFR Companion Volume* (CoE, 2018), and the updated series of the *CEFR manual 2020* (Council of Europe, 2020). These documents helped familiarize the test developers with constructs of targeted language competencies and specified tasks that elicit language production for the targeted proficiency levels.

This article describes the evaluation conducted on SMSPT towards validating its alignment to CEFR. The evaluation is informed by the Council of Europe (CoE) manual published in 2009 and 2020, which systematically delineates "procedures in a cumulative process to situate examinations in relation to the CEFR" (Council of Europe, 2009, p. 9). The article proceeds to describe the content analysis of SMSPT to determine its cognitive and context validity concerning CEFR, guided by Weir's (2005) socio-cognitive framework for language test validation. Finally, the article illuminates contrasts found between SMSPT and the CEFR model while highlighting implications for changes to SMSPT and the sets of procedures essential towards aligning the test to CEFR.

METHODOLOGY: TOWARDS ALIGNING SMSPT TO CEFR

The first step towards aligning a test to CEFR requires test developers to show how their tests can be related to CEFR in terms of "test content and assessment criteria, and how performance on the language test is interpreted" (Council of Europe, 2011, p. 7). According to the CoE, relating an examination or a test to CEFR "entails implementing five inter-related sets of procedures" (Council of Europe, 2009, p. 9), as depicted in Figure 1. It includes familiarisation, specification, standardisation and benchmarking, standard-setting, and validation processes.

The subsequent section describes the extent to which SMSPT adhered to these five inter-related sets of procedures. The evaluation of this adherence was conducted by an external group of CEFR experts in relating the extent to which SMSPT is aligned to the criteria features of CEFR.

Hazita Azman, Zarina Othman, Chairozila Mohd. Shamsuddin, Wahiza Wahi, Mohd Sallehuddin Abd Aziz,
Wan Nur'ashiqin Wan Mohamad, Shazleena Othman and Mohd Hafiszudin Mohd Amin
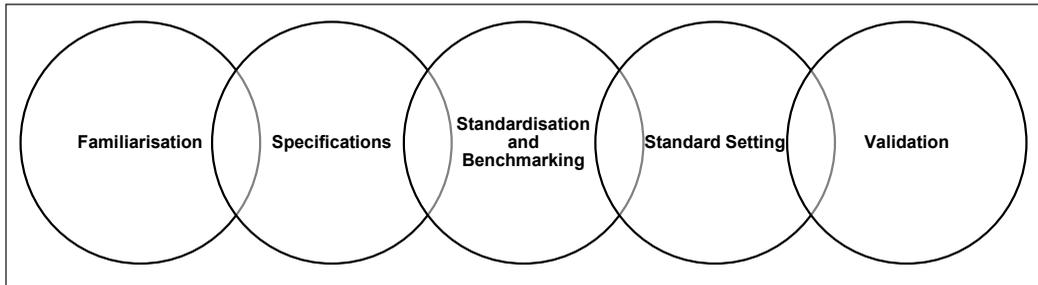
Figure 1. Five inter-related sets of procedures for relating an examination to the CEFR
Source. Relating Language Examination to the CEFR: A Manual (Council of Europe, 2009)

Linking SMSPT to the Five Inter-related Sets of Procedures

It is helpful to begin this section with a brief description of how CEFR views speaking competency. First, the CEFR model makes a distinction between spoken production and spoken interaction that is the ability to speak individually and to interact with two or more people, respectively, on a variety of topics, from familiar to less familiar, situated in domains ranging from social, educational, and occupational, and extended degrees of formality (Council of Europe, 2018).

In CEFR, the spoken production encompasses the ability to produce sustained monologue in the form of "short and simple" directional speech to detailed descriptions and "presentations on complex subjects" in long turn forms (Council of Europe, 2018, p. 68). These monologic tasks may include describing experiences, giving opinions or information, putting a case forward or addressing familiar to complex topics situated in various contexts and domains.

On the other hand, spoken interaction illustrates the ability to interact with verbal exchanges in pairs and groups. The speaker demonstrates the competency in turn-taking skills to initiate, maintain and end the interaction and intervene in ongoing exchanges when appropriate (Council of Europe, 2018). These interaction tasks may include conversations, dialogues, interviews, and group discussions that elicit short turns and joint constructions of discourse to manage and sustain turn-taking in the pair or group interactions.

While CEFR descriptors specify what language learners can do at different proficiency stages (Council of Europe, 2011), it does not clarify or illustrate what materials or tasks should be designed to elicit these abilities for assessment. Furthermore, it does not explain, as it was never intended to do in the first place, how learners can develop their knowledge of spoken English to get to the next CEFR level (Don, 2020). Herein lies the gap between the CEFR specifications and how to operationalize them in translating them into language curriculum, teaching the targeted level and assessing the targeted proficiency. This section addresses the aspect concerning assessment in this lacuna. It describes how we attempted to interpret the specifications and translate them into test items towards

relating SMPST to CEFR within the frames of the five inter-related sets of procedures.

Familiarisation

Familiarisation is a procedure where "the language test developer must demonstrate an in-depth knowledge and understanding of the CEFR descriptors that illustrate the salient features of the language proficiency in different skills at the different levels" (Council of Europe, 2009, p. 17).

The SMSPT test developers comprised ESL experts, who received a one-week familiarisation training conducted by the Cambridge Assessment English (CAE) experts. Based on their shared understanding of CEFR obtained through related documents and training, the test developers derived task specifications from the B2 CEFR descriptors. Based on these specifications, 80 monologue task items were developed, and only 50 were selected for SMSPT after a pilot test analysis. Test items were randomly selected from this selection by an interlocutor during the speaking test.

Specifications

Specification procedure requires "detailed descriptions of the test, profiling its test specifications for content analysis and verification of the abilities that are tested can be related to the relevant CEFR descriptors, categories and levels" (Council of Europe, 2009, p. 29). The specifications specify (1) the speaking production abilities that can be assessed at the targeted levels of proficiency, (2) the types of real-world speaking purposes that the targeted abilities and level of proficiency will fulfil, and (3) the rating descriptors that distinguish one level of proficiency from another to rate the performance of these competencies as concisely and comprehensibly as possible.

The CEFR model (Council of Europe, 2011) identifies five spoken production tasks—*addressing audiences, public announcements, describing an experience, giving information,* and *putting a case*. The SMSPT test developers *described the experience* and *put a case* as the two categories of production tasks that are assessed. It is mainly because these speaking tasks are commonly practised in English language proficiency courses at the university.

Standardisation and Benchmarking. Standardisation training is an extended part of the familiarisation cycle where test examiners or raters work with exemplar performances and test tasks to achieve an adequate understanding of CEFR levels and develop an ability to relate the local test tasks and performances to those levels (Council of Europe, 2009). For SMSPT, both standardisation and benchmarking were conducted in the same session, following the procedures explicated in the CEFR manual (Council of Europe, 2009, p. 40-53). In addition, rater standardisation documents containing selected exemplar performance from validated pilot sessions, sample tasks, rating scales, and sample marks were compiled for the one-day examiner training session.

Hazita Azman, Zarina Othman, Chairozila Mohd. Shamsuddin, Wahiza Wahi, Mohd Sallehuddin Abd Aziz,
Wan Nur'ashiqin Wan Mohamad, Shazleena Othman and Mohd Hafiszudin Mohd Amin

The benchmarking session progresses with sample tasks from the actual test, where test examiners practice rating the production videos individually and in small groups. Finally, a plenary group discussion is conducted to reach a consensus regarding assigning a particular performance to a CEFR level. A single assessment criterion that referenced CEFR with bands corresponding to A2-C1 levels was used to rate the test performance. To confirm inter and intra rater reliability, training of interlocutors and test examiners was conducted to ensure standardisation in the rating of test performances across examiners.

Standard-Setting. Standard-setting procedures (Council of Europe, 2009) is related to establishing the overall validity and reliability of the test concerning its alignment to CEFR standards, categories and levels. Concerning SMSPT, the performance level standards are drawn mainly from the "Can do" statements in CEFR for monologue spoken production descriptors (Council of Europe, 2018, pp. 68-73). The assessment criteria used for SMSPT covers levels A2 to C1, and the standardisation training provided shows cased exemplars of test performances that were gauged at the said levels. Of course, the concern with the standard-setting results applied for SMSPT is whether the CEFR level allocated to the student performances is trustworthy.

We now turn to the discussion about the validation process and procedures, the fifth and final phase in the process of linking a test to CEFR. However, this discussion after that will be restricted to the examination of the SMSPT test items, mainly to highlight salient aspects of the monologue tasks, in terms of cognitive, context, scoring and criterion-related validity (Weir, 2005), and ways in which the test can more clearly be linked to CEFR.

Validation. The validation procedure conducted on SMSPT involved a content analysis of its test items to determine the extent to cognitive validity, context validity, scoring validity, and criterion-related validity can be linked to the CEFR descriptors and established standards. Hereafter, the scope of discussion related to validation is limited to highlighting the points of similarity and contrast between SMSPT test items and CEFR descriptors. To this point, the framework analysis of the evaluation conducted on SMSPT to justify its link to CEFR is explicated.

Framework Analysis of the Evaluation Conducted on SMSPT

Content Analysis of SMSPT Test Items. As part of the validation, the procedure to link SMSPT to CEFR, a content analysis of the test was conducted with three primary purposes in mind: 1) To examine the test items for evidence of cognitive validity and context validity (Cambridge Assessment English, 2019) in order to validate the extent to which the monologue tasks elicit the competencies described in CEFR for level B2 specifically. 2) To investigate the scoring validity of the test to determine the reliability of the judgment by the test

examiners in rating the test performances to a CEFR level. 3) To reference the test, for evidence of criterion-related validity, to external validations. In the case of SMSPT, this entailed comparing the judgments of external experts trained with CEFR knowledge with the scores allocated by SMSPT test examiners.

The mediating theoretical framework employed for the validation process is Weir's (2005) socio-cognitive framework for language test validation. It is in line with the use of language for social purposes as defined in CEFR.

The framework adopts an interactionist position in defining language ability construct where "ability is defined both in terms of cognitive abilities and mental processing of individual learners as well as the interaction of these abilities with the surrounding social and contextual factors" (Cambridge Assessment English, 2019, p. 8). Weir (2005) identifies five critical

components of test validity as indicated by the darkened boxes in Figure 2. Only four components, namely cognitive, context, scoring and criterion-related validity, will be predominantly discussed in the findings. Consequential validity would require an extensive study which is beyond the scope of this article.

An explanatory sequential mixed-methods approach (Creswell & Clark, 2011) was adopted for the validation process of the test item analysis as "it allows for qualitative methods to establish a rich explanation of the quantitative results from the participants' perspectives" (Zeiglar & Kang, 2016, p. 56). Thus, the first phase of the data collection was quantitative, and which was then fed into a qualitative focused stage before both the quantitative and qualitative data were combined to provide an integrated interpretation of the findings, as shown in Table 1.
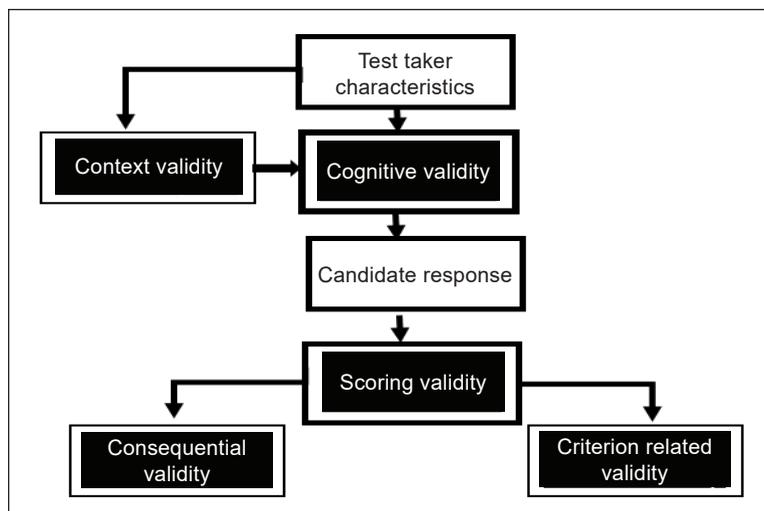
Figure 2. A socio-cognitive framework
Source. Weir, C. J. (2005), pp. 44-47

Hazita Azman, Zarina Othman, Chairozila Mohd. Shamsuddin, Wahiza Wahi, Mohd Sallehuddin Abd Aziz,
Wan Nur'ashiqin Wan Mohamad, Shazleena Othman and Mohd Hafiszudin Mohd Amin

Table 1
Sequential explanatory mixed-method research design

Focus of analysis	Phase 1: Quantitative ⟹	Phase 2: Qualitative ⟹	Integration & Interpretation
Test appraisal of monologue task items by evaluators	• 50 Test items were randomly selected and examined by a panel of 10 experts. • The panel comprised CEFR trained experts with more than ten years of experience in assessing speaking. • The panel were asked to judge and assign each item a CEFR level. It was done separately and individually.	1. Data gathered from the quantitative phase was processed and reviewed. 2. The quantitative data were referred to in a focus group discussion with the panel of experts. 3. Annotations from this phase are transcribed and inform the interpretation and integration of the data. 4. The discussion was framed in line with aspects of the socio-cognitive framework.	Quantitative and qualitative data are combined to provide a rich review of the content analysis of the test items and the extent to which they can be linked to the CEFR.
Test performance appraisal: video and scores allocated	• Video recordings of Test performances taken during SMSPT were viewed and judged by the panel of 10 experts. • The panel were asked to rate and allocate a score to each performance based on a single-criterion scale drawn from the CEFR 2017 descriptor tables for spoken language. • These judgements were compared with the scores already allocated by examiners of the SMSPT.		

FINDINGS AND DISCUSSION

Cognitive Validity

The construct of cognitive validity establishes the types of cognitive processing or cognitive load that is activated by the test question and the extent to which the cognitive processes required to respond to the question are appropriate for the target language level, candidates, and purpose of the test (Taylor, 2011). Field (2011) proffers a cognitive processing model for speaking, depicting six stages of how a speaker processes information in preparation for speech production: conceptual, syntactic, lexical, phonological, phonetic, and articulatory stages. The model is depicted in Figure 3. The following section summarises the findings concerning the cognitive processing triggered by test items of SMSPT.

Cognitive Processing Triggered by Test Items in SMSPT

According to Taylor (2011), the design of a speaking task must be mindful of the cognitive demands that the given task may have on the test taker. Thus, the test taker's performance is highly dependent on whether the speaking task required of them is familiar and is pitched at a suitable level in terms of ideas or topics and linguistic complexity. Table 2 illustrates descriptors in the CEFR scale for overall spoken production that offers ideas for speaking tasks. At the B1 level, for instance, the focus of the speaking tasks should be on personal and everyday information, 'within his/her field of interest, presenting it as linear sequence of points.' In contrast, at the B2 level, the items should prompt developed descriptions on a wide range of subjects, expanded with supporting ideas and relevant examples on familiar and less familiar topics of interest.

Regarding CEFR, topic familiarity and any other reliance on content knowledge that can facilitate rather than inhibit performance are important features to carefully consider when selecting ideas for topics of spoken production assessment (Alderson, 2000; Galaczi & French, 2011). For example, with
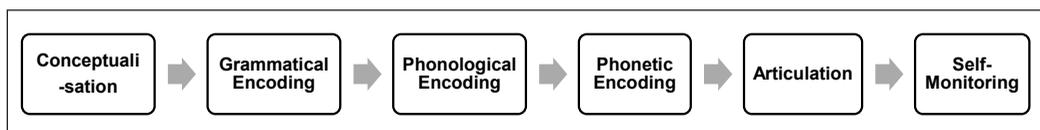
Figure 3. A model of cognitive processing for speaking assessment
Source. Cognitive validity (Field, J., 2011)

Table 2
Overall spoken production CEFR scale for A2-C1 levels (Council of Europe, 2018, p. 75)

2	Can give clear, systematically developed descriptions and presentations, with appropriate highlighting significant points and relevant supporting details. Can give clear, detailed descriptions and presentations on a wide range of subjects related to their field of interest, expanding and supporting ideas with subsidiary points and relevant examples.
B1	Can reasonably fluently sustain a straightforward description of one of a variety of subjects within their field of interest, presenting it as a linear sequence of events.

SMSPT, it was found that conceptualisation of the topics and themes for the monologue speaking tasks were relatable to candidates' living experiences and are therefore suitable.

However, as the test is a one-level criterion-referenced test targeted at B2 spoken production level, all of the 50 test items should be comparable within the test. An evaluation of each of the test items examined by the evaluators found potential issues in this regard. The evaluation revealed that 20% or only 10 test items are estimated to target at B2, whereas the majority of the items or n=24 (48%) is found to be estimated between B2 and C1 levels. Meanwhile, 8 or 16% of the items are estimated to be between B1–B2 levels, and four items or 8% were found to be ranged between C1-C2 levels. However, another four items were found to be unsuitable to the CEFR category of topics. This level of parity in terms of idea provision is a pressing concern as the overall data shows that a total of 38 (76%) of the 50 test items appear to skew towards high B2-C2 levels. Figure 4 illustrates the findings by evaluators in their comparative estimation of three SMSPT items, from the theme about learning a foreign language to the CEFR level.

As Field (2011) observes, the distinction in cognitive load between a B1 and B2 task item is often most noted in the wording of the test items themselves. The way the test question is posed entails grammatical encoding to be applied by the candidates as they attempt to comprehend the purpose of the speaking task and trigger the related linguistic patterns required to perform the task successfully.

Compare, for example, the sample from SMSPT as illustrated in Figure 4. Although the three items displayed were originally designed to be comparable at the B2 level, the analysis revealed that the way the questions were worded could potentially raise the cognitive and linguistic demand of the task, resulting in the disparity. For instance, the way a candidate may respond to "What are some benefits of learning a foreign language?" and to "Students should be required to learn a foreign language. Do you agree or disagree?" would elicit oral competency of differing levels. While the likely response to the former question is factual and may elicit a simple listing of positive factors drawn from personal experiences or opinion, the latter question, by comparison, is somewhat evaluative, inviting an appraisal of personal, public, and national policy perspectives. Thus, triggering B1 and B2 levels of competencies, respectively. Similarly, the third question in
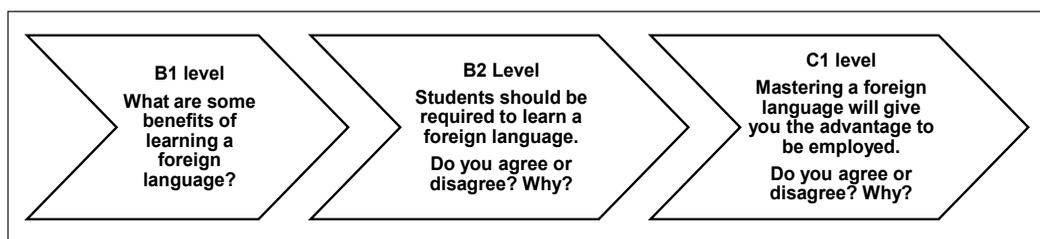
Figure 4. Evaluation of 3 items from theme on Learning a Foreign language and their CEFR level estimates

Figure 4 above is more evaluative than it is factual. It is because, as commented by an expert panel: "… the item invites nuanced views and the use of complex structures to express and defend an opinion(s)," estimated at the C1 level.

Another recurring theme that was found lacking in SMSPT is in its holistic assessment criteria scale. The current descriptions in the SMSPT scale do not reference specific spoken competencies sufficiently to provide an accurate measure of the 'cognitive processes which would prevail in a natural context' (Field, 2011, p. 66). Some of these include criterial features of phonological encoding, articulation, and self-monitoring. CEFR specifically references these features as they are also viewed as indicators of proficiency in spoken production. Table 3 shows findings

related to evaluating SMSPT holistic assessment criteria regarding these criterial features that needed improvement.

Context Validity

Weir's socio-cognitive framework identifies specific aspects to context validity for speaking. Salient aspects of these contextual factors are addressed below regarding the SMSPT task items, highlighting the extent to which the evaluation of the characteristics of the test items and their administration are appropriate to the target candidates, levels, and test purpose.

The evaluation found that the long-turn monologic task format in SMSPT is a semi-controlled response format that tends to 'elicit predominantly informational functions', typical of an examiner-candidate format. In appraising the test items, the

Table 3

Comparable findings between CEFR and SMSPT for spoken competency criterial features

Cognitive Processing features indicating spoken competence	CEFR descriptors for fluency at B2 level (CoE, 2020)	SMSPT assessment descriptors
Phonological encoding: Use of pre-assembled chunk, length of run, duration of planning pauses, frequency of hesitation and pauses	Can produce stretches of language with a fairly even tempo; although the speaker can be hesitant as he or she searches for patterns and expressions, there are noticeably long pauses.	It does not refer to a tempo, hesitations, searches for patterns and expressions and pauses.
Articulation: Use of appropriate intonation, stress, sound articulation, L1 interference, intelligibility rather than accuracy	Can generally use appropriate intonation, place stress correctly, and articulate individual sounds clearly; accent tends to be influenced by speaker's L1 but has little or no effect on intelligibility.	It does not explicitly reference the quality of articulation in speech. For example, while the descriptors refer to candidates giving 'clear information' with 'few language slips', there is no direct reference to pronunciation of what is produced.
Self-monitoring: Use of self-initiated production strategy to self-repair	Can often retrospectively self-correct occasional slips and errors in sentence production that the speaker becomes conscious of.	It refers to 'correction of slips,' but it is not clear how the ability to self-monitor is assessed.

evaluators pointed out the need for each task developed to reflect real-life skills the test taker may use or need. As Shaw and Weir (2007, p. 71) point out, 'appropriateness of task purpose enhances the authenticity of the assessment because it is imbued with a real-world purpose which goes beyond the ritual display of knowledge for assessment'. Furthermore, it underlines the main point that different purposes require different cognitive processes, which impact the difficulty of a task. Furthermore, the evaluation found that while the current SMSPT items are mainly informational functions, the 'types of talk' (Galaczi & French, 2011, p.163) can be grouped into a range of functions as listed in Table 4.

Table 4
Types of informational functions found in SMSPT

Informational functions in the CEFR	SMSPT Test items
• Expressing opinion	10
• Justifying opinions	10
• Describing	7
• Expressing preferences	7
• Providing personal information	5
• Suggesting	5
• Comparing	5
• Speculating	5

SMSPT's prevalent types of items emerges as "expressing opinions" and "justifying opinions". However, Galaczi and French (2011) note that while functions are present across various levels, some such as 'comparing' and 'speculating' are only tested at higher levels. Hence items types that ask for test-takers to "compare" and "speculate" (5 items each) need to be revised.

It was also argued that a greater range of interaction types that provide adequate coverage of open-ended formats with the interaction between peers should be included in the SMSPT test response format. The evaluation raised the issue of whether the single-question task provides adequate scaffolding for both weaker and stronger candidates. In line with the CEFR descriptors, it was suggested as well that visual and text prompts be provided as scaffolding support to facilitate the cognitive demand of the abstract questions.

Scoring Validity

Taylor and Galaczi (2011) suggest that cognitive, contextual, and scoring validity form the core of the socio-cognitive framework. As pointed out, by focusing on these three core dimensions, test developers can better develop a collection of theoretical, logical, and empirical evidence to support validity claims and arguments about the quality and usefulness of the test (p. 172). In the case of SMSPT, the evaluators made the following observations about the scoring criteria used:

- The wording of the scales is often very negative in tone. CEFR descriptors focus on what the students can do.
- There is a mismatch between some of the tasks and the descriptors in the scale. The tasks need to be revised to match the descriptors measured.
- The link between one scale level and the subsequent need to be

made more evident and specific to show gradation incompetence as illustrated in the CEFR scales.

• The contextual parameters in the current scale should include specific features such as discourse management, grammatical control, phonological encoding, articulation, self-monitoring and mediation following the updated 2020 CEFR scales.

Criterion Related Validity

Evidence of criterion-related validity can be obtained from relating a test to an external standard such as the CEFR model (Khalifa & Salamoura, 2011). For example, in the case of SMSPT, the evaluation found evidence that the test selected production functions focused on the CEFR, specifically concerning the types of talks that elicit "describing experiences and giving information" as well as "putting a case".

As noted earlier, there are aspects that SMSPT needs to further emphasise in its assessment criteria, such as include discourse management strategies, which is viewed as an indicator of fluency and competency in CEFR. For example, production strategy such as self-monitoring, i.e. "Can correct mix-ups with tenses…" at B1 level, can help distinguish competency from B2 level, where a speaker "Can correct slips and errors if he/she becomes conscious of them…." Likewise, a C1 competency who "Can back track when he/she encounters a difficulty…" can be compared with a C2 level speaker who "Can back tract and

restructure around a difficulty smoothly…" (Council of Europe, 2018, p.78). Thus the SMSPT scale should be revised to include such aspects of performance for better criterion-related validity towards CEFR alignment.

DISCUSSION AND CONCLUSION

Having explored various aspects of SMSPT, this section discusses implications of the findings and lessons learned by way of conclusion. This discussion is focused on four salient issues about the aim of relating SMSPT to CEFR.

Firstly, in terms of cognitive validity, SMSPT could be enhanced by including a variety of interaction patterns-from controlled to semi-controlled and open communication. It would broaden the construct being assessed while enabling a broader spoken production to assess, ensuring a more valid assessment of the oral performance and its criterial features. Additionally, to achieve more significant cognitive validity, SMSPT should consider including prompts, visual and textual, to lessen the cognitive demand on candidates in tackling the speaking task. It will assist in balancing support for weaker candidates while allowing stronger candidates to show the full range of their speaking ability.

Secondly, the evaluation found a notable disparity in the cognitive demands of individual questions in terms of discourse mode, nature of information, lexical and functional resources required, and topics selected. It has been noted elsewhere that greater control is needed concerning the

relative demand of one question prompt versus another if they are to be genuinely comparable and show a clearer adherence to the assumptions of the CEFR model. Thus, item analysis of test tasks will be conducted to review each speaking task's construction carefully. Vocabulary analytical tools such as Text Inspector will be applied to gauge the CEFR level of each of the test structures or rubric.

The length of familiarisation training provided for the SMSPT test developers and test examiners was inadequate. There is a need to provide prolonged training to ensure a satisfactory level of familiarisation is reached before specifications of the tasks and standardisation of judgements in rating performances can be aligned to the CEFR standards. Therefore, in reviewing SMSPT, retraining the test developers is imperative and revised test validation is a vital criterion. Better rater training would also improve the delivery of the test and encourage more consistent standardised rating.

Thirdly, the current rating scale used for SMSPT, holistic for ease of use, given many candidates to be evaluated, needs to be revised. The rating scale needs more specific descriptions related to production strategies and management discourse subskills, which further distinguishes the competent speaker from the less competent according to CEFR. In addition, it calls for a more nuanced rating system measuring aspects of production strategies such as pauses, compensating and self-correcting. Furthermore, it was pointed out that an analytic scale based on CEFR's multiple

illustrative scales for communicative activities, communication strategies, communicative language competence, and plurilingual and pluricultural competences (Council of Europe, 2020) is more valid than a holistic assessment scale. However, regarding SMSPT's test purpose and aims, aspects of plurilingual and pluricultural competences remain unnecessary for inclusion in the revised scale.

Finally, the fourth aspect for SMSPT to consider in its revision is that a speaking performance within the CEFR framework must reflect the underlying assumption that production, reception and interaction, as well as mediation, should be viewed as co-occurring facets of language use rather than activities which happen in isolation (Taylor, 2011). It suggests that there is a need for SMSPT to include a variety of interaction patterns and tasks to be presented to the candidate to ensure better test validity and enhance its coverage of the CEFR. Thus, a monologue speaking production test on its own is limited in its capacity to be linked to the CEFR completely. Considering these revelations, while remaining a sustained monologue test, SMSPT will instead expand its test tasks to address familiar to complex topics situated in various contexts and domains, using visual and text prompts.

In conclusion, the evaluation of SMSPT has revealed some contentious points of contrast between the test items and the language demand that each item prompted in production. Consequently, selected items will be improved or deleted to ensure the appropriate competency levelled at B2-

C1 are correctly prompted. The findings also underlined the imperative need for adherence to procedures for justifying the test aligned to CEFR. Finally, it emerged that familiarisation training of CEFR and its illustrative descriptors is a fundamental prerequisite for attaining this alignment.

ACKNOWLEDGEMENT

REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Cambridge Assessment English. (2019). *Review of a monologue spoken production test*. [Unpublished research technical report]. University of Cambridge.

Council of Europe. (2009). *Relating language examination to the CEFR: A manual.* Council of Europe Publishing.

Council of Europe. (2011). *Common European Framework of Reference for Languages: Learning, teaching, assessment: A structured overview of all CEFR scales*. Council of Europe Publishing.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume with new descriptors.* Council of Europe Publishing.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume*. Council of Europe Publishing.

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage Publications.

Don, Z. M. (2020). The CEFR and the production of spoken English: A challenge for teachers. *The English Teacher, 49*(3), 77-88.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 65-111). Cambridge University Press.

Galaczi, E., & French, A. (2011). Context validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 112-170). Cambridge University Press.

Khalifa, H., & Salamoura, A. (2011). Criterion-related validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 259-292). Cambridge University Press.

Khan, A. M. A., Aziz, M. S. A. & Stapa, S. H. (2019). Examining the factors mediating the intended washback of the English language school-based assessment: Pre-service ESL teachers' accounts. *Pertanika Journal of Social Science & Humanities, 27*(1), 51-68.

Manokaran, J., Soh, O. K., & Azman, H. (2021). Lecturers' perceptions towards students' English language proficiency: A preliminary study. *International Journal of Academic Research in Progressive Education and Development, 10*(2), 957-963. https://doi.org/10.6007/IJARPED/v10-i2/10461

Ministry of Higher Education Malaysia. (2015). *English language education reform in Malaysia: The roadmap 2015-2025.* Ministry of Education.

Hazita Azman, Zarina Othman, Chairozila Mohd. Shamsuddin, Wahiza Wahi, Mohd Sallehuddin Abd Aziz,
Wan Nur'ashiqin Wan Mohamad, Shazleena Othman and Mohd Hafiszudin Mohd Amin

Ministry of Higher Education. (2018). *the ecosystem for English language learning and assessment in higher education*. Ministry of Education.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Taylor, L. (Ed.) (2011). *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30). Cambridge University Press.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 234-258). Cambridge University Press.

Uri, N. F. M., & Aziz, M. S. A. (2020). The appropriacy and applicability of English assessment against CEFR global scale: Teachers' judgement. *3L The Southeast Asian Journal of English Language Studies, 26*(3), 53-65. https://doi.org/10.17576/3L-2020-2603-05

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Zeigler, N., & Kang, L. (2016). Mixed methods design - Chapter 4. In A. J. Moeller, J. W. Creswell & N. Saville (Eds.), *Studies in language testing 43: Second language assessment and mixed methods research* (pp. 51-83). Cambridge University Press.

SOCIAL SCIENCES & HUMANITIES

The Effects of Different Rater Training Procedures on ESL Essay Raters' Rating Accuracy

Souba Rethinasamy

Faculty of Language and Communication, Universiti Malaysia Sarawak, Kota Samarahan, 94300 Sarawak, Malaysia

ABSTRACT

The study investigated the effects of three commonly employed rater training procedures on the rating accuracy of novice ESL essay raters. The first training procedure involved going through a set of benchmarked scripts with scores, the second involved assessing benchmarked scripts before viewing the scores. The third was a combination of the former and latter. A pre, post and delayed post-experimental research design was employed. Data were collected before, immediately after and one month after training. Actual IELTS scripts with benchmarked scores determined by subjecting expert IELTS raters' scores through Multi-Faceted Rasch (MFR) analysis were used for the training and data collection purposes. Sixty-three TESL trainees were selected based on their pre-training rating accuracy to form three equally matched experimental groups. The trainees' scores for the essays before, immediately after and one month after the assigned training procedure were compared with the official scores for the operational essays. Although the findings indicate that generally, rater training improves raters' rating accuracy by narrowing the gap between their scores and the official scores, different training procedures seem to have different effects. The first training procedure significantly improved raters' rating accuracy but showed a decreasing effect with time. The second training procedure showed immediate as well as delayed positive effects on raters' rating accuracy. The third training did not lead to significant immediate improvement, but rating accuracy improved significantly after some time. This paper discusses the implications of the findings in planning efficient and effective rater training programmes.

Keywords: Assessing writing, rater training, rating accuracy, standardisation, validity and reliability

INTRODUCTION

Assessment is often seen as a crucial and integral part of teaching and learning. Assessment in education has been going through a major shift from traditional assessment of cognitive knowledge only to performance-based assessments. The scores derived from assessments conducted by educational institutions and testing bodies usually have critical implications on both the test takers and the stakeholders. However, subjectivity in assessing performance assessments, including written essays, poses a major threat to validity (Barkaoui, 2011; Lumley, 2002; Messick, 1994; Shabani & Panahi, 2020; Xie, 2015).

While a common yardstick referred to as rating scale or rubrics help reduce subjectivity in scoring even when multiple assessors are involved (Ragupathi & Lee, 2020), rubrics alone are insufficient to improve standardisation in scoring (Brown, 2009; Reddy & Andrade, 2010). Rater training has been an important component of assessment literary and is often recommended to increase the validity and reliability of scoring from a rubric. Rater training is also known as 'standardisation', 'moderation', 'calibration', 'parity' and 'norming' sessions (Hamilton et al., 2001; Hodges et al., 2019; Kondo, 2010; McIntyre, 1993; Schoepp et al., 2018). During rater training, raters are calibrated towards a common rubric with exemplar scripts to guide them to interpret the rubrics in a similar manner (Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010). Rater training is suggested to decrease subjectivity in rating,

keep score variations within acceptable limits and assist raters to assess according to standards set by the testing organisation. According to Alderson et al. (1995), "the training of examiners is a crucial component in any testing programme, since if the marking of a test is not valid and reliable then all of the other work undertaken earlier to construct a 'quality' instrument will have been a waste of time" (p. 105).

In the last two to three decades, rater training has become widely accepted and implemented by many educational institutions and language testing organisations such as Cambridge ESOL which is responsible for the International English Language Testing System (IELTS), Educational Testing Services (ETS), which is responsible Test of English as a Foreign Language (TOEFL), and Malaysian Examination Council which is responsible for the Malaysian University English Test (MUET) (Brown, 2000; Chan & Wong, 2004; Furneaux & Rignall, 2002; Wei & Llosa, (2015).

A literature review shows that empirical studies on rater training only started to gain some attention in the 1990s. For example, Shohamy et al. (1992) and McIntyre (1993) reported that training improves raters' ratings, particularly inter-rater and intra-rater reliability. Weigle (1998) found the training to be more beneficial to improving intra-rater reliability than inter-rater reliability. On the other hand, Engelhard (1992, 1996) reported significant differences in rater severity and accuracy even among highly trained raters. Myford

and Wolfe (2009) found significant positive and negative drift in rater accuracy over time for a small proportion of the raters. Despite similarities, the studies have reported some rather contradictory findings.

While examining rater, essay and environment effects, Freedman (1981) unexpectedly found that subtle differences in approach and input during training could lead to significant differences in rating. For example, the training that raters in Weigle's (1994, 1998, 1999) study went through consisted of the following procedures:

- reading through exemplar essays with their official scores
- assessing a set of essays and compare own scores with the official scores
- explaining reasons for own scores that differ from the official scores and reaching an understanding of the reason for the official score

Weigle (1994) also added that a complete description of the training session was not possible and that a certain amount of 'informal training' also took place as the ratings were done in a group setting where raters could see the scores given by the previous rater and receive feedback on their ratings. Trainers also did speak to the raters whose ratings were aberrant in some ways.

On the other hand, in Lumley's (2000, 2002) study, the rater training involved the following two major procedures.

- practise assessing several sample essays using the rating scale
- discuss the scores and the reasons for and against different scores,

by the trainers and/or the other members of the group

An extensive review of studies shows that rater training programmes seem to employ several procedures in various ways (Attali, 2015). The most common procedures utilised in rater training programmes are

- going through exemplar essays with their official scores (Furneaux & Rignall 2002; Knoch et al., 2007; McIntyre 1993; O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Raczynski et al., 2015; Wang et al., 2017; Weigle, 1998; Weigle, 1999)
- practise rating exemplar essays and comparing own scores with official scores (Erlam et al., 2013; Furneaux & Rignall 2002; Knoch et al., 2007; Lumley, 2000; Lumley, 2002; O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Weigle, 1998; Weigle, 1999; Wolfe & McVay, 2010)
- discuss reasons for scores (Kim et al., 2017; Knoch et al., 2007; Lumley, 2000; Lumley, 2002; O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Shaw, 2002; Weigle, 1994; Weigle, 1998; Weigle, 1999)

As cautioned by Freedman (1981), the differences in the training procedures employed during rater training sessions raise the question of whether they had contributed to the inconsistencies in the effects of rater training in language performance assessment.

Although many studies have compared the effects of different rating procedures, especially in the field of performance appraisal (Ellington & Wilson, 2017; Rosales-Sánchez et al., 2019; Tziner et al., 2000), studies comparing the effects of different rater training procedures in assessing language performances seem scarce. Despite several calls to investigate the effect of procedures used for training language performance raters so that these procedures could be put to best use (Freedman 1981; Furneaux & Rignall, 2002; Hamp-Lyons, 1990; McIntyre 1993; O'Sullivan and Rignall, 2001), only one type of training procedure, i.e. feedback, that too as a post-training procedure, has received some attention (O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Shaw 2002; Wigglesworth 1993). Although studies by Leckie and Baird (201l) and Gyagenda and Engelhard (2009) have focused on rater training in language performance assessment, they did not study the effects of the training.

Wigglesworth (1993) experimented with the potential effect of Multi-Faceted Rasch (MFR) based bias analysis feedback as a form of post-training procedure on a speaking test. The study found some evidence of improvement in rater consistency following the feedback and recommended the implementation of the bias analysis feedback into rater training. O'Sullivan and Rignall (2001) conducted an experimental study to explore Wigglesworth's (1993) suggested use of MFR based bias analysis feedback as a form

of post-training procedure in the context of writing assessment. The MFR bias analysis feedback had an additional brief written description to make it self-explanatory. Twenty trained and experienced IELTS examiners with at least two years of rating experience were involved in this study. The study utilised 81 essays written by candidates who sat for the IELTS Writing Module in 2002. The findings showed that only written feedback had a limited effect on the Feedback Group's rating performance.

O'Sullivan and Rignall (2001) hypothesised that feedback delivered systematically over a period may result in more consistent and reliable examiner performance. Shaw (2002) investigated the effect of feedback delivered over a period. The feedback given to the participants in this study was based on the official scores for the essays, with notes explaining the reasons for the scores. The participants were the Certificate of Proficiency in English (CPE) examiners. Data were collected on five successive rating occasions. The results showed a small gain in the percentage of rating with 0 band difference and a small gain in the percentage of rating with 0 and 1 band difference. Shaw (2002) attributed the small improvement in accuracy of the experienced raters to the possible inherent standardisation quality of the revised scoring rubrics.

While the studies above investigated the effect of feedback as a post-training procedure, the effects of the different procedures employed 'during' training on raters' rating have not been addressed

sufficiently. In addition, researchers have highlighted that a great deal remains unknown about the effects of different rater training procedures on raters' rating accuracy (Azizah et al., 2020; Leckie & Baird, 2011; Raczynski et al., 2015; Wolfe & McVay, 2010).

The emphasis of research on rater training in the 21st century shifted to the emergence of web-based rater training programmes. An early study by Hamilton et al. (2001) described a pilot online rater training programme and investigated the raters' attitudes toward the programme. A similar study conducted by Elder et al. (2007) also canvassed raters' responses towards the effectiveness of an online rater training programme. Knoch et al. (2007) compared the effectiveness of an online rater training programme and face-to-face rater training in a large-scale writing assessment. On the other hand, Attali (2015) compared the effect of web-based rater training between inexperienced and experienced raters. While these studies indicate the practical alternative to face-to-face rater training, the effects of the different rater training procedures employed during training to train the raters remain unanswered. Thus, it creates a huge gap in designing effective rater training courses and calls for focused investigation in this area (Shabani & Panahi, 2020).

The present study attempts to address this gap and shed light on the effects of some of the commonly employed procedures during essay rater training on raters' rating accuracy. The general research question that the study aimed to address is;

"How do the different rater training procedures affect raters' rating accuracy?":

This study investigated the effects of different essay rater training procedures on the rating accuracy of novice ESL raters.

The study attempted to answer the following research questions

RQ1. To what extent do the different rater training procedures affect ESL raters' rating accuracy immediately after training?

H_{01a}: There will be no significant difference between the rating accuracy of the Training Procedure A group immediately after training compared to before training.

H_{o1b}: There will be no significant difference between the rating accuracy of the Training Procedure B group immediately after training compared to before training.

H_{o1c}: There will be no significant difference between the rating accuracy of the Training Procedure C group immediately after training compared to before training.

RQ2. To what extent do the different rater training procedures affect ESL raters' rating accuracy stability over time?

H_{o2a}: There will be no significant difference between the rating accuracy of the Training Procedure A group one month

after training compared to before training.

H_{o2b}: There will be no significant difference between the rating accuracy of the Training Procedure B group one month after training compared to before training.

H_{o2c}: There will be no significant difference between the rating accuracy of the Training Procedure C group one month after training compared to before training.

Materials and Methods

The study employed a matched pairs quasi-experimental design with three rater training procedures, three rating occasions and three experimental groups. The first rating was done before training, the second was completed immediately after training, and the third rating was done one month after training.

Participants

Shohamy et al. (1992) and Weigle (1998) highlighted that raters' background could influence their rating. Thus, a homogeneous group of raters with similar backgrounds were selected as participants for this study. The study involved all the penultimate and final year undergraduates taking a degree in Teaching English as a Second Language (TESL) at a public university in Malaysia. The requirements to be accepted into the TESL programme are that an applicant must have obtained good grades in the Malaysian

equivalence of the GCSE and A-Level examinations. In addition, candidates also must have a distinction in the GCSE English language papers and at least a band 3 in the Malaysian University English Test (MUET). The demographic data obtained from the participants confirmed that the participants of the study fulfilled these language requirements. Also, they have not had any formal training in assessing written essays. Thus, the participants could be classified as novice essay raters.

Instruments

The materials used in carrying out the planned study must go through several proper construction stages, vetting and testing. In turn, it would help ensure that the findings from the study are not affected by the problems related to the writing task, scoring rubrics or the scripts. The International English Language Testing System (IELTS) test was mainly chosen because it is a long-established high-stakes test used in assessing international students' English language proficiency. In addition, the tasks and rating scales of the test have gone through several years of rigorous experimentation and validation.

The IELTS is designed to assess "the language ability of candidates who intend to study or work where English is used as the language of communication" (IELTS, 2003, p. 3). The IELTS test's ability of test-takers all four language skills—Reading, Writing, Listening and Speaking. IELTS provides a nine defined band level that ranges from Non-User to Expert-User as a guide for

interpreting the band scores (Green, 2003; O'Sullivan & Rignall, 2001)

The IELTS test for Writing Task 2 Version 42 (a retired version) was utilised for this study. The topic for the writing Task 2 Version 42 reads as below:

> Only parents can offer the care and attention that is necessary to a child's development. It is, therefore, wrong for both parents in a family to expect to pursue a career: one of them, whether it is the father or mother, should stay at home and look after the children.
>
> Do you agree or disagree?
>
> Give reasons for your answer.
>
> You should write at least 250 words.

Benchmarked Scripts

The study utilised IELTS essays as benchmarked scripts for training. On request, 81 essays written by candidates on the topic and the rating scale were provided by Cambridge ESOL. The official scores of the essays were determined by subjecting the certified IELTS raters' scores to Multifaceted Rasch analyses. The essays were rated using the IELTS rating scales. The rating scales are not made public; however, a public version is available at https://www.ieltsanswers.com/wp-content/uploads/2016/09/Essay-writing-criteria-official.pdf .

From the 81 essays, a total of 36 essays were selected for the study. For training purposes, two parallel sets (matched in terms of their scores) consisting of nine essays each were selected. First, the two sets

were labelled as Set A and Set B. Then, 18 essays were selected to form the operational set (Set C) used for actual rating purposes. Table 1 and Table 2 provide the list of essays and their band scores.

Table 1
List of the parallel Set A and Set B and their official bands

SET A		SET B	
Essay ID	Global band score	Essay ID	Global band score
21	3	08	3
13	4	53	4
03	5	67	5
16	5	54	5
84	6	65	6
66	6	40	6
22	7	39	7
38	7	82	7
33	8	69	8

Table 2
List of operational essays (Set C)

Essay ID	Global band score
43	3
27	4
87	4
02	5
37	5
78	5
85	5
04	6
07	6
14	6
36	6
30	7
31	7
44	7
64	7
24	8
28	8
29	8

Data Collection Procedure

This experimental study involved 3 rating occasions. Figure 1 illustrates the data collection procedure for the study.

Prior to the training rating occasion, a total of 103 TESL trainees assessed the operational essays (Set C). The scores given by the trainees were compared with the official scores for the essays to determine their rating accuracy. Based on the before training rating accuracy of the trainees, 63 of them were selected. The selected participants were randomly divided into three equally matched experimental groups consisting of 21 raters each. Thus, each group had the same rating accuracy before training. Then, each group was assigned to different training procedures. Finally, the training for each group was conducted two weeks after the first rating occasion.
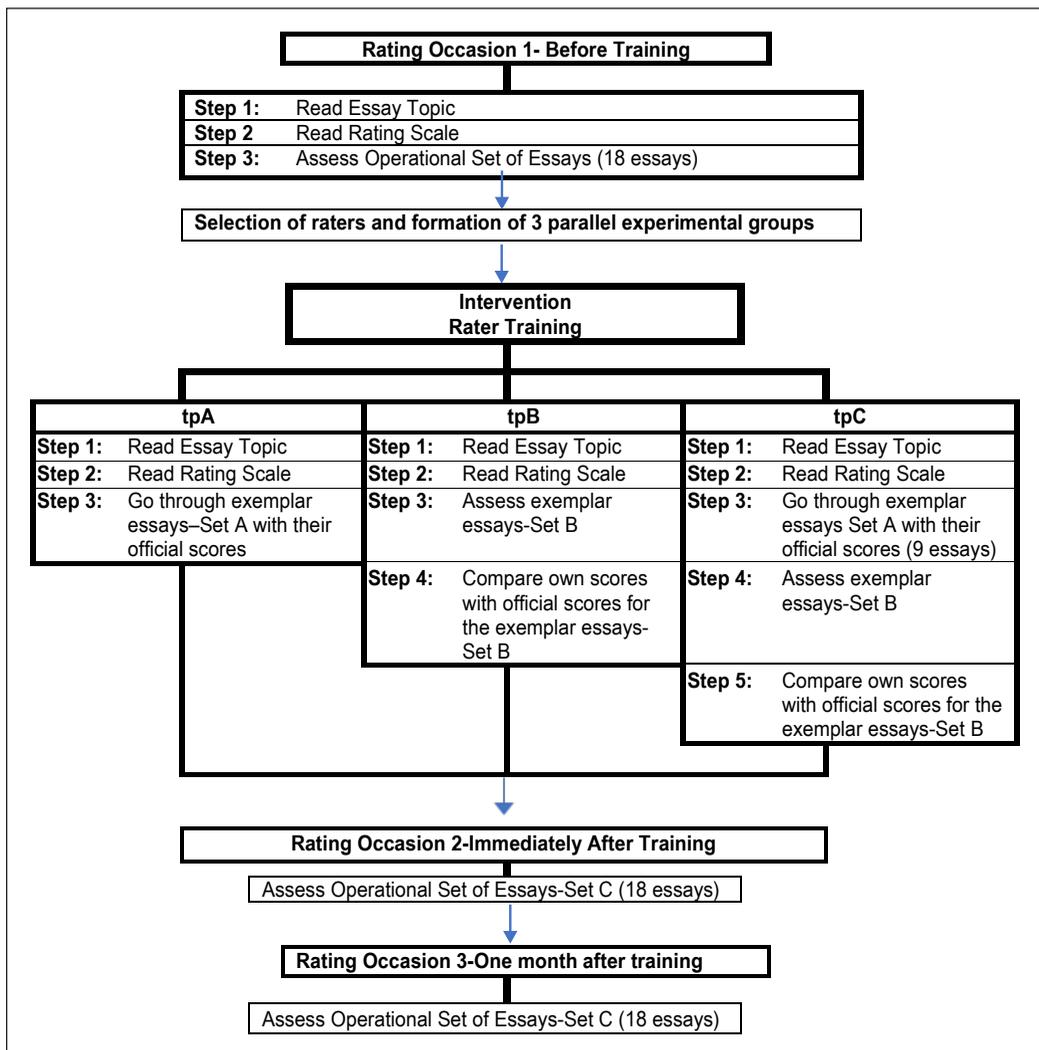
Figure 1. Data collection procedure

The first group, the Training Procedure A group (TpA), first read the topic for the essay and scoring rubrics. Then, they went through a set of exemplar essays (Set A) consisting of 9 benchmarked essays with their official scores. It took them approximately 45 minutes to complete TpA. The second group, the Training Procedure B (TpB) group, also read the essay topic and the scoring rubrics. Then, they assessed a set of benchmarked essays (Set B) and compared their scores with the official scores for the essays. TpB group took approximately 1.5 hours to complete the training. The third group, the Training Procedure C (TpC), went through a combination of TpA, followed by TpB. TpC took approximately 2.5 hours. Since the time taken to complete the training and assess the scripts is rather long, the raters were supplied drinks, snacks, and were allowed to have short breaks.

After the assigned training, each group assessed the same set of operational essays (Set C). Then, one month after training, each group assessed the operational essays again. The raters' rating accuracy before training, immediately after training and one month after training formed the dataset.

Data Analysis

For rating performance, two categories were initially formed using the data on the band difference distribution. The categories were Rating Accuracy and Rating Deviation.

Rating accuracy category refers to scores with no difference or one band difference with the Official Band. This category included all essays that differed by '0', '-1' and '+1 band from the Official Band. The reason for this category is to provide an alternative measure for accuracy by allowing a small variation from 'On-Standard' (0 band difference) as practised by most test organisations (Weigle, 2002). The number of essays scored 0–1 was calculated and converted to percentage [(number of essays within 0-1 difference/total number of essays scored)*100]. The higher the percentage of essays in this category indicates that the rater's, or in this study, the group's rating accuracy is high as the differences in scores are within an acceptable range, which in turn suggests that the quality of the group's scoring is good.

Rating Deviation refers to scores with '≥2 band difference'. This category included all essays that differed by two and more bands, regardless of whether the band difference is 'minus (-)' as when assessed harshly or 'plus (+)' when assessed leniently. When the percentage of essays in this category is high, the group's rating is deviant from the acceptable range. Thus, the accuracy is low, suggesting that the quantitative quality of the group's scoring is poor.

The 'Within 0–1' and '≥2' categories comprise the number of essays assessed in each rating occasion. In other words, when the percentage of essays in these two categories are added, they make up 100%. Thus, the two categories dovetail with each other, and so an increase in one of these categories corresponds to a decrease in the other.

The percentage of essays scored for the 'within 0-1' category and '≥2' category was calculated to determine rating quality for descriptive statistical analysis. In addition, inferential statistics were performed to examine how each of the training procedures affects rating accuracy. For this purpose, the data were input into an SPSS file and subjected to One-Way Repeated Measures ANOVA analysis with three levels of rating occasions. They are Before Training (BT), Immediate After Training (IAT), and One Month After Training (OMAT) as the within subject-factor. In addition, a post-hoc test using the Bonferroni Adjusted Pairwise Comparison procedure was also performed to determine the extent of differences between the rating occasions. The threshold p value for this study was pre-determined at .05 ($p < .05$) because the commonly used p value is .05 for educational studies (Best, 1977; Lodico et al., 2010).

RESULTS

The purpose of this experimental study was to investigate the effect of the different rater training procedures on raters' immediate and delayed rating performance compared to before training. Thus, the rating performance for each experimental group was calculated after every rating occasion, i.e., before training, immediately after training and one month after training. All the three experimental groups had baseline similarity, as indicated by the rating accuracy percentage before training (BT) in Table 3 and Table 4. The experimental groups had the same rating accuracy for the "within 0–1" category before training, i.e., 63% for within 0-1 band difference and 37% with '≥2 band difference. The three experimental groups were equally matched in terms of rating accuracy before training. Each experimental group's rating accuracy immediately after training and one month after training were compared to the rating accuracy before training to determine the effect of the different rater training procedures on raters' rating.

Rating Accuracy

As shown in Table 3, immediately after training (IAT), the rating accuracy for 'within 0-1" for TpA increased to 81%, TpB to 83% and TpC to 74%. However, one month after training (OMAT), TpA's rating for within 0-1 accuracy dropped 4% to 77%, TpB's increased 1% to 84%, while TpC's increased to 78%.

Table 3
Rating accuracy (%) 'Within 0-1 band difference'

Training Procedure	BT	IAT	OMAT
TpA	63	81	77
TpB	63	83	84
TpC	63	74	78

Rating Deviancy

As shown in Table 4, immediately after training (IAT), all three groups' rating deviancy for essays that were scored with two or more band differences with the official scores decreased to 19%, TpB to 17% and Tp C 26%. One month after training, TpA's rating deviancy was 23%, TpB's 16% and TpC 22%. As mentioned

earlier, the results for Rating Deviancy dovetails with the results for rating accuracy. Thus, the increase in rating accuracy within the 0–1 category corresponds with the decrease in rating deviancy.

Table 4
Rating Deviancy (%) '≥2 band difference'

Training Procedure	BT	IAT	OMAT
TpA	37	19	23
TpB	37	17	16
TpC	37	26	22

Figure 2 illustrates the results for rating accuracy and deviancy in graphical form.

Additionally, to examine how each training procedure affects Rating Accuracy, employing an alpha level of 0.05, the data for within 0-1 band difference were subjected to One-Way Repeated Measures ANOVA analysis, with three levels of rating occasions (BT, IAT and OMAT) as the within subject-factor. A post-hoc test using the Bonferroni Adjusted Pairwise Comparison procedure was also performed to determine the extent of differences between the rating occasions. Tables 5 shows the post hoc test results.

The results in Table 5 show that TpA's rating accuracy was at p = .005 ($p < .01$), indicating a highly significant difference immediately after training compared to before training. Furthermore, one month after training, the p value was at p =0.037 ($p < 0.05$), indicating a significant difference one month after training. Thus, H_{01a} and H_{02a} are rejected.

Although the rating accuracy on both occasions post-training was significantly different compared to before training, a clear inspection of Table 5 shows that the
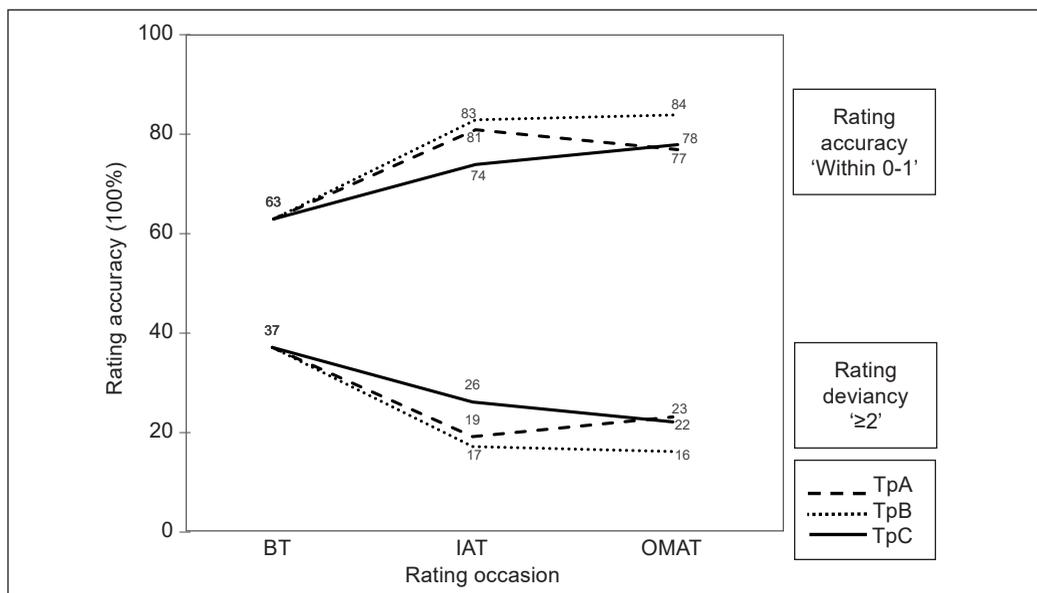
Figure 2. Graphical representation of TpA, TpB and TPC's rating accuracy before (BT), immediately after (IAT) and one-month after training (OMAT)

Table 5
Post-hoc results for rating accuracy

Training Procedure	(I) OCCASION	(J) OCCASION	Mean Difference (I-J)	Std. Error	Sig.(a)	99 Confidence Interval for Difference(a)	
						Lower Bound	Upper Bound
TpA	IAT	BT	18.254(**)	5.059	.005	5.037	31.471
	OMAT	BT	13.492(*)	4.903	.037	.682	26.302
TpB	IAT	BT	20.370(**)	4.360	.000	8.980	31.761
	OMAT	BT	21.164(**)	5.081	.001	7.890	34.437
TpC	IAT	BT	11.376	5.207	.123	2.228	24.980
	OMAT	BT	12.170(**)	3.378	.005	3.346	20.994

significance level dropped from highly significant ($p < .01$) immediately after training to significant ($p < .05$) at one month after training. It suggests that the effect of TpA is showing signs of fading.

TpB results show that raters' rating accuracy at $p = .000$ (p < .01) was highly significant immediately after training and the p value was at $p = .001$ ((p < .01) indicating that the improvement remained highly significant one month after training. Thus, H_{01b} and H_{02b} are rejected.

On the other hand, TpC results show that the rating accuracy was at $p = .123$ ($p > .05$), indicating no statistical significance immediately after training compared to before training. However, the rating accuracy became highly significant one month after training with a p value at .005 ($p < .01$). Therefore, based on the results for TpC, H_{01c} is accepted, whereas H_{02c} is rejected.

DISCUSSION

The descriptive statistics results (frequency results) showed that all three rating procedures helped improve the raters' accuracy immediately after training and reduced the number of deviant scripts. Behind the improvement in rating accuracy lies a reduction in rating deviancy. The improvement in rating accuracy and decrease in rating deviancy indicate that raters can better understanding the scoring rubrics and standards for each band level after going through rater training. Thus, it enables them to score closer to the standard set by the testing body. These findings are consistent with previous findings that training improves raters' rating performance as reported by Attali (2015), Fahim and Bijani (2011), Furneaux and Rignall (2002), Knoch et al. (2007), Tajeddin and Alemi (2014), Wang et al. (2017) and Weigle (1994, 1998, 1999). Therefore, the results from the present study affirm that for writing performance assessment, rater training is a must to ensure better validity and reliability of the scores awarded.

The post hoc test results revealed the differences in the effect of the training procedures. The results from the scores given immediately post-training showed that TpA and TpB groups' rating accuracy

improved significantly immediately after training, surprisingly the improvement for TpC group's rating accuracy was not significant. However, one month after training, all three experimental groups' rating accuracy was significantly high, indicating a positive delayed post-training effect.

It is also interesting to note that the post host test results for TpA showed highly significant ($p < 0.01$) improvement immediately post-training but dropped slightly and became significant ($p < 0.05$) during the delayed post-training. It suggests that TpA, which involved raters going through a set of exemplar scripts and official scores for each script, has an immediate positive and delayed effect, but it also revealed signs of fading. The results for TpA suggest that the positive effect of TpA may not be retained over a long time. Thus, retraining would be required to maintain stability in scoring. Lumley (2000, 2002), Shohamy et al. (1992) and Wang et al. (2017) have also highlighted that the effect of training may not last long and reinforced the need for retraining.

In contrast, TpB seemed to have a highly significant immediate and delayed effect on raters' rating accuracy. TpB required raters to assess the exemplar scripts and then compare their scores with the official scores. In TpA, raters were only asked to go through the exemplar essays with the official scores without rating the essays. In contrast, TpB is more hands-on because

the TpB group had to score the exemplar scripts before the official scores were shown to compare scores. The results indicate that rating the scripts and comparing their scores with the official scores acts as a feedback to raters on their rating performance. Although the study's feedback is not in verbal form, Hoskens and Wilson (2001) and Leckie and Baird (2011) reported a similar effect of verbal feedback on raters' drift toward the mean leading to the homogeneity of raters' scoring.

The post hoc results for TpC were unexpected. TpC, which is a combination of TpA and TpB involved longer training and more exposure to the standards. However, TpC did not seem to improve raters' rating significantly immediately after training but showed a highly significant follow-up effect. This result is rather puzzling. The reason for this could be lethargy. Since TpC is a long training session because it is a combination of TpA and TpB, it is likely that the raters became tired and could not fully concentrate on their operational scoring immediately after training. However, the significant increase in rating accuracy during delayed post-training suggests that the effect of what the TpC group have learned during the training seem to surface sometime after training. Although the improvement in rater performance over months of rating was also reported by Lim (2011), the finding on the post-training effect of TpC in this study needs further investigation, perhaps with longer breaks during the training.

Among the three training procedures, TpB, which involved raters rating the exemplar scripts before comparing their scores with the official scores seems to have a greater effect on raters' immediate and follow-up rating accuracy. Thus, the effect of training that involves more hands on or active involvement of raters tend to have more immediate as well as longer effect on raters' rating accuracy. It is consistent with the recommendation made by Wang et al. (2017).

Overall, the results for rating accuracy seem to suggest that TpB, which is longer and more hands-on than TpA and but less time consuming than TpC, appears to be more effective for immediate and follow-up positive effects. Nevertheless, TpA would be sufficiently effective and adequate for immediate rating that does not prolong over a long period. It is also crucial to remember that long training may be exhausting to the raters and detrimental to their immediate operational rating performance. Thus, if training takes long, raters should not be asked to assess operational scripts immediately. These findings have significant implications for practical and effective rater training courses, as Shabani and Panahi (2020) emphasised.

CONCLUSION

According to Reed and Cohen (2001), the rating is itself a performance, just as important as the test-takers performance and is thus worthy of investigation. This study adopted an experimental pre–post-follow-up approach to investigate the effects of different procedures employed during the training of writing raters on immediate and delayed rating accuracy. The findings from the study show that although different rater training procedures have a different effect on raters' rating accuracy, rater training does help raters to assess more accurately according to the standard set by the organisation, especially assessments that involve multiple raters. Considering the important decisions that educational institutions and organisations make using assessment scores, perhaps it is not an exaggeration to say that test organisations must train their raters not only to meet their professional obligations but more so for moral reasons.

In this study, the scoring rubrics, the exemplar scripts for training and operation scripts for scoring were chosen from a set of scripts from an established examination, i.e., IELTS. In addition, the official scores for the scripts were determined through MFR analysis of the scores given by expert IELTS raters. These could have contributed to the effectiveness of the training and consequently the raters' understanding of the standard required for scoring. It also highlights that for rater training to be effective, such careful and meticulous selection of benchmarked exemplar scripts are crucial. Nevertheless, the findings from this study offers crucial insights on the effects of different rater training procedures on ESL essay raters' ability to access according to the standard set by the testing organisation. Since rater training is not only time consuming but a costly process

(McIntyre 1993; Hamilton et al., 2001), it is hoped that the empirical evidence the study provides will help inform practitioners, test developers and test administration organisations in designing training for raters effectively and efficiently way. It is also hoped that the study will broaden the scope of research in the direct assessment of writing and other performance assessments such as speaking, in which similar rating procedures are typically used.

Previous research by Eckes (2008), Cumming (1990), and Wolfe et al. (1998) reported that more experienced raters considered factors that were not in the scoring rubrics. However, the raters in this study were novice ESL raters. Therefore, the exposure to the rubrics and the benchmarked scripts to these novice raters may have contributed to their adherence to the standards they were exposed to during the training. Consequently, this could have contributed to the positive effect of the rater training procedures. However, the effect may not be the same with expert raters. Thus, further research with raters of different rating backgrounds and test contexts are necessary.

Finally, the present study employed a quantitative approach to investigate the effects of rater training. However, it cannot be denied that quantitative similarities may camouflage differences in rating judgement, i.e., the reasons for awarding the scores. Thus, future studies could focus on investigating rater training effects on raters' qualitative judgement.

REFERENCES

Alderson, J. C., Clapman, C., & Wall, D. (1995). *Language test construction and evaluation*. Ernst Klett Sprachen.

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115. https://doi.org/10.1177/0265532215582283

Azizah, N., Suseno, M., & Hayat, B. (2020). Severity-leniency in writing assessment and its causal factors. In *International Conference on Humanities, Education and Social Sciences (IC-HEDS) 2019* (pp. 173-185). Knowledge E. https://doi.org/10.18502/kss.v4i14.7870

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy and Practice, 18*(3), 279-293. https://doi.org/10.1080/0969594X.2010.526585

Best, J. W. (1977). *Research in education* (3rd ed.). Prentice-Hall Inc.

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In *International English Language Testing System (IELTS) Research Reports 20003* (Vol. 3, pp 49-84). IELTS Australia

Brown, G. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In S. D. L. H. Meyer, H. Anderson, R. Fletcher, P. M. Johnston & M. Rees (Eds.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40-48). Ako Aotearoa.

Chan, S. H., & Wong, B. E. (2004). Assessing oral skills of pre-tertiary students: The nature of the communicative act. In *Proceedings of the International Conference on English Instruction and Assessment* (pp. 33-48). National Chung Cheng University.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51. https://doi.org/10.1177/026553229000700104

Eckes, T. (2008). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64. https://doi.org/10.1177/0265532207071511

Ellington, J. K., & Wilson, M. A. (2017). The performance appraisal milieu: A multilevel analysis of context effects in performance ratings. *Journal of Business and Psychology*, *32*(1), 87-100. https://doi.org/10.1007/s10869-016-9437-x

Engelhard, J. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191. https://doi.org/10.1207/s15324818ame0503_1

Engelhard, J. (1996) Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

Erlam, R., Ellis, R., & Batstone, R. (2013). Oral corrective feedback on L2 writing: Two approaches compared. *System, 41*(2), 257-268. https://doi.org/10.1016/j.system.2013.03.004

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing, 1*(1), 1-16.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245-255.

Furneaux, C., & Rignall, M. (2002) *The effect of standardisation-training on rater judgments for the IELTS Writing Module*. Cambridge University Press.

Green, A. (2003). *Test impact and English for academic purposes: A comparative study in backwash between IELTS preparation and university professional courses* [Unpublished Doctoral dissertation]. University of Surrey.

Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement, 10*(3), 225-246.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perception of on-line rater training and monitoring. *System, 29*(2001), 505-520. https://doi.org/10.1016/S0346-251X(01)00036-7

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.009

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., & McTigue, E. (2019). Developing and examining validity evidence for the Writing Rubric to Inform Teacher Educators (WRITE). *Assessing Writing, 40*, 1-13. https://doi.org/10.1016/j.asw.2019.03.001

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the golden state examination. *Journal of Educational Measurement, 38*, 121-

145. https://doi.org/10.1111/j.1745-3984.2001.tb01119.x

IELTS. (2003). *IELTS annual review 2001/2002*. IELTS Australia.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(20), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kim, Y. S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and writing, 30*(6), 1287-1310. https://doi.org/10.1007/s11145-017-9724-6

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing writing*, *12*(1), 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics, 14*(2), 1-23.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560. https://doi.org/10.1177/0265532211406422

Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2010). *Methods in educational research: From theory to practice* (Vol. 28). John Wiley & Sons.

Lumley, T. (2000). *The process of assessment of writing performance: The rater's perspective.*

[Unpublished Doctoral dissertation]. University of Melbourne.

Lumley, T. (2002). Assessment criteria in large scale writing test: What do they really mean to raters? *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples* [Master's Research thesis, University of Melbourne]. Minerva Access. http://cat.lib.unimelb.edu.au/record=b1849170

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. https://doi.org/10.3102/0013189X023002013

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

O'Sullivan, B., & Rignall, M. (2002). *A longitudinal analysis of the effect of feedback on rater performance on the IELTS General Training writing module*. Cambridge ESOL/The British Council/ IDA Australia: IELTS Research Report.

O'Sullivan, B., & Rignall, M. (2001). *Assessing the value of bias analysis feedback to raters for the IELTS writing module*. Cambridge ESOL/The British Council/ IDA Australia: IELTS Research Report.

Raczynski, K. R., Cohen, A. S., Engelhard Jr, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301-318. https://doi.org/10.1111/jedm.12079

Ragupathi, K., & Lee, A. (2020). Beyond fairness and consistency in grading: The role of rubrics in higher education. In *Diversity and inclusion in global higher education* (pp. 73-95). Palgrave Macmillan. https://doi.org/10.1007/978-981-15-1628-3_3

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language testing essays in honour of Alan Davies* (pp. 82-96*)*. Cambridge University Press.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Rosales-Sánchez, C., Díaz-Cabrera, D., & Hernández-Fernaud, E. (2019). Does effectiveness in performance appraisal improve with rater training? *PloS one*, *14*(9), Article e0222694. https://doi.org/10.1371/journal.pone.0222694

Schoepp, K., Danaher, M., & Kranov, A. A. (2018). An effective rubric norming process. *Practical Assessment, Research, and Evaluation, 23*(1), 1-12.

Shabani, E. A., & Panahi, J. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing Asia, 10(1),* 1-25. https://doi.org/10.1186/s40468-020-00111-4

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes, 8*, 13-17.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Tajeddin, Z., & Alemi, M. (2014). Pragmatic rater training: Does it affect non-native L2 teachers' rating accuracy and bias. *Iranian Journal of Language Testing*, *4*(1), 66-83.

Tziner, A., Joanis, C., & Murphy, K. R. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perception, and ratee satisfaction. *Group & Organization Management*, *25*(2), 175-190. https://doi.org/10.1177/1059601100252005

Wang, J., Engelhard Jr, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36-47. https://doi.org/10.1016/j.asw.2017.03.003

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, *12*(3), 283-304. https://doi.org/10.1080/15434303.2015.1037446

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (1999). Investigating Rater/Prompt Interactions in Writing assessment: Quantitative and Qualitative Approaches. *Assessing Writing, 6*(2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing writing*. Ernst Klett Sprachen. https://doi.org/10.1017/CBO9780511732997

Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language*

Testing, 10(3), 305-335. https://doi.org/10.1177/026553229301000306

Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training context* (White paper). Pearson Assessments.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*, 465-492. https://doi.org/10.1177/0741088398015004002

Xie, Q. (2015). "I must impress the raters!" An investigation of Chinese test-takers' strategies to manage rater impressions. *Assessing Writing*, *25*, 22-37. https://doi.org/10.1016/j.asw.2015.05.001

The Influence of Test Preparation Programs on IELTS Test Performance among Bangladeshi Students' Studying in Malaysia

Ayesha Aktar*, Nooreen Noordin and Lilliati Ismail

Department of Language and Humanities Education, Faculty of Educational Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

Test preparation programs namely coaching, mock tests, and repetitive test-taking aid students in achieving language skills and comprehending the International English Language Testing System (IELTS) test procedure. This paper attempted to find out the influence of test preparation programs on candidates' IELTS test performance. A quantitative method was used, and an online questionnaire survey was conducted to obtain the data. The population of the study was the Bangladeshi students at Universiti Putra Malaysia. Probability sampling, specifically simple random sampling techniques were used to draw the sample. Data were collected from a total of 100 students, and SPSS was employed to analyze the data. The finding showed that mock test ($r = 0.450$), coaching ($r = 0.496$), and repetitive test-taking ($r = 0.369$) have a positive and moderate correlation with IELTS test performance. In contrast, the mock test, coaching, and repetitive test-taking have 23.4%, 35.3%, and 21.3% influence on IELTS test performance, respectively. This study has implications for candidates' teaching and learning opportunities for competitive English language test programs. Furthermore, it will deliver a diverse viewpoint on the preparation programs and show their effectiveness for future reference. This study can be used as a guideline for future research to improve test preparation programs for better test performance.

INTRODUCTION

Test performance, test-taking skills, and test preparation courses are interrelated aspects that influence whether a test taker will perform to the best of his or her ability

in a test (Powers, 2017). As a practice, students attend preparatory programs to achieve standard grades for the test (Saif et al., 2021). Bangladeshi students who have intentions of studying abroad, such programs especially related to IELTS are excitingly popular. Performance in the test means candidates must perform a task or activity rather than only simply answering the questions (Sultana, 2019). Performance may include various capacities and the implementation of skills in a real test (Farooqui, 2020).

The results obtained through test preparation courses will help candidates in their admission into a reputable educational institution. Preparation programs claim that they can raise learners' scores in a test and reduce test-taking anxiety (Paul, 2012). In addition, the International English Language Testing System (IELTS) is deemed one of the most popular standard examinations in measuring a learner's English language proficiency. This test includes separate language skills testing and the learner's knowledge of speaking, listening, writing, and reading (Davis & Vehabovic, 2018), and the results are reported based on a nine-band scale. Due to its popularity as a high-stakes English language proficiency test, IELTS test-takers invest for upgraded performance. Consequently, they attain from test preparation programs, namely coaching, mock test and recently repetitive test taking also included in the test preparation course. Coaching in this context refers to intensive care and receiving private lessons and directions from an expert (Roza, 2019).

The administration mimics the questions and answers of the actual test. Language Assessment experts will mark these mock test papers to help students (Knoch et al., 2020). Repetitive test-taking requires learners to face the real test situation to gain experience to do better and secure good scores (Green & Van Moere, 2020). Most students go through test preparation courses in Bangladesh to prepare themselves for the IELTS test (Kar, 2013).

The test preparation program is believed to guarantee students' success in attaining good results (Powers, 2017). For these reasons, students from Bangladesh tend to choose the IELTS test despite the fact that the standard of English language teaching in Bangladesh is still not up to par (Hamid, 2011). Moreover, English teachers often assess their students' performance without a standard (Ghorbani et al., 2008b; Sato, 2019), encouraging more students to opt for test preparation courses for language skills development.

According to statistics on higher education in Malaysia, more than 28000 Bangladesh students were studying in different universities across the country in 2017, accounting for one out of every four international students studying in Malaysia (Afterschool.my, 2018). Public universities in Malaysia treat English language proficiency with absolute importance, especially in the admission process, where candidates' four language skills are measured before approving admission (Samad et al., 2008). The current study uses the concept of language management theory where its main features are discussed

in relation to the IELTS test performance of Bangladeshi students. Thus, the aims of this study are to find out the influence of Mock test, Coaching and Repetitive test taking on candidates' performance on their IELTS exam performance.

Background of the Study

Recent studies from various parts of the world suggest that more students have engaged in test preparation courses, and Bangladeshi is no different (Sultana, 2019). Test preparation courses, especially for IELTS tests, are becoming very common (Kabir, 2018). Learners attended these courses without hesitation or thinking about how far they could benefit from them in the long run (Sultana, 2018). Furthermore, the scores obtained from students' examinations are considered crucial in demonstrating their capacity as a learner as well as a platform to determine their future (Zakaria et al., 2013). Traditional evaluations provide one-off and indirect experiments with little input for students, as well as test assignments that are decontextualized. This indicates that traditional teaching is not sufficient for learners' needs, encouraging test preparation programs (Singh & Samad, 2013).

Teachers' knowledge influenced the exam preparation courses. Generally, students want more feedback from teachers, who also encourage student to take these programs (Ghorbani et al., 2008a). There have been many contrary opinions about the test preparation course for IELTS, and yet, not many researches have been carried out in this sector to clarify that the test preparation course can help learners to do better in the real IELTS test (Kar, 2013). Tests like IELTS are frequently thought of as anxiety-provoking and can create pressure on students as learners do not know exactly how the tests are carried out (Hu & Trenkic, 2019). In Bangladesh, for countless students, a test preparation program is the ultimate solution to their all troubles. Many courses favor test items over profundity, and students may be left with only a speedy summary of terms, topics, and theories without sufficient time and proper design materials. Therefore, it is necessary to find out the actual influence of test preparation courses on Bangladeshi's students' performance in their IELTS test. The problems and controversies of the test preparation course for the IELTS test will remain if it is not dealt with and could result in learners not being familiar with the preparation courses for IELTS and the actual test itself. Unfortunately, no such elaborate study has been conducted in Bangladesh for the investigation of such issues. Therefore, it created a gap in knowledge that all concepts and issues regarding coaching, repetitive test-taking, and mock tests found in previous research are the same in the Bangladeshi context or different from the Bangladeshi context or vice versa.

LITERATURE REVIEW

IELTS

IELTS is jointly owned by the British Council, IDP: IELTS Australia, and Cambridge Assessment English (IELTS, 2021) and is considered a high stakes test

that has been administered in over 135 countries, and the scores are accepted by over 7,000 educational institutions. This test is designed to evaluate students' English language skills critically. This test is famous for immigration, study and work accreditation and is particularly known in almost all educational sectors (Arcuino, 2013).

Test Performance

Test performance entails performing well in the examination, recognizing the test items, associating the items with the correct response, and managing the entire procedure. In an attempt to investigate the relationship between test performance and language proficiency, researchers discovered that the low performance was not a general pattern, but rather one that revealed a loose verbal–nonverbal contradiction in which bilingual persons' performance was noticeably lower than test means that were similar in the previous kind but closer to or at the mean on tests categorized in the latter (Sotelo-Dynega et al., 2013). However, there may not seem to be any scientific data to suggest that such ratings alone are fundamentally accurate estimates of true abilities or that they better predict academic success than those that involve linguistically difficult activities (Lohman et al., 2008). Moreover, the absence of scientific evidence for categorizing bilingual people's results as a straightforward dichotomy shows that the association between language competence and test scores is more complex than previously assumed.

Research on the Preparation Course for IELTS

It has been suggested that Bangladesh needs to revise its policies and emphasize more on tertiary level English language proficiency in developing good communicative skills among its workforce (Rahman & Pandian, 2018). However, the efficacy of the test preparedness course has not been discussed at length, considering its prevalence in test preparation and its accuracy in assessing test-takers' English language skills. Power's (1993) meta-analysis of coaching impact on SAT scores found only questionable proof of arguments made by coaching companies and publishers of test content.

Instruction in test preparation connotes test-centric teaching, which includes an unfair and restrictive collection of activities that limit students' reading learning opportunities (Davis & Vehabovic, 2018). Because of this, test value and test anxiety has varying degrees of impact on students' confidence and test performance (Chou, 2019). Brown (1998) studied the influence of test tag findings of the course in preparing IELTS students and found that the IELTS training course is more effective than the EAP in preparing students for the original test. Read and Hayes (2003) conducted a study on IELTS classes in two Auckland language schools where courses A offered intensive teacher training for students, while Courses B paid attention to language skills. Nevertheless, Read and Hayes' analysis did not indicate any correlation between the different IELTS preparatory styles/formats and IELTS examiners' actual results.

The connection between intensive research in English for academic purposes and the IELTS ranking was also explored by Elder and O'Loughlin (2003). A total of 112 non-English students participated in the study, which utilized comprehensive English courses provided by one of Australia's four independent language centers and New Zealand's four independent language centers. Their findings show that it is easier to move from one stage to the last of the IELTS scale for both the overall score and the individual sub-skills. Regarding the impact of instructional variables, Elder and O'Loughlin's study found that the quality of instruction was a key predictor of the success of a group of variables. The above studies have been conducted in Australia or New Zealand, where the IELTS university examination is mandatory for immigrants from Asian countries, including Japan, China, and Korea. Thus, the previous studies found that preparation courses are significant for test performance and may need more investigation.

Test Preparation Influences on Test Performance

In language assessment, studies on test preparation centered mainly on the impact of washback, particularly examining learning and teaching (Xie & Andrews, 2013; Xie, 2013). However, several studies have investigated the extent to which test preparation improves test scores. Xie and Andrews (2013) also found a higher degree, though limited in influence, of the training of 807 Chinese students in College Test Band 4 (CET4).

They defined the types of test preparation that contributed to scores: test management, memorization, and drill practice learning, which most likely also enhanced the test-takers' knowledge. The influence of test-coaching on test scores was reported by Farnsworth (2013) for oral competence assessments. Farnsworth noted that coaching is another word for an intensive test preparation program where students practice similar question formats. ESL students took two separate oral tests (Basic English Skills Plus and Versant English Test) for pre-testing and randomized coaching sessions related to each of the two assessments listed in the Farnsworth report. Both examinations were taken as post-tests. Both participants improved their grades during the post-test, regardless of the test they had qualified for; however, the benefit was greater when contrasted with coaching. To analyze the relationship between test preparation and TOEFL iBT outcomes, Liu (2014) performed large-scale regression analyses with 14,593 Chinese test participants performing online survey questionnaires.

According to the findings, TOEFL iBT scores were also predicted by the school attendance of the test participants; however, the author stressed that its contribution to the test score was too small to endorse the impact of school attendance coaching (an increase of 1.86 points out of 120). The vocabulary and the simulation training substantially predicted the coaching program's total scores and subsection scores, particularly the simulation courses for reading and hearing. It helped to improve the reading

and listening test scores total points by more than 1 in 30. Likewise, substantial predictors of total test scores and subsection scores were formed among test-specific strategies, vocabulary memorization, and practice research. In general, the relationship between the implementation of the technique and the test results was unique to the field under study. For example, listening techniques, like reading sketch strategies, were more likely to contribute to listening ratings. As mentioned before, the test developers tried to make the tests resistant to these coaching influences (Spaan, 2007). Test planning has seen a significant uptick in high-stakes testing situations. Under these situations, the key goal of test planning exercises is to improve student grades (Gebril & Eid, 2017).

However, the actual influences of the preparation of the exam may need to be further investigated. From Xie and Andrew's (2013) findings, it is hard to see how much real language acquisition has occured alongside test planning. Although official English instruction was suspended outside the test preparation, there was no real monitoring group in the study, which would have shown the magnitude of expected improvements over the 10-week duration without preparation for the test. In addition, Farnsworth's study is limited in its scale and thus needs replication (15 or 19 participants in both oral test groups). Since the study concentrated on detailed questionnaires online, Liu's (2014) research is also not straightforward. Such aggregated data are likely to blur many details on an individual basis. Based on the findings, it would be

appropriate and timely to investigate if the same findings apply to IELTS test runners, another high-stakes academic language test for ESL students (Tulloh & Wood, 1998). The only helpful thing test-centric instructional methods may provide students is increased knowledge of how tests are written. Test-writers employ vocabulary and forms that some pupils are not familiar with.

Hence, teaching students how to feel at ease with the language of the exam which enables them to show their understanding rather than attempt to decipher the author's code, might prove valuable and beneficial (Davis & Vehabovic, 2018; Winke & Lim, 2017).

Mock Tests

A study conducted in Iran has shown that mock tests can positively influence the training of IELTS candidates in the context of EFL. Erfani (2012) discovered that the ability to complete exams such as IELTS and TOEFL encouraged both language teachers and learners to use them as training tests in the classroom. Test-based language classes, however, have not always been preferred. The use of tests in language classes and the practice of testing techniques were indicators of a negative washback influence, as argued by Watanabe (2004). Moreover, the results of this research have shown that the practice of technique testing is more effective than teaching the content of courses in terms of high stakes tests, such as IELTS. However, a more comprehensive study is needed to determine the magnitude of the washback influence in these circumstances. The

washback influence may be due to Iranian EFL learners not having enough practice in strategies to take the test, as mentioned by Mohammadi (2016), where it was concluded that candidates' time is wasted on the exam.

In their study, Yang and Badger (2015) also concluded that IELTS training courses that assess students provide them with a sense of protection, as most students want to learn how to do well in the assessment and get high scores. In test preparation courses, the authors, therefore, promote the use of mock evaluations. Lumley and Stoneman (2000) concluded that the tests were central to the preparation of IELTS courses and had a more positive influence than focusing on the linguistic features of the language in those courses. This study's findings align with Lumley and Stoneman's (2000) study concluded that real examination practice had a greater influence than the IELTS training courses. Naseri et al. (2014) noted that practicing speed reading strategies with IELTS candidates will enhance their understanding of reading in the test. As stated by Chung and Nation (2006), speed reading techniques include skimming and scanning techniques. Bell (2001) also mentioned that such techniques are better developed in a test situation, increasing the test-taker reading comprehension and scores. The above discussion indicates that a mock test has a constructive influence on test performance.

Coaching

Specialized exam preparation services are also known as test coaching for students who need to pass high-level examinations. Using several different mechanisms, they can improve test scores (Kabir, 2018). Any of these do not make the legitimacy of ratings a challenge. For example, by removing irrelevant variables, practices that familiarize students with the test format will decrease anxiety and thus increase validity. By helping to build the underlying skills of the exam, other operations can increase scores (Koh et al., 2018). However, some test preparation practices can compromise the validity of the scores. These activities improve the score by limiting the program to focus exclusively on the subject and the type of questions that can be measured (Powers, 2017). Activities of this kind can improve scores without correspondingly enhancing the underlying ability, sacrificing the validity (strength of inference from tested to untested behavior) of score extrapolation.

Other operations can boost the scores by helping to build underlying abilities that assessments offer (Arendasy et al., 2016). These activities raise the scores by restricting the curriculum to only concentrating on the contents and types of test questions. Activities such as this can improve scores without improving the underlying skills, thereby compromising the validity of scoring extrapolation. In Green's (2007) report, the related consequences of curriculum reduction strategies were observed based on the IELTS writing scores obtained from 476 international students in the UK.

Although no characteristic distinctions have been made between the different types of programs (Davis & Vehabovic, 2018),

the IELTS class behaviors were the only course parameter that was uniquely positive. Based on what was observed in Green's (2007) study, these are precisely the types of practices that have been adopted by the booming English test prep industry (Matoush & Fu, 2012). Test centers concentrate on routine practice using parallel tests to prepare applicants for a 'probability game' (Ma & Cheng, 2015).

The reliability of IELTS results is further strengthened by a study conducted on 45 Chinese students. The candidates' progress with various other activities is indexed online, including checking in Oxford Dictionary, vocabulary assignment, and written sentences. Besides, their performance on these tests was very similar to the community of 44 subjects in the controlled group, both before and after the exercise. At that time, they had no interest in any training. After the IELTS evaluation, competence measurement for the test-prep group exceeded the control group. The results obtained showed that IELTS can increase approximately half the band scoring with this adjustment, as other English proficiency tests introduce a 4-week coaching program without this change. The research showed that certain test preparation activities may undermine the validity of the extrapolation of scores and that previous concerns about the ability of the test preparation courses to achieve the required return may not have been taken into account (Green, 2007). The more intensive accomplishment of the test coaching industry depends on their effective planning

and teaching. In a nutshell, coaching can provide the learners the facilities to improve their test performance.

Repetitive Test-Taking

Listening is an important skill in effective communication. In Bangladesh, most of the students belong to Bangla medium education, which causes problems among students to understand English medium instruction (Abedin et al., 2009). For this reason, to achieve good scores in the IELTS test, students take repetitive tests to prepare for the IELTS. The required language score for an unconditional first-time university offer is not attained by most international students (Li, 2013). Therefore, the majority of the candidates would usually have to repeat the test at least once (Sultana, 2018). IELTS advertises on its website that test repeaters are unlikely to raise their test scores further if they do not strengthen their English language skills (Green, 2007).

The IELTS results can be confirmed and accessed within 90 days. However, in 2006, the legislation was repealed, forcing applicants to take a test and repeat until adequate changes were made in their skills to be worth a higher rating. For this reason, students need to repeat the IELTS test, and there have been reports indicating that some students had to repeat the test 14 times in eight months, with three attempts in one month to develop the language skills and scores (Hamid, 2016).

While the major changes in language testing alone are not likely to occur, at least some evidence is that small improvement are

likely to exist. For example, Zhang (2008) found evidence of minor but consistent score improvements (effect sizes varying from 0.12 to 0.17 SD for test components and 0.17 SD for the test as a whole) by examining the outcomes of about 12,000 candidates who repeated the TOEFL within one month. However, it is doubtful that the improvements would have resulted from improves proficiency, considering the limited time the test was replicated (Sato, 2019).

However, repetitions may have resulted from increased knowledge of the test format (Koh et al., 2018). Although it is unlikely that major changes in language tests alone can be due to repeated tests, there is at least some proof that small improvements may be made. Overview of the latest review by Hu & Trenkić (2019) observed that extensive IELTS preliminary programs, immediately tested after the intervention, would increase IELTS scores by about 0.5 bands with the corresponding increase in alternative skill measures. This report recruited Chinese students at the UK University to investigate how the IELTS practice of test preparation impacts students' skills when arriving and how the IELTS ranking predicts academic performance. Three investigation questions have been answered in the report, which shows that IELTS preliminary programs increase the scores. Previous researchers consistently found a positive link between repetitive test-taking and test performance.

Students generated their research schedules and time frames, which necessitated purchasing test planning guides, question banks, and commercial courses on

their own time and expense, which helped them improve test performance (Schwartz et al., 2018). Academic success gaps may be attributed to a variety of external causes, including academic readiness measures. In addition, when readiness (e.g., SAT or high-school grade-point average) is considered, student characteristics may forecast accomplishment in certain disciplines and preparation for an exam (Salehi et al., 2019). Previous research has shown that a goal-oriented strategy is correlated with lower test anxiety. Students who were goal-oriented in their research and planned their learning were less stressed (Yusefzadeh et al., 2019).

It is only normal that the world's most widely used English exam is critical for students who often participate in test-prep courses or pursue tutoring to obtain their desired results (Minakova, 2020). All types of coaching, mock tests, and repetitive test-taking are common for learners' study careers. However, no definite research shows these three variables together to compare with each item and determine the effective application of these variables. These issues are never addressed properly in Bangladesh, which motivates this study.

From the literature review, it can be said that all the studies mentioned are based on only the authors' context. So far, no prominent research has been conducted on Bangladeshi students, which created a research gap in the Bangladeshi context. Moreover, whether these types of programs are beneficial or have no impact on students' learning is unclear or received less attention. Therefore, it

is essential to study the influence of the test preparation program on Bangladeshi students' IELTS test performance. The findings from the current study are hoped to address the pertinent issues and add to the body of knowledge regarding IELTS test preparation.

Based on the study, it is evident that coaching, repetitive test-taking, and mock tests help learners improve their test performance. These preparatory programs are the instrument to develop learners' language skills. It is, therefore, hypothesized that

> H1: There is a positive influence of mock tests on candidates' test performance on the IELTS exam
>
> H2: Coaching has a positive influence on candidates' test performance on the IELTS exam
>
> H3: There is a positive influence of repetitive test-taking on candidates' test performance on the IELTS exam

Theoretical Details of the Study

Language Management Theory. Language management theory (LMT) explains that the initial stage in language planning addresses the problems concerning the context. Then, the activities must be carried out to solve all the problems and suggest completing the planning process (Neustupný, 1994). According to the theory, if hindrances persist, then language implementation is not possible. Therefore, learners' test performance will not be satisfactory if prior preparation is not taken on aspects such as practicing speed reading (Chung & Nation

2006), good language proficiency (Abedin et al., 2009), and different mechanisms of the test (Kabir, 2018). The first step of simple management occurs when an individual notes something in her or his own or the interlocutor's utterance. Then, the speaker evaluates the phenomenon (Nekvapil, 2015), and if the phenomenon is evaluated negatively, it is referred to as inadequacy in LMT. If the phenomenon is evaluated positively, it is referred to as gratification in LMT (Kimura, 2014).

It will progress to the next step, known as adjustment design, in which the speaker, for example, begins to consider rephrasing her/his utterance. Even at this stage, the procedure will stop or move on to a new step in which the speaker applies or implements the proposed change design in the context of the current discussion (Marriott & Nekvapil, 2012); management will become cyclic (Lanstyák, 2014).

Using the Language management theory (LTM), coaching trainers, mock tests, and repetitive test-takers will first take notes on candidates' IELTS-related problems. Then they will evaluate the entire problem in order to overcome these problems. A planning or adjustment process will be followed by trainers implementing the plan to resolve IELTS exam-related issues. In the last stage, learners will give feedback on their test performance for further evaluation to check their performance.

METHODOLOGY

A quantitative research method was used to conduct this study. Data were collected

via an online survey questionnaire. The research employed the probability sampling method, in which a simple random sampling technique was used to select the sample. Universiti Putra Malaysia (UPM) was selected for this study because UPM is one of the leading research universities that offer undergraduate and postgraduate courses focusing on agriculture and other related fields, including English language teaching and learning. The total population of participants for this study was 140 students who are either studying for their Bachelor, Masters or Ph.D. in UPM, and previously they attended IELTS test preparation programs. As the population is 140, the expected sample size for the study will be a minimum of 100, according to Kish (1995), while the confidence level is 95% and the margin of error is 5%. Therefore, based on their homogeneity test results and their performance on IELTS, 100 Bangladeshi learners were selected as the sample size for this study. The usage questionnaire for the dependent variable named "test performance" was adopted from Zhengdong (2009). The independent variables titled "coaching," "mock test," and "repetitive test-taking" were adapted from Farnsworth (2013), Khodabakhshzadeh & Zardkanloo (2017), and Hu & Trenkic (2019), respectively. The relationship among the selected variables was investigated by Zhengdong (2009) through the survey questionnaire, and the question was regarding the influence of test preparation programs on the learners' test performance. The items in this study are based on the 5-Likert scale, ranging from 1=strongly disagree to 5=strongly agree. Descriptive and inferential statistics are employed to analyze the data, and scores obtained are tabulated using SPSS version 26.

DATA ANALYSIS

Demographic Profile

The demographic profile consists of respondents' age, gender, and level of study. Most of the respondents (48%) were between the ages of 20–30 years old, 29% were between 31–40 years old, and 23% were between 41–50 years old. Regarding gender, 37% of respondents were male, while 63% of respondents were female. Most of the respondents (73%) studied for their bachelor's degree, 16% studied for their master's degree, while the remaining respondents (11%) were Ph.D. or Doctor of Philosophy students.

Correlational Analysis

Correlation is a method to investigate the relationship between two variables linearly. Pearson correlation was adopted to examine the association between two variables. Thus, the change in one variable eventually will lead to a change in another variable. The Pearson coefficient (r) stated the direction, magnitude, and significance of the correlation. A perfect positive relationship occurs if the r value is +1.0, and a perfect negative relationship if r is -1. The result, as shown in Table 1, illustrates that there was a moderate positive correlation between

Table 1
Pearson correlations coefficient between variables

	(Y1)	(Y2)	(Y3)	(Y4)
Performance (Y1)	1	0.450**	0.496**	0.369**
Mock test (Y2)	0.450**	1	0.411**	0.335**
Coaching (Y3)	0.496**	0.411**	1	0.221**
Repetitive test taking (Y4)	0.369**	0.335**	0.221**	1

**. Correlation is significant at the 0.01 level (2-tailed).

performance and repetitive test-taking (r = 0.369), performance and coaching (r = 0.496), and performance and mock test (r = 0.450). If the mock test, coaching, and repetitive test-taking were increased, the IELTS performance would increase moderately. Additionally, if the mock test, coaching, and repetitive test-taking decreased, the IELTS performance would also decrease moderately because of the moderate relationship among the variables.

Multiple Regression Analysis

Multiple regression analyzes whether a mock test, coaching, and repetitive test-taking explain the IELTS test performance. Researchers have conducted multiple regression analysis because (1) the relationship between independent variables (mock test, coaching, and repetitive test-taking) and dependent variable (IELTS examination performance) are linear, (2) there is no multicollinearity in data, As the VIF is below ten and the tolerance is higher than 0.1, as shown in Table 4, (3) the values of the residuals are independents and normally distributed and (4) the variance of residuals are constant.

The values presented in Table 2 of the regression coefficient (r²) is 0.359

(0.359x100= 35.9%), which indicates the degree of variance in the performance is explained by the repetitive test-taking, coaching, and mock test. It also means that repetitive test-taking, coaching, and mock test explains 35.9% of the variance in performance.

Table 2
Model summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.599[a]	0.359	0.349	0.36143

a. Predictors: (Constant), Repetitive test-taking, Coaching, Mock test

With F = 38.787 at 99 degrees of freedom, the test is very significant. Hence, it can be concluded that there is a relationship between predictors and dependent variables in the model (Table 3).

Pallant (2005) details out that the greater value of beta and less value of significance level (p<.05) of independent variables will show the contribution to the dependent variable. For example, the standardized coefficient (Beta) value is 0.234, 0.353, and 0.213 for a mock test, coaching, and repetitive test-taking, respectively, while p is 0.000, as shown in Table 4. Therefore, the mock test (23.4%), coaching (35.3%),

and repetitive test-taking (21.3%) have a significant influence on performance. If the mock test, coaching, and repetitive test taking have been increased one unit, the IELTS test performance will increase 23.4%, 35.3%, and 21.3%, respectively.

Table 3
Result of ANOVA[a]

	Sum of Squares	df	Mean Square	F	Sig.
Regression	15.200	3	5.067	38.787	0.000[b]
Residual	27.171	96	0.131		
Total	42.372	99			

a. Dependent Variable: Performance
b. Predictors: (Constant), Repetitive test-taking, Coaching, Mock test

Table 4
Coefficients [a]

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	1.416	0.253		5.604	0.000		
Mock test	0.220	0.060	0.234	3.692	0.000	0.769	1.301
Coaching	0.274	0.048	0.353	5.765	0.000	0.823	1.214
Repetitive test-taking	0.177	0.049	0.213	3.593	0.000	0.879	1.137

a. Dependent Variable: Performance

DISCUSSION

The purpose of the study was to find out whether mock tests, coaching, and repetitive IELTS exam-taking influence an IELTS candidate's overall performance. The findings reflect the importance of addressing test preparation programs' issues as thousands of test-takers pay considerably high fees for the preparation programs every year.

The study's first hypothesis was a positive influence of mock tests on candidates' test performance on the IELTS exam. This study's results show that mock test programs are vital to providing students with effective approaches to prepare for the actual exam and guidance on strategically managing their time in the actual test. Finally, the mock test enabled candidates to be competent in recognizing test items and answering them correctly. Thus, hypothesis H1 is accepted. This finding is also similar to the previous research conducted by Erfani (2012), Naseri et al. (2014) as well Mohammadi (2016).

The second hypothesis of this study was that coaching positively influences candidates' test performance on the IELTS

exam. The finding indicates that coaching has a significant relationship with the real exam scores and candidate performance. Test preparation coaching programs help students to perform well in real test. The result in this study shows that after attaining the test preparation coaching program, candidate's language skills are enhanced to produce better answers in real exams as well as help to understand the real test mechanism which further assists to scores better. It is able to enhance language skills and obtain better scores in the real exam. After coaching listening, speaking, writing, and reading, test-takers language skills in English were enhanced to produce better answers in real exams which altogether increase the IELTS scores and performance. Therefore, hypothesis H2 is accepted. This finding is also similar to the findings found in Brown (1998), Elder and O'Loughlin (2003), Xie and Andrews (2013), and Spaan's (2007) studies. Students' strong motivation to pass the IELTS combined with professional and dedicated instructors, a teaching emphasis on both improving English language learning and targeted programs, and students' positive mindset to pass the IELTS, could lead to a good program in these circumstances, with possible beneficial results for students' learning outcomes (e.g., increased English language proficiency) (Saif et al., 2021; Schwartz et al., 2018).

The third hypothesis regarding the positive influence of repetitive test-taking on candidates' tests is performed on the IELTS exam. The findings demonstrate that repetitive test-taking has a positive influence on candidates' test performance. The result indicates that repetitive tests taking increases the performance positively in the exam. Therefore, the third hypothesis is accepted. This research finding corresponded to the previous findings found in Li (2013) and Zhang's (2008) studies. Repetitive test-taking increased candidates' test scores and reduced their test anxiety. Repeat testing is an established technique for increasing memory and retention (Roediger III & Butler, 2011). When knowledge of the degree to which output is satisfactory is provided, the ability to carry out test tasks consistently is maximized, or experience and expertise are increased (Boyd et al., 2019). On the other hand, if teachers and students use repetitive tests to study, it may remove the limitation on what is covered and learned to improve the test performance (Green, 2019).

It can be concluded that, in general, all participants agreed that coaching, mock tests, and repetitive test-taking had positively increased their performance in the test, which is an indication of test preparation programs' significant influence on test performance. The findings show that mock tests have a moderate (23.4%) significant influence, coaching has a moderate (35.3%) significant influence, which is generally a good influencer, and repetitive test-taking has a low (21.3%) significant influence on test performance, indicating that more research should be done to find out issues related to repetitive test-taking. All three variables have positive impact but coaching

has much more influence on students' test performance. In other words, coaching should be emphasized more and carefully designed as a preparation tool for test-takers.

This study's findings can be applied in the test preparation programs industry to guide them to choose a correct program according to the candidates' needs. Students will concentrate on tasks that can contribute to their target if they divide their training/ studying into levels. Using a range of study methods can help students improve their abilities. Discussions with tutors, peer use of the student resource of taking revision documents, reviewing examiners' papers, and practicing exam-type questions are only a few examples of test preparation aids (O'Sullivan et al., 2021).

Since this form of private, supplemental tutoring imitates the mainstream test method in that the tutoring offered is matched to the exam being prepared for, the metaphor of "shadow" is used. North America, Western Europe, Asia, and Africa, according to a report conducted by the Sutton Trust in the United Kingdom (Kirby, 2016), are the main markets for private tutoring/ coaching, but for various purposes and with different economic scopes. This study, with its finding, also extended the previous researches.

As a researcher from the same field, the outcome of this study may suggest that students learn and practice in a stress-free setting where they may do things like taking tests and receive individual help with homework, including a step-by-step outline for readability. It is all done

with them being guided by professionals. Students will develop confidence heading into the test after taking a lesson. If an exam question has emerged that they are unprepared for, students might resort to what they have learned from their supplemental classes. They can use it to approach the term, unraveling it until they reach the ideal description. It indicates that the researcher found new information regarding test preparation programs. It will benefit both the researcher and other researchers, teachers, students, and general people to acquire knowledge to organize future programs.

This research will also be beneficial for researchers in other fields because it will uncover various aspects of test preparation programs that will enrich their knowledge and encourage them to explore new areas of interest for the elaboration of learning and development.

This research also extended the findings of previous research with its investigated topics. Most previous research was conducted on general issues or any one of the varieties of preparation programs. Nevertheless, this study specifically addressed three items: coaching, mock test, and repetitive test-taking, to examine the influence of these items on the candidates' test performance in the IELTS exam.

The results found by Zhengdong (2009) show that lower-scoring students are more likely to attend the IELTS test preparation course. Zhengdong (2009) investigated exam scores and students' university learning experiences. However, this study expanded on previous findings by focusing

on three specific items related to actual test performance. Conversely, the research by Farnsworth (2013), Khodabakhshzadeh & Zardkanloo (2017), and Hu & Trenkic (2019), show that these test preparation programs have a future influence on students' learning. Differences in study settings exist because test preparation is still up for debate.

Watanabe (2004) analyzed several factors relating to this topic, such as negative and positive influences. Nonetheless, this research tries to determine if there are any benefits students could harvest from these programs. The results indicate that these programs are useful in the Bangladeshi context and help students prepare for the challenge. Furthermore, it enriches previous literature with new information in different regions and nations and uncovers the opportunity to conduct further research for evidence to enhance educational standards.

LIMITATIONS AND FUTURE RESEARCH

There are certain limitations in this report. For example, firstly, only one Malaysian university was chosen to collect data for this research. Therefore, the findings could not apply to the other foreign community of Malaysian universities. It is suggested that future studies involve several universities around Malaysia, thus increasing the sample size and amount of responses received. Second, the predictor constructs included in this analysis are not comprehensive; future studies should provide more specific predictors, such as test anxiety and students'

English language proficiency, to improve the research's predictive capacity. Finally, only Bangladeshi students studying in Malaysia were selected for this study. Students from other countries studying in Malaysia could be included in future research to create a more detailed paper.

CONCLUSION

This study shows that test preparation programs that include coaching, mock tests, and repetitive test-taking significantly influence candidates' test performance in a test. The findings from this study clearly shows that mock tests in IELTS test preparation programs have a positive influence on Bangladeshi students. Additionally, the findings also indicate that the coaching program for the IELTS test has a positive significant influence on Bangladeshi candidates' test performance on the IELTS test, which corroborates with findings from previous studies. Moreover, repetitive test-taking also has an affirmative influence on the candidates' IELTS test performance. Since English is essential as the medium of instruction in today's university education, appropriate language testing is crucial for assessing candidates' language abilities (Samad, 2019). Traditionally many institutions conduct language tests on a large scale. However, teachers are rarely involved in communicative-based instruction and do not teach communicative language test courses based on the proper structure. As a result, it is difficult for teachers to determine

the pathways to include the subject or compellingly present the test material, providing more learning opportunities.

The learner's attention may focus entirely on the study and mastery of the types of test subjects, thus neglecting the primary purpose of language learning. Furthermore, materials for preparing for the test, which contains different parts of the exams, are often meaninglessly taught, and the exam is based on scores. Thus, it is characterized not by the development of knowledge and analytical thought but by rote learning. Despite the discrepancies in making conclusions concerning the test preparation programs and their influence on test-takers performance, it is undeniable that these programs have a statistically significant influence on students' test scores (Singh et al., 2015).

Hence, in an ongoing effort to better understand the implemented test preparation practices and disseminate those practices that produce the best results, more focus should be given to content and test-taking strategies. From the findings obtained in this study, it is clear that integrating content review and subject-specific test preparation activities into the programs is expected to improve learners' test scores. The test preparation programs can provide test-takers with various benefits beyond the learning and testing environment. They can be considered a valuable studying technique that does not only help ease anxiety but also in brushing up on areas in preparation for language tests.

ACKNOWLEDGEMENT

REFERENCES

Abedin, M. M., Majlish, S. H. K., & Akter, S. (2009). Listening skill at tertiary level: A reflection. *Dhaka University Journal of Linguistics, 2*(3), 69-90. https://doi.org/10.3329/dujl.v2i3.4144

Afterschool.my. (2018). Here's why Bangladeshi students choose to study in Malaysia. Retrieved May 2, 2021, from https://afterschool.my/international-students-why-study-inmalaysia/here-s-why-bangladeshi-students-choose-to-study-in-malaysia

Arcuino, C. L. T. (2013). *The relationship between the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) scores and academic success of international Master's students* [Doctoral dissertation, Colorado State University]. Mountain Scholar. http://hdl.handle.net/10217/78818

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence*, *55*, 44-56. https://doi.org/10.1016/j.intell.2016.01.004

Bell, T. (2001). Extensive reading: Speed and comprehension. *The Reading Matrix*, *1*(1), 1-13.

Boyd, E., Green, A. B., Hopfenbeck, T., & Stobart, G (2019). *Effective feedback: The key to successful*

assessment for learning. Oxford University Press.

Brown, J. D. H. (1998). An investigation into approaches to IELTS preparation, with particular focus on the academic writing component of the test. *IELTS Research Reports, 1*, 20-37.

Chou, M. H. (2019). Predicting self-efficacy in test preparation: Gender, value, anxiety, test performance, and strategies. *The Journal of Educational Research, 112*(1), 61-71. https://doi.org/10.1080/00220671.2018.1437530

Chung, M., & Nation, I. S. P. (2006). The effect of a speed-reading course. *English Teaching, 61*(4), 181-204.

Davis, D. S., & Vehabovic, N. (2018). The dangers of test preparation: What students learn (and don't learn) about reading comprehension from test-centric literacy instruction. *The Reading Teacher, 71*(5), 579-588. https://doi.org/10.1002/trtr.1641

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *IELTS research reports, 4*(6), 207-254.

Erfani, S. S. (2012). A comparative wash back study of IELTS and TOEFL iBT on teaching and learning activities in preparation courses in the Iranian context. *English Language Teaching, 5*(8), 185-195. https://doi.org/10.5539/elt.v5n8p185

Farnsworth, T. (2013). Effects of targeted test preparation on scores of two tests of oral English as a second language. *TESOL Quarterly, 47*(1), 148-155. https://doi.org/10.1002/tesq.75

Farooqui, S. (2020). The methods of learning English language: A critical evaluation of test-focused teaching in Bangladesh. In *The Routledge Handbook of English Language Education in Bangladesh* (pp. 130-148). Routledge. https://doi.org/10.4324/9780429356803-9

Gebril, A., & Eid, M. (2017). Test preparation beliefs and practices in a high-stakes context: A teacher's perspective. *Language Assessment Quarterly, 14*(4), 360-379. https://doi.org/10.1080/15434303.2017.1353607

Ghorbani, M. R., Samad, A. A., & Gani, M. S. (2008). The washback impact of the Iranian university entrance examination on pre-university English teachers. *Journal of Asia TEFL, 5*(3), 1-29.

Ghorbani, M. R., Samad, A. A., Hamzah, M. S. G., & Noordin, N. (2008). All that glitter is not gold: Curriculum alignment and improving students' test scores. *Iranian Journal of Language Studies, 2*(1), 19-40.

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education, 14*(1), 75-97. https://doi.org/10.1080/09695940701272880

Green, A. (2019). Washback in language assessment. In C. A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 1159-1193). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1274.pub2

Green, A., & Van Moere, A. (2020). Repeated test-taking and longitudinal test score analysis. *Language Testing, 37*(4), 475-481. https://doi.org/10.1177/0265532220934202

Hamid, M. O. (2011). Planning for failure: English and language policy and planning in Bangladesh. In J. A. Fishman & O. Garcia (Eds.), *Handbook of language and ethnic identity* (Vol. 2, pp. 192-203). Oxford University Press.

Hamid, M. O. (2016). Policies of global English tests: Test-takers' perspectives on the IELTS retake policy. *Discourse: Studies in the Cultural Politics of Education, 37,* 472-487. https://doi.org/10.1080/01596306.2015.1061978

Hu, R., & Trenkic, D. (2019). The effects of coaching and repeated test-taking on Chinese candidates'

IELTS scores, their English proficiency, and subsequent academic achievement. *International Journal of Bilingual Education and Bilingualism, 24*(10), 1-16. https://doi.org/10.1080/13670050.2019.1691498

IELTS. (2021). *IELTS is the International English Language Testing System, the world's proven English language test.* Retrieved August 2, 2021, from https://www.ielts.org/info-pages/about-us

Kabir, S. A. (2018). IELTS writing test: Improving cardinal test criteria for the Bangladeshi context. *Journal of NELTA, 23*(1-2), 76-89. https://doi.org/10.3126/nelta.v23i1-2.23352

Kar, A. (2013, May 04). IELTS preparation. *The Daily Star.* https://www.thedailystar.net/news/ieltspreparation.

Khodabakhshzadeh, H., & Zardkanloo, R. (2017). The effect of mock tests on Iranian EFL learners' test scores. *International Journal of Education and Literacy Studies, 5*(3), 47-51. https://doi.org/10.7575/aiac.ijels.v.5n.3p.47

Kimura, G. C. (2014). Language management as a cyclical process: A case study on prohibiting Sorbian in the workplace. *Slovo a slovesnost, 75*(4), 225-270.

Kirby, P. (2016). *Shadow schooling: Private tuition and social mobility in the UK.* The Sutton Trust.

Kish, L. (1995). *Survey sampling.* Wiley-Interscience.

Knoch, U., Huisman, A., Elder, C., Kong, X., & McKenna, A. (2020). Drawing on repeat test takers to study test preparation practices and their links to score gains. *Language Testing, 37*(4), 550-572. https://doi.org/10.1177/0265532220927407

Koh, K., Burke, L. E. C. A., Luke, A., Gong, W., & Tan, C. (2018). Developing the assessment literacy of teachers in Chinese language classrooms: A focus on assessment task design. *Language teaching research, 22*(3), 264-288. https://doi.org/10.1177/1362168816684366

Lanstyák, I. (2014). On the process of language problem management. *Slovo a slovesnost, 75*(4), 325-351.

Li, L. (2013). Chinese candidates need to sit 3 IELTS before studying abroad and 60% of them are expected to achieve an overall of at least 7. http://learning.sohu.com/20130209/n365924751.shtml.

Liu, O. L. (2014). Investigating the relationship between test preparation and TOEFL iBT® performance. *ETS Research Report Series, 2014*(2), 1-13. https://doi.org/10.1002/ets2.12016

Lohman, D. F., Korb, K., & Lakin, J. (2008). Identifying academically gifted English language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. *Gifted Child Quarterly, 52*, 275-296. https://doi.org/10.1177/0016986208321808

Lumley, T., & Stoneman, B. (2000). Conflicting perspectives on the role of test preparation in relation to learning? *Hong Kong Journal of Applied Linguistics, 5*(1), 50-80.

Ma, J., & Cheng, L. (2015). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth, and signi cance. *TESL Canada Journal, 33*(1), 58-79. https://doi.org/10.18806/tesl.v33i1.1227

Marriott, H., & Nekvapil, J. (2012). Language management approach: Probing the concept of "noting". *Journal of Asian Pacific Communication, 22*(2), 155-317. https://doi.org/10.1075/japc.22.2.01mar

Matoush, M. M., & Fu, D. (2012). Tests of English language as significant thresholds for college-bound Chinese and the washback of test-preparation. *Changing English, 19*(1), 111-121. https://doi.org/10.1080/1358684X.2012.649176

Minakova, V. (2020). Dynamic assessment of IELTS speaking: A learning-oriented approach to test

preparation. *Language and Sociocultural Theory, 6*(2), 184-212. https://doi.org/10.1558/lst.36658

Mohammadi, L. (2016). *The effect of applying time constraints on IELTS candidates' writing error types and their attitudes* [Unpublished Master's thesis]. Payame Noor University.

Naseri, S., Maghsoudi, M., & Rajabi, P. (2014). The effect of speed reading on IELTS EFL learners' reading comprehension ability. *International Journal of Language Learning and Applied Linguistics, 5*(10), 506-514.

Nekvapil, J. (2015). Language management theory as one approach in language policy and planning. *Current Issues in Language Planning, 17*(1), 11-22. https://doi.org/10.1080/14664208.2016.1108481

Neustupný, J. V. (1994). Problems of English contact discourse and language planning. In T. Kandiah & J. Kwan-Terry (Eds.), *English and language planning* (pp. 50-69). Academic Press.

O'Sullivan, B., Dunn, K., & Berry, V. (2021). Test preparation: An international comparison of test takers' preference. *Assessment in Education: Principles, Policy & Practice, 28*(1), 13-36. https://doi.org/10.1080/0969594X.2019.1637820

Pallant, J. (2005). SPSS survival guide: A step by step guide to data analysis using SPSS for windows (3rd ed.). Open University Press.

Paul, A. M. (2012, April 13). How to be a better test-taker. *The New York Times*. https://www.nytimes.com/2012/04/15/education/edlife/how-to-be-a-better-test-taker.html

Powers, D. E. (1993). Coaching for the SAT: A Summary of the Summaries and an Update. *Educational Measurement: Issues and Practice, 12*, 24-30. https://doi.org/10.1111/j.1745-3992.1993.tb00530.x

Powers, D. E. (2017). Understanding the impact of special preparation for admissions tests. In

Advancing human assessment (pp. 553-564). Springer. https://doi.org/10.1007/978-3-319-58689-2_17

Rahman, M. M., & Pandian, A. (2018). A critical investigation of English language teaching in Bangladesh: Unfulfilled expectations after two decades of communicative language teaching. *English Today*, *34*(3), 43-49. https://doi.org/10.1017/S026607841700061X

Read, J., & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. *IELTS Research Reports, 4*(5), 153-205.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

Roza, D. (2019). The challenges and strategies of teachers in teaching TOEFL and IELTS test preparation. *Journal of English for Academic, 6*(2), 1-13. https://doi.org/10.25299/jshmic.2019.vol6(2).3067

Saif, S., Ma, J., May, L., & Cheng, L. (2021). Complexity of test preparation across three contexts: Case studies from Australia, Iran and China. *Assessment in Education: Principles, Policy & Practice, 28*(1), 37-57. https://doi.org/10.1080/0969594X.2019.1700211

Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., Harcombe, W., McGaugh, S., Wassenberg, D., Yonas, A., & Ballen, C. J. (2019). Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety. *Frontiers in Education, 4*, Article 107. https://doi.org/10.3389/feduc.2019.00107

Samad, A. A. (2019). *Essentials of language testing for Malaysian teacher*. UPM Press.

Samad, A. A., Rahman, S. Z. S. A., & Norbaiti, S. (2008). Refining English language tests for

university admission: A Malaysian example. *Asian Journal of University Education (AJUE), 4*(1), 57-68.

Sato, T. (2019). An investigation of factors involved in Japanese students' English learning behavior during test preparation. *Papers in Language Testing and Assessment*, *8*(1), 69-95.

Schwartz, L. F., Lineberry, M., Park, Y. S., Kamin, C. S., & Hyderi, A. A. (2018). Development and evaluation of a student-initiated test preparation program for the USMLE step 1 examination. *Teaching and learning in medicine, 30*(2), 193-201. https://doi.org/10.1080/10401334.2017.1386106

Singh, C. K. S., & Samad, A. A. (2013). Portfolio as an assessment tool and its implementation in Malaysian ESL classrooms: A study in two secondary schools. *Pertanika Journal of Social Sciences & Humanities, 21*(4), 1255-1273.

Singh, C. K. S., Samad, A. A., Hussin, H., & Sulaiman, T. (2015). Developing a portfolio assessment model for the teaching and learning of English in Malaysian L2 classroom. *English Language Teaching, 8*(7), 164-173. https://doi.org/10.5539/elt.v8n7p164

Sotelo-Dynega, M., Ortiz, S. O., Flanagan, D. P., & Chaplin, W. F. (2013). English language proficiency and test performance: An evaluation of bilingual students with the Woodcock-Johnson III tests of cognitive abilities. *Psychology in the Schools, 50*(8), 781-797. https://doi.org/10.1002/pits.21706

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly, 4*(3), 279-293. https://doi.org/10.1080/ 15434300701462937

Sultana, N. (2018). Test review of the English public examination at the secondary level in Bangladesh. *Language Testing in Asia*, *8*(1), 1-9. https://doi.org/10.1186/s40468-018-0068-1

Sultana, N. (2019). Language assessment literacy: An uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia*, *9*(1), 1-14. https://doi.org/10.1186/s40468-019-0077-8

Tulloh, R., & Wood, S. (1998). *IELTS Research Reports* (Vol. 1). ELICOS Association.

Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 19-36). Lawrence Erlbaum Associates.

Winke, P., & Lim, H. (2017). The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly*, *14*(4), 380-397. https://doi.org/10.1080/15434303.2017.1399396

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly, 10*(2), 196-218. https://doi.org/10.1080/15434303.2012.721423

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49-70. https://doi.org/10.1177/0265532212442634

Yang, Y., & Badger, R. (2015). How IELTS preparation courses support students: IELTS and academic socialisation. *Journal of Further and Higher Education, 39*(4), 438-465. https://doi.org/10.1080/0309877X.2014.953463

Yusefzadeh, H., Iranagh, J. A., & Nabilou, B. (2019). The effect of study preparation on test anxiety and performance: a quasi-experimental study. *Advances in Medical Education and Practice, 10*, 245-251. https://doi.org/10.2147/AMEP.S192053

Zakaria, N. A., Samad, A. A., & Omar, Z. (2013). Pressure to improve scores in standardized English examinations and their effects on

classroom practices. *International Journal of English Language Education, 2*(1), 45-56. https://doi.org/10.5296/ijele.v2i1.4524

Zhang, Y. (2008). *Repeater analyses for TOEFL® iBT* (ETS Research Memorandum). Educational Testing Service.

Zhengdong, G. (2009). IELTS preparation course and student IELTS performance: A case study in Hong Kong. *RELC Journal, 40*(1), 23-41. https://doi.org/10.1177/0033688208101449

Language Education for Orang Asli Children in Malaysia

Hema Letchamanan[1]*, Nur Surayyah Madhubala Abdullah[2] and Kamal Solhaimi Fadzil[3]

[1]*School of Education, Taylor's University, No. 1 Jalan Taylor's, 47500 Subang Jaya, Selangor, Malaysia*
[2]*Department of Language and Humanities Education, Faculty of Educational Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[3]*Department of Anthropology and Sociology, Faculty of Arts and Social Sciences, University of Malaya, Jalan Ilmu, 50603 Kuala Lumpur, Malaysia*

ABSTRACT

Literacy among Orang Asli children is found to be well below the national average. This paper explores the connection between language learning and meaning-making and its relation to the problem of language education among Orang Asli children. In so doing, the paper shows that language learning should be situated within their environment. This paper uses the findings from the observation made in an exploratory case study of Temuan children aged 7–12 in an Orang Asli village in Selangor and their struggles with language. Taking a critical view of the challenges faced by the children, this study surmised that a proper recognition of the Orang Asli community in language education is needed for effective meaning-making to ensure their genuine participation. The insight adds to the discussion within decolonisation of education on the importance of indigenisation of language education for Orang Asli children.

Keywords: Environment, indigenous, language education, literacy, meaning-making

INTRODUCTION

At its basic level, literacy in language learning involves the acquisition of listening, speaking, reading, and writing skills. It makes literacy a crucial first step towards children's ability to make sense of the world around them. Children's early literacy development is crucial for their learning opportunities and success in school (Hare, 2011; Roberts et al., 2005; Sénéchal et al., 2006; Whitehurst & Lonigan, 2002).

Studies on literacy of indigenous children in countries such as Canada, Australia, and New Zealand have found an incongruence in the development of literacy skills between indigenous and non-indigenous children (Cowley & Easton, 2004; Frigo et al., 2004; Hare, 2011) that has led to indigenous people not being able to participate meaningfully in society. Further, Article 14.1(3) of United Nations Declaration on the Rights of Indigenous People (cited from Bauer, 2007, p.13) specifies the right of indigenous children to receive education "in their own culture and...in their language". In the case of Malaysia, while much has been done to support the education of Orang Asli children, there are still gaps in their performance and achievement in schools compared to non-Orang Asli children (Wan, 2020).

The incongruence in performance between indigenous and non-indigenous children is of concern considering the fact that two-thirds of the world's indigenous population lives in Asia (Errico, 2017). Compared to non-indigenous children, indigenous children do not have access to the same quality of education (Shay & Sarra, 2021). Indigenous children and children from other marginalised communities experience literacy differently from the literacy practices and expectations in school (Arnold & Doctoroff, 2003; Heath, 1983; MacNaughton, 2006; Neuman, 2006). Therefore, the children, do not perform well in school because of this disconnect from their reality (Valdés,

1996) contributes to indigenous children facing more educational challenges than children from other communities (Anderson et al., 2016). Edo et al. (2013) refer to the disconnect with indigenous students' reality as social exclusion. Here they refer to a wider exclusion to include politics, economy, and education, contributing further to the marginalisation of the indigenous community. Policies aimed at improving the lives of Orang Asli, including the establishment of the Department of Aborigines, known as JAKOA (Department of Orang Asli Development), worked towards alienating the community from mainstream society. More importantly, it also disempowered Orang Asli to customary land and self-representation (Dentan et al., 1997).

For Orang Asli children, their lives are deeply rooted in their customary land, which is culture imbued with nature and its elements. They also mostly come from an oral tradition (Nicholas, 2004), so the children's early years may develop literacy in their language through an oral framework. The Orang Asli children's performance in school has been affected by a lack of recognition in the school curriculum on the importance of their culture (Errico, 2017). Not fully recognising the community's literacy experiences has contributed to the language education in Malaysia not adequately allowing Orang Asli children to engage in meaningful learning in schools (Renganathan, 2013). Meaningful learning

in this context refers to recognition of the Orang Asli community, which would lead to genuine participation by the children. Meaningful learning allows children to retain their knowledge better when learning new things (Ausubel & Ausubel, 2010) as they need to connect new information with their pre-existing concepts (Vallori, 2014). When the focus is not on achieving meaningful learning, Orang Asli children tend to be discouraged. Therefore, they face difficulty in making deep connections and relating the learning to themselves and their environment; in other words, *meaning-making* does not occur.

Language education is seen as a tool to facilitate the development of literacy skills to ensure their effective participation in society, for example, accessing services, articulating their rights as citizens, and contributing to nation-building. As a marginalised community, the Orang Asli's ability to participate meaningfully in society is critical to empower the community as citizens. At present, the Orang Asli population in Malaysia is disenfranchised and underrepresented in society (Nicholas, 2021; Sato, 2019). Even though the Orang Asli population in Peninsular Malaysia is not a homogeneous group, the use of the singular term "community" in literature on language education suggests a disconnect between learning and meaning-making in standardising language education across the different ethnic groups. A standardised language education raises concern about whether meaningful participation among the Orang Asli children can take

place. Therefore exploring whether a contextualised language education that considers the Orang Asli children's lived experiences and environment could help address the imbalance in the community's education.

Article 28 of the Convention on the Rights of the Child, which recognises the right to education for all children, supports the examination of this issue Hence, standardisation of it will pose some challenges to indigenous children. The notion of standards has always been conceptualised as a top-down process, with communities having little to no say in developing a curriculum. In the case of the Orang Asli community, their perspectives and environment need to be included to support literacy development (Kral, 2009; Renganathan, 2013).

Studies on Orang Asli and their learning have lacked focus on education (Edo et al., 2013; Wan, 2020), literacy, and meaning-making, yet it is an important consideration in ensuring children can relate to and engage with their learning. Poor literacy levels are often associated with low test scores, poor attendance, and disciplinary problems. While these factors are important, they do not provide a good sense of the problem of literacy faced by OA children in language education. This paper proposes meaning-making in language learning as an important aspect in understanding this problem. The case of Temuan children aged 7-12 in an Orang Asli village in Selangor is used to present the problem of language education for these children.

An Issue in Education of Orang Asli Children

Orang Asli is a legal category defined under the Orang Asli Act 134 (1954) (as contained in Government of Malaysia, 2010). Applying a cultural definition, the Act defines an Orang Asli as any person who speaks an Orang Asli language, practices Orang Asli cultures, and remains a member of an Orang Asli community. The recent official census places the total number of Orang Asli at approximately 178,197 (DOSM, 2019; JAKOA, 2018), or less than 1% of the national population (Nicholas, 2021). They consist of 19 sub-groups and vary in population size, distribution, political, and social organisation, and traditional economic practices. The three broad groups are; the Senoi, Proto Malay, and Negrito. The largest group is the Senoi, which makes up approximately 54% of the population. The second largest group is the Proto-Malay, which makes up approximately 43% of the overall population. Finally, the smallest group is the Negrito, with approximately 3% of the overall Orang Asli population.

Government policy has been consistent in that it looks to incorporate Orang Asli communities within the larger Malaysian society. In 1961, the Malaysian government formulated a cohesive policy to address integration and development of the Orang Asli in the "Statement of Policy Regarding the Long-Term Administration of the Aboriginal Peoples in the Federation of Malaya". In the 1970s, the government proposed developing a settlement scheme termed *Rancangan Pengumpulan Semula*

(RPS), modelled after the FELDA (Federal Land Development Authority) scheme. The main focus is on agriculture-based development programs. The scheme includes an administrative hub, pre and primary level schooling, and medical facilities. Families would also receive government-built houses and other basic amenities. Today there are about 17 RPS throughout the Peninsular. By the 1990s, keeping in line with the general development policy, government policy for Orang Asli development focused on growth led by the private sector through entrepreneurship initiatives. The Strategic Plan for Orang Asli Development 2011–2015 outlined six development thrusts, one of which was a human development model and its relation to education/literacy. The most recent strategic plan is a collaboration between Jabatan Kemajuan Orang Asli and the United Nations Development Programme (UNDP) that aims to translate the Sustainable Development Goals as key strategies for Orang Asli development until 2030. However, in empowering Orang Asli as citizens with equal representation, much more is needed to engage Orang Asli as a stakeholder on policy matters that affect their community; one area being education.

Indigenous people are historically accepted as being present before the 1400s (Andaya & Andaya, 2017). Despite being one of the earliest inhabitants of Malaysia, the quality of education received remains below the national average. The dropout rate for Orang Asli children in 2017 was 26%, compared to the national average, which was consistently below 4% from

2016 to 2018 (Wan, 2020). In 2015, the Ministry of Education reported that the Orang Asli children comprised 4% of the national student population (MOE, 2016b). Recent statistics suggest that out of 26,571, only 13,155 enrolled in secondary school (JAKOA, 2018). Out of those who go on to secondary school, not all complete their secondary education.

Based in the preceding paragraph, the outlook for Orang Asli children's education in achieving the National Education Blueprints targets is worrying. While some claim that the problem of education among Orang Asli is because of the lack of awareness among the community, specifically the parents (Mazzlida & Ruhizan, 2016; Sawalludin et al., 2020), there are other studies (IDEAS, 2020; Nicholas, 2005) that suggest otherwise. The problems that the Orang Asli community faces in education are multi-faceted and require deeper analysis.

While dropout rates between 2016 and 2018 seem to have been falling, the struggle of Orang Asli children with education should still be of concern. There is a lack of substantive data to formulate better policy and inform practice on language education for Orang Asli. The diversity among Orang Asli communities suggests the need for inter and intra-group research. Indeed, field studies indicate that the education problem among Orang Asli children is more complex and should be explored from different angles to inform policy implementation better.

Quality education requires an inclusive and equitable policy for it to be effective and sustainable. While policy on education in general, and specifically for Orang Asli, is encouraging, achievement of results requires attention to challenges of the target population in implementation as they affect the achievement of the policy outcomes. The empowerment of Orang Asli depends on substantive approaches to policy and practice that consider the actual situation of marginalised Orang Asli children in the development of language education. The problem faced by indigenous children suggests that policy and practice which lack recognition of marginalised Orang Asli children's learning hinder their development (Romero-Little, 2010). Furthermore, a lack of research that considers language education from the perspective of the situation of marginalised Orang Asli children themselves fails to significantly contribute to addressing the problem of literacy. Research should focus on providing clarity to the problem of literacy among Orang Asli children to identify what actions can assist policy and practice on language education in moving forward (Ainscow, 2020). This study focuses on whether language education currently adopted is properly framed to the needs of the children to better support the educational needs of indigenous communities in Peninsular Malaysia.

Decolonisation and Indigenisation of Education for Orang Asli Children

Orang Asli children's learning, despite the efforts made by the relevant authorities, has not seen a significant improvement.

One reason for this is the lack of inclusion of their culture and indigenous identity in mainstream education (Rosnon & Talib, 2019)

Decolonisation of education in the context of indigenous children in Malaysia involves looking into education standards and their impact on their learning. Decolonising involves dismantling the assumed knowledge children should acquire in schoolings, such as the content, skills, and values taught in schools (Pratt & Vries, 2018). The focus has to be on the challenges these children face in learning literacy in schools to ensure education responds to the learning needs of indigenous children. It can then provide a nuanced narrative on the need to indigenize language education. To better understand the need for the decolonisation of Orang Asli children's literacy education, it is imperative to shed some light on their present learning and education.

Children develop their literacy skills by making sense of the world around them (Husbye & Dorner, 2017). For Orang Asli children, their connection to the environment may be different; thus, this requires consideration in language education. However, the question that needs to be asked is to what extent does the current language education accommodate their meaning-making process. For example, as shown in Figure 1, meaning-making might be lost in this lesson when the notion of a pet is understood differently amongst children from different socio-cultural backgrounds. This situation arises because a pet can be understood as a domestic animal kept for companionship by an individual or a family or could be seen in a broader sense as an animal collectively kept by a community. Therefore, when Orang Asli children are asked to talk about their pets, it is important to consider that their understanding of pets

MATERIAL ADAPTATION AND SUGGESTED TEACHING AND LEARNING ACTIVITY

Warm Up:
1. Recall the previous lesson on pg 58 of the textbook.
2. Get pupils to talk about their pets.

Presentation:
1. Show a picture of a pet (not from the textbook).
2. Ask pupils questions on the pet.
3. Write the answers on the board.
4. Show pupils how to construct sentences based on the answers of the pupils.

Practice:
1. Get pupils to sit in groups of 4.
2. Pupils complete exercises on pg 59 of the textbook.
3. Provide other pictures of other possible pets.
4. Pupils write sentences about the pets.

Reinforcement / remedial:
Pupils draw their dream pets in their exercise books.

Figure 1. Taken from Common European Framework of Reference for Languages (CEFR) - English Language Handbook for Primary School Teachers (MOE, 2016a, p.3)

may differ from that of other children in the classroom. Moreover, drawing on their cultural norms to guide the meaning-making process would help them acquire literacy skills. If this is not the case, meaning-making may be lost in the language classroom, and children may not participate in the learning process.

Conceptual Framework

Several factors influence the learning of indigenous children and their educational development; socio-economic status, home environment, school context, and individual child's life experiences (Doyle & Hill, 2008; Ockenden, 2014). This paper focuses on the influence of school context, specifically language education, on Orang Asli children's language learning. Based on Bronfenbrenner's Ecological Systems Theory (EST) and Bruner's idea of active learning, these aspects are situated within the social ecology of the children, interconnected with their language learning, and related to how they make sense of the learning.

EST (Bronfenbrenner, 1977) shows the different systems that influence a child's development and how these systems are interrelated, both within and between the systems. According to EST, the interaction between and within the five systems influences how a child develops and grows (Wilson et al., 2020). Therefore, the social ecology in which a child is situated should be considered to ensure meaningful learning occur. In the context of education for indigenous children, EST

points to the importance of considering the ecology in articulating what education, and specifically language education, should be for indigenous children. It suggests that the environment, for instance, the curriculum, would affect their learning (Matengu et al., 2019). Thus, progress in language learning for indigenous children is influenced by where language education is situated for the children and to what extent it considers their ecology.

Bruner (1966a) adds a dimension about children's learning to the framework. His views on learning as an active process, where the child constructs their knowledge based on current and past experiences, is important for the discussion on language learning for indigenous children. In constructing their knowledge, the children are making sense of what they are learning. However, meaning-making is a process that extends beyond just learning to also include identity and emotions (Zittoun & Brinkmann, 2012). Meaning-making occurs at three levels: semantics, pragmatics, and existential (ibid). Children identify and associate words, symbols, and sounds with their conceptual meaning associated with cultural understanding at the semantics level. Pragmatic meaning occurs when children identify with the social practices, which continuously changes and expands. Finally, children make sense of learning at the existential level through their lived experiences that shape their emotions and identities. When this meaning-making process is hindered at any one of these levels, it affects language learning. One

example of how this may occur in language learning for indigenous children is through cultural differences. Language imposes certain cultural norms that interfere with or hinder meaning-making when the said norms are not part of the children's everyday lives or environment.

Based on the preceding explanation, the premise of this study is that: (1) meaningful language learning occurs when children can make sense of their learning; (2) children's sense-making occurs when they can relate to their environment (environment in the context of Orang Asli children is understood in a broad sense which includes family, community, village, respective indigenous culture, own languages, and individual experiences); and therefore (3) if language education for Orang Asli children takes into consideration their environment, it is more likely to aid the children in their language learning.

METHODOLOGY

The study adopted a philosophical approach that applied a critical lens to the issue of language education for Orang Asli children. The issue was identified from an exploratory case study on Temuan children in one Orang Asli village in Selangor. Based on the conceptual framework, a critical lens that involved personal reflection, observation, and authority/experience as educators and experts in the field of education was applied to the problems these children face in language learning. The conclusions reached were used in the discussion on the issue of language education for Orang Asli children.

The exploratory case study was conducted to obtain preliminary data (Mills et al., 2010; Yin, 2018) on the situation faced by the children in language education. As part of the study, a diagnostic test was administered to determine the children's literacy level in four subjects; Bahasa Melayu, English, Mathematics, and Science. Studies have shown that language proficiency impacts performance in other subjects such as Mathematics and Science (Bayat et al., 2014; Henry et al., 2014; Neri et al., 2019; Prediger et al., 2018). In addition, observations were made about problems encountered by the children when responding to the test questions. Participants for the exploratory case study were chosen through convenience sampling. These participants have had some experiences of language learning in school. Therefore, they were able to provide narratives of their language learning. There were 20 participants aged between seven and twelve years ranging from Year One to Year Six: nine were in the upper primary (Year 4–Year 6), and 11 were in the lower primary (Year 1–Year 3). Most of these children attend a national school where the medium of instruction is Bahasa Melayu. Out of the 20 children, three had already dropped out of school for various economic and social reasons. For the remaining 17, their school attendance was poor and academic performance was below the expected level for their age group. The ones in the upper primary were also faced with the possibility of not transitioning to secondary school because of poverty, distance to school, the

perspective of studying, and concerns of performance in secondary school.

The problem faced by these children in language learning and meaning-making was identified using observation and their performance in these tests. The observations also provided a context for understanding these children's interests, response to the tests, and performance (and non-performance) in the tests. The test was administered over four weekends using topics taken from the Year Two National Primary School Standard Curriculum on English, Bahasa Melayu, Science, and Mathematics. These tests were based on the key learning outcomes of the respective subjects. Thus, a year two-level test provided a suitable perspective that could be used to develop the narrative on language learning experiences across the different schooling levels of the children. The tests were administered towards the end of the school year; therefore, it was assumed that the children in Year One would be able to respond to at least some of the questions. These responses would demonstrate the problems of language education that affect meaning-making.

A non-participant, semi-structured observation was conducted using a protocol that looked at how the children responded to the tests to understand the nature of reading and writing in Bahasa Melayu and English. The observation protocol was framed on the Classroom Language Observation Checklist (CLOCK) (CAL, 2015) and adapted for this study. Four aspects were included in the protocol; (1) vocabulary, (2) comprehension, (3) language control, and (4) fluency. For vocabulary, the focus was on the ability of the children to use grade-level words in answering the test questions. Comprehension was observed through the ability of the children to understand the instructions and questions in the test. Language control focused on how well the children used words, phrases, and sentences in Bahasa Melayu and English in answering the test questions. Finally, fluency considered the ease of understanding the instructions and questions.

Themes were identified from the observation on the four aspects stated in the protocol using thematic analysis. The thematic analysis was carried out through coding, looking for commonalities and contrasts. From this analysis, two themes emerged; (1) challenges faced when responding to the diagnostic test; and (2) the connection between language learning and meaning-making. Finally, the conceptual framework was applied to these themes to suggest the areas of concern in language education for Orang Asli children.

The Case

The Temuan are of the Proto-Malay ethnic subgroup of peninsular indigenous people. The first Temuan families settled in this village around 60 years ago. To date, around 20 families are living in the village (information obtained from the village headman, also known as *Tok Batin*). JAKOA and other organisations built some houses to accommodate the families, a kindergarten and a community hall. However, the

villagers only received electricity in 2017, having relied on a generator previously. In addition, there is no proper road, and access by vehicles is limited. Thus, despite its proximity to some of the more affluent suburbs in Selangor, the infrastructure and facilities in this village are still of concern. This scenario together with poverty have affected the Temuan children's schooling and contributed to their low literacy level.

FINDINGS

A critical analysis was done on the two themes identified in the observation. The two themes were: (1) challenges in responding to tests; and (2) response to tests, and they are described below. Themes and inferences drawn took into consideration that the child's learning is affected by meaning-making and the child's environment is relevant to their learning.

Challenges in Responding to Tests

Observation on the children's approach to the diagnostic tests revealed the varied nature of their responses in reading and writing in Bahasa Melayu and English. There was a sense of attentiveness among the children to focus and complete the tests. However, some became distracted and gave up answering the questions. Children who could read and write proceeded to do the tests with some assistance from the facilitators. They attempted to read and answer all the questions within the time given. The children who seemed to have difficulty reading and understanding the questions tried initially to attempt the test

but soon left their table to do other things, such as play with other children who were not involved in the test and scribble on the board. As these tests were administered over four weeks, it was noted that some students were motivated and looked forward to the next test. When the students completed one test, they asked about the next one, the subject, what would be tested and when it would be conducted. Some did not want to participate in the next test because they felt the questions were too difficult to answer. The children who had already dropped out of school were not motivated in taking the tests. When this observation was done, three children from the group had dropped out of school (two girls and one boy). The girls had to stop schooling so that they could look after their younger siblings. They were nine and twelve years old, respectively. The boy, aged ten, had stopped schooling so that he could help his mother collect bamboo. When asked, they did not see the test's purpose as they were no longer in school.

Response to Tests

Regarding the children's response to tests, their engagement with the questions depended on their ability to read and understand the instructions and questions. When questions were read out to them, the children were able to respond verbally, and in most instances, provided correct answers. It was especially evident in the Mathematics and Science tests. However, in the Bahasa Melayu and English language tests, more than half of the children struggled to comprehend the meaning of the sentences

Soalan 7, berdasarkan gambar di bawah.

7. Farhan menyapu tingkap dengan
    menggunakan se_____ bulu
    ayam.

    A   helai
    B   tangkai
    C   batang

Figure 2. Sample Bahasa Melayu diagnostic test
question

even when they were read aloud to them.
Another important observation was that
these children's understanding was hindered
by the cultural context in which the questions
were situated. For example, the following
question was taken from the Bahasa Melayu
diagnostic test.

Another important observation was
that these children's understanding was
hindered by the cultural context in which
the questions were situated. For example,
the following question was taken from the
Bahasa Melayu diagnostic test. When this
question and the options were read out to
the children, they could not associate the
image with the action of cleaning a window
(Figure 2). The children asked what the
image was since they had never seen it
(Figure 2). When asked what they usually
use to clean the windows, they replied with
a piece of cloth, and this answer was stated
correctly in Malay. Some children also
expressed their concern about the source of

the feathers. Having developed an empathy
for the livestock they have grown up with,
they looked out of the window for their
chickens roaming around the village and
could be seen to be visibly affected.

DISCUSSION

The findings suggest two matters for
discussion:

Orang Asli Children and their Meaning-making

Observations on the children's approach
to the diagnostic tests demonstrate their
low literacy level, which is shown by their
ability and inability to cope with reading
and understanding the test questions. The
problem of meaning-making here is situated
at two levels; semantics and pragmatics. The
third level, existential, was not included
in the study conducted because it was an
exploratory study. At the semantics level,
it is a problem of decoding but inability to
associate the words to their meaning. At
the pragmatics level, there is difficulty for
the children to relate to questions situated
in a particular cultural context that may
be foreign to them (Zittoun & Brinkmann,
2012). However, children who showed some
understanding at the semantic level could
not cope with the pragmatic understanding.
This situation points towards a difficulty in
making sense of language use. Therefore,
language learning within indigenous
children's cultural context is more likely to
assist them in meaning- making (Bruner,
1966b; Siekmann et al., 2017).

The Interconnectedness of Language Learning and Environment

The findings suggest that more needs to be said about language learning within the context of Orang Asli children's environment. It should be understood in a broader and deeper sense that includes tangible and intangible aspects such as family, customary land, identity, and individual life aspirations. Referring to the Ecological Systems Theory discussed in Section Four of this paper, language education that would benefit Orang Asli children's learning is situated within their environment as learning and environment are interconnected. As Romero-Little (2010) points out, children from indigenous communities need to recognise their environment in articulating education, and failure to do so may hinder their learning. As she observes, indigenous children's environment may vary widely from mainstream children, therefore of itself should not be a barrier to learning. As such, if this is adequately considered or recognised, it could support their learning. For the Orang Asli children, if the aspect of the environment is better understood, then language education can be better conceptualised for these children, and therefore benefit their language learning.

Overall, the mismatch observed between language learning and Orang Asli children's meaning-making in the case cited points to the lack of recognition of the environment as one of the factors that can hinder their language education. Furthermore, language learning is often not placed within the children's cultural and environmental contexts. It shows a need to decolonise language education by considering what is understood by literacy learning and how it is presented to Orang Asli children. For instance, the context in which questions are framed, such as test questions, is important for children to understand not just at the semantic but also at the pragmatic level to make sense of what is asked of them in terms of learning. In addition, what is asked of them in terms of learning does not adequately offer an opportunity to include their identity, which poses a problem of existential understanding. Failure to consider the required levels in how language education for Orang Asli children is conceptualised and delivered can contribute to performance and achievement not only in language but also in other school subjects (Bayat et al., 2014; Henry et al., 2014; Neri et al., 2019; Prediger et al., 2018).

IMPLICATION AND CONCLUSION

Whilst this paper acknowledges that the consideration for this study is based on observations from a preliminary study, it offers an insight into the issue of language education for Orang Asli children. There is a tendency in some studies to articulate the problem of education among Orang Asli children as being situated in the children and their communities (Sawalludin et al., 2020). However, the findings of this study suggest that this may not be the only case. Therefore future studies on language education for Orang Asli children should take into account their environment in better understanding the problem of their language learning.

This paper questions the current narrative on language education in Malaysia, in particular the disconnect between mainstream language education and the Orang Asli children. Despite the many measures taken to ensure the continuity of learning for Orang Asli children, the teaching of language in school should be further explored to adequately situate it within the needs of indigenisation of Orang Asli children's education. One way is for language education to represent Orang Asli's cultural practices and norms, allowing for effective meaning-making. It should also be linguistically and culturally appropriate to bridge their transition from home to school as their home language may be different (Ball, 2009).

One way forward to address this implication would be to further investigate the Orang Asli children's educational needs from their perspective, as Nicholas (2010) suggested. It means starting from their literacy traditions and perspectives on life and living. Returning to Ball (2009), this requires consideration of elements of meaningful learning for Orang Asli children. This consideration should also take into account that the Orang Asli community comprises 19 sub-ethnic groups, and therefore should not be viewed as a homogenous group in the research and development of language education. Improvements to their language learning would have significance to the broader problem of low literacy levels among Orang Asli children. The exploratory nature of this study limits the discussion on meaning-making in language education among Orang Asli children. A longitudinal study that encompasses both the home and school environments could further enrich the discussion.

ACKNOWLEDGEMENT

REFERENCES

Ainscow, M. (2020). Promoting inclusion and equity in education: Lessons from international experiences. *Nordic Journal of Studies in Educational Policy*, *6*(1), 7-16. https://doi.org/10.1080/20020317.2020.1729587

Andaya, B. W., & Andaya, L. Y. (2017). *A history of Malaysia (3rd Ed)*. Red Globe Press.

Anderson, I., Robson, B., Connolly, M., Al-Yaman, F., Bjertness, E., & King, A., Tynan, M., Madden, R., Bang, A., Coimbra Jr, C. E. A., Pesantes, M. A., Amigo, H., Andronov, S., Armien, B., Obando, D. A., Axelsson, P., Bhatti, Z. S., Bhutta, Z. A., Bjerregaard, P., … & Yap, L. (2016). Indigenous and tribal peoples' health (The Lancet–Lowitja Institute Global Collaboration): A population study. *The Lancet*, *388*(10040), 131-157. https://doi.org/10.1016/s0140-6736(16)00345-7

Arnold, D., & Doctoroff, G. (2003). The early education of socioeconomically disadvantaged children. *Annual Review of Psychology*, *54*(1), 517-545. https://doi.org/10.1146/annurev.psych.54.111301.145442

Ausubel, D., & Ausubel, D. (2010). *The acquisition and retention of knowledge*. Kluwer Academic Publishers.

Ball, J. (2009). supporting young indigenous children's language development in Canada: A review of research on needs and promising practices. *The Canadian Modern Language Review*, *66*(1), 19-47. https://doi.org/10.3138/cmlr.66.1.019

Bauer, A. (Eds.). (2007). United Nations declaration on the rights of indigenous peoples. International Journal of Cultural Property, *14*(4), 399-408. https://doi.org/10.1017/s0940739107070270

Bayat, N., Şekercioğlu, G., & Baker, S. (2014). The relationship between reading comprehension and success in science. *Journal of Education and Science*, *39*(176). https://doi.org/10.15390/eb.2014.3693

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513-531. https://doi.org/10.1037/0003-066x.32.7.513

Bruner, J. S. (1966a). *Toward a theory of instruction*. Belknap Press of Harvard University.

Bruner, J. (1966b). *The culture of education.* Harvard University Press.

CAL. (2015). *CLOCK: Classroom observation checklist*. Centre for Applied Linguistics.

Cowley, P., & Easton, S. T. (2004). *Report card on aboriginal education in British Columbia.* Fraser Institute.

Dentan, R. K., Endicott, K., Gomes, A.G., & Hooker, M.B. (1997). *Malaysia and the original people – A case study of the impact of development on indigenous people*. Massachusetts: Allyn and Bacon

DOSM. (2019). *Current population estimates, Malaysia, 2018-2019*. Department of Statistics, Malaysia.

Doyle, L., & Hill, R. (2008). *Our children, our future - Achieving improved primary and secondary education outcomes for Indigenous students: an overview of investment opportunities and approaches*. AMP Foundation, Effective Philanthropy & Social Ventures Australia.

Edo, J, Awang, S. N., Fadzil, K. S., Zainol, R., & Ibrahim, W. S. W. (2013). Pendidikan dan hak asasi kanak-kanak: Satu tinjauan cabaran dan halatuju pendidikan awal kanak-kanak Orang Asli di Kompleks Belum-Temenggor [Education and human rughts of chidren; A review of challenges and direction of Orang Asli early childhood education in Belum-Temenggor Complex]. *Manusia dan Masyarakat, 23*, 1-14.

Errico, S. (2017). *The rights of indigenous peoples in Asia - Human rights-based overview of national legal and policy frameworks against the backdrop of country strategies for development and poverty reduction.* https://www.ilo.org/wcmsp5/groups/public/---dgreports/---gender/documents/publication/wcms_545487.pdf

Frigo, T., Corrigan, M., Adams, I., Hughes, P., Stephens, M., & Woods, D. (2004). *Supporting English literacy and numeracy learning for Indigenous students in the early years*. ACER Press.

Government of Malaysia. (2010). *Aboriginal people Act 1954 (Act 134), Pub. L. No Act 134 (1954)*. Government Printer.

Hare, J. (2011). They tell a story and there's meaning behind that story: Indigenous knowledge and young indigenous children's literacy learning. *Journal of Early Childhood Literacy*, *12*(4), 389-414. https://doi.org/10.1177/1468798411417378

Heath, S. B. (1983). *Ways with words*. Cambridge University Press.

Henry, D., Nistor, N., & Baltes, B. (2014). Examining the relationship between math scores and English language proficiency. *Journal of Educational Research and Practice, 4*(1), 11-29.

Husbye, N., & Dorner, L. (2017). Teaching strategies to develop inquiry and literacy skills: Languaging in foreign language immersion education. *L2*

Journal, *9*(1), 36-48. https://doi.org/10.5070/l29128532

IDEAS. (2020). *IDEAS annual report.* https://www.ideas.org.my/wp-content/uploads/2021/04/IDEAS_AnnualReport_2021_V10.pdf

JAKOA. (2018). *Demografi Orang Asli Mac 2018.* https://www.data.gov.my/data/ms_MY/dataset/demografi-orang-asli-mac-2018.

Kral, I. (2009). *The literacy question in remote Indigenous Australia: CAEPR Topical Issue 06/2009.* Centre for Aboriginal Economic Policy Research.

MacNaughton, S. (2006) *Considering culture in research based interventions to support early literacy.* In S. B. Neuman & D. K. Dickinson (Eds.), Handbook of early literacy research (Vol. 2, pp. 229-242). The Guilford Press.

Matengu, M., Korkeamäki, R., & Cleghorn, A. (2019). Conceptualizing meaningful education: The voices of indigenous parents of young children. *Learning, Culture and Social Interaction*, *22*, Article 100242. https://doi.org/10.1016/j.lcsi.2018.05.007

Mazzlida, M. D., & Ruhizan, M. Y. (2016). Quality education of Orang Asli in Malaysia. *International Journal of Academic Research in Business and Social Sciences, 6*(11), 233-240.

Mills, A. J., Durepos, G., & Wiebe, E. (2010). *Encyclopedia of case study research.* SAGE Publication.

MOE. (2016a). *Common European Framework of Reference for Languages (CEFR) - English language handbook for primary school teachers.* Ministry of Education Malaysia.

MOE. (2016b). *Malaysia education blueprint 2013-2025*. Ministry of Education.

Neri, N. C., Guill, K., & Retelsdorf, J. (2019). Language in science performance: Do good readers perform better? *European Journal of Psychology of Education*, *36*(1), 45-61. https://doi.org/10.1007/s10212-019-00453-5

Neuman, S. B. (2006) *The knowledge gap: Implications for early education.* In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 29-40). The Guilford Press.

Nicholas, C. (2010) *Orang Asli: Rights, problems, solutions*. Human Rights Commission of Malaysia (SUHAKAM).

Nicholas, C. (2021). Malaysia. In D. Marno (Ed), *The indigenous world 2021* (35th Ed) (pp. 256-265). The International Work Group for Indigenous Affairs (IWGIA).

Nicholas, H. (2005). Introduction: Putting industrialization back into development. *Development And Change*, *36*(6), 1031-1033.

Nicholas, C. (2004, August 20). Stories of a People – Asserting Place and Presence via Orang Asli Oral Tradition. *Center for Orang Asli Concerns*. http://www.coac.org.my/main.php?section=articles&article_id=18

Ockenden, L. (2014). *Positive learning environments for Indigenous children and young people.* Closing the Gap Clearinghouse.

Pratt, B., & Vries, J. (2018). Community engagement in global health research that advances health equity. *Bioethics*, *32*(7), 454-463. https://doi.org/10.1111/bioe.12465

Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., & Benholz, C. (2018). Language proficiency and mathematics achievement. *Journal Für Mathematik-Didaktik*, *39*(S1), 1-26. https://doi.org/10.1007/s13138-018-0126-3

Renganathan, S. (2013). 'A Pet Rabbit, Bunny': Teaching English to Orang Asli children, an indigenous community in Malaysia. *International Proceedings of Economics Development and Research (IPEDR), 68*(25), 145-151.

Roberts, J., Jergens, J., & Burchinal, M. (2005). The role of home literacy practices in preschool children's language and emergent literacy skills. *Journal Of Speech, Language, And Hearing Research*, *48*(2), 345-359. https://doi.org/10.1044/1092-4388(2005/024)

Romero-Little, M. (2010). How should young indigenous children be prepared for learning? A vision of early childhood education for indigenous children. *Journal of American Indian Education, 49*(1/2), 7-27.

Rosnon, M., & Talib, M. (2019). Indigenous education rights: The Malaysian case. *International Journal of Academic Research in Business and Social Sciences*, *9*(10), 149-167. https://doi.org/10.6007/ijarbss/v9-i10/6470

Sato, D. (2019). Indigenous land rights and the marginalization of the Orang Asli in Malaysia. *Lehigh University*, *37*(09), 56-64.

Sawalludin, A. F., Lim, J. M., & Ishar, M. I. M. (2020). The struggle of Orang Asli in education: Quality of education. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 5*(1), 46-51. https://doi.org/10.47405/mjssh.v5i1.346

Sénéchal, M. (2006). Testing the home literacy model: Parent involvement in kindergarten is differentially related to grade 4 reading comprehension, fluency, spelling, and reading for pleasure. *Scientific studies of reading*, *10*(1), 59-87. https://doi.org/10.1207/s1532799xssr1001_4

Shay, M., & Sarra, G. (2021). Locating the voices of indigenous young people on identity in Australia: An indigenist analysis. *Diaspora, Indigenous, And Minority Education, 15*(3), 166-179. https://doi.org/10.1080/15595692.2021.1907330

Siekmann, S., Webster, J., Samson, S., & Moses, C. (2017). Teaching our way of life through our language: Materials development for Indigenous immersion education. *Cogent Education*, *4*(1), Article 1362887. https://doi.org/10.1080/2331186x.2017.1362887

Valdés, G. (1996). *Con respeto* [With respect]. Teachers College.

Vallori, A. B. (2014). Meaningful learning in practice. *Journal of Education and Human Development, 3*(4), 199-209. https://doi.org/10.15640/jehd.v3n4a18.

Wan, Y. S. (2020). *Education policies in overcoming barriers faced by Orang Asli children: Education for all.* IDEAS Policy Research Berhad (IDEAS).

Whitehurst, G. J., & Lonigan, C. J. (2002). Emergent literacy: Development from prereaders to readers. In S. B. Newman & D. K. Dickinson (Eds), *Handbook of early literacy research* (pp. 11-29)*.* Guilford Press.

Wilson, R., Dumuid, D., Olds, T., & Evans, J. (2020). Lifestyle clusters and academic achievement in Australian Indigenous children: Empirical findings and discussion of ecological levers for closing the gap. *SSM - Population Health*, *10*, Article 100535. https://doi.org/10.1016/j.ssmph.2019.100535

Yin, R. (2018). *Case study research and applications*. SAGE Publication.

Zittoun, T., & Brinkmann, S. (2012). Learning as meaning making. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1809-1811). Springer. https://doi.org/10.1007/978-1-4419-1428-6_1851

REFEREES FOR THE PERTANIKA
JOURNAL OF SOCIAL SCIENCES AND HUMANITIES

VOL. 29 (S3) 2021

AUC   : The American University in Cairo

AUT   : Auckland University of Technology

UEL   : University of Economics and Law (Vietnam
             National University)

UKM   : Universiti Kebangsaan Malaysia

UM   : Universiti Malaya

UMP   : Universiti Malaysia Pahang

UniSZA   : Universiti Sultan Zainal Abidin

UNITEN   : Universiti Tenaga Nasional

UPM   : Universiti Putra Malaysia

UPSI   : Universiti Pendidikan Sultan Idris

UQ   : The University of Queensland

USM   : Universiti Sains Malaysia

UTeM   : Universiti Teknikal Malaysia Melaka

UTHM   : Universiti Tun Hussein Onn Malaysia

UTM   : Universiti Teknologi Malaysia

UUM   : Universiti Utara Malaysia

Pertanika Journal of Social Sciences & Humanities

Our goal is to bring high-quality research to the widest possible audience

INSTRUCTIONS TO AUTHORS

(SPECIAL ISSUE)

(Manuscript Preparation & Submission Guide)

Revised: December 2020

Please read the *Pertanika* guidelines and follow these instructions carefully. The Chief Executive Editor reserves the right to return manuscripts that are not prepared in accordance with these guidelines.

MANUSCRIPT PREPARATION

Manuscript Types

Pertanika accepts submission of regular articles for peer-review.

Regular article

Regular article is a full-length original empirical investigation, consisting of introduction, methods, results, and discussion. Original research work should present new and significant findings that contribute to the advancement of the research area. *Analysis and Discussion* must be supported with relevant references.

Size: Generally, each manuscript is **not to exceed 6000 words** (excluding the abstract, references, tables, and/ or figures)**,** a maximum of **80 references**, and **an abstract of less than 250 words.**

Language Accuracy

Pertanika **emphasises** on the linguistic accuracy of every manuscript published. Articles can be written in **English** or **Bahasa Malaysia** and they must be competently written and presented in clear and concise grammatical English/ Bahasa Malaysia. Contributors are strongly advised to have the manuscript checked by a colleague with ample experience in writing English manuscripts or a competent English language editor. For articles in Bahasa Malaysia, **the title, abstract, and keywords should be written in both English and Bahasa Malaysia.**

Author(s) **may be required to provide a certificate** confirming that their manuscripts have been adequately edited. **All editing costs must be borne by the authors**.

Linguistically hopeless manuscripts will be rejected straightaway (e.g., when the language is so poor that one cannot be sure of what the authors are really trying to say). This process, taken by authors before submission, will greatly facilitate reviewing, and thus, publication.

Similarity Index

All articles received must undergo the initial screening for originality before being sent for peer-review. The Guest Editor should check all the manuscripts for possible plagiarism using Turn-It-In before sending them out for review. *Pertanika* does not accept any article with similarity index **exceeding 20%.**

MANUSCRIPT FORMAT

The paper should be submitted in **one-column format** with 1.5 line spacing throughout. Authors are advised to use Times New Roman 12-point font and *MS Word* format.

1. Manuscript Structure

The manuscripts, in general, should be organised in the following order:

Page 1: Running title

This page should **only** contain the running title of your paper. The running title is an abbreviated title used as the running head on every page of the manuscript. The running title **should not exceed 60 characters, counting letters and spaces.**

Page 2: Author(s) and Corresponding author's information

General information: This page should contain the **full title** of your paper **not exceeding 25 words**,

with the name of all the authors, institutions and corresponding author's name, institution and full address (Street address, telephone number (including extension), handphone number, and e-mail address) for editorial correspondence. **The corresponding author must be clearly indicated with a superscripted asterisk symbol (*).**

Authors' name: The names of the authors should be named **in full without academic titles.** For Asian (Chinese, Korean, Japanese, Vietnamese), please write first name and middle name before surname (family name). The last name in the sequence is considered the surname.

Authors' addresses: Multiple authors with different addresses must indicate their respective addresses separately by superscript numbers.

Tables/figures list: A **list** of the number of **black and white/colour figures and tables** should also be indicated on this page. See *"5. Figures & Photographs"* for details.

Example (page 2):

The Mediating Role of Cognitive Emotion Regulation Strategies in the Development of Social Behavior among Adolescents

Samsilah Roslan[1]*, Noorhayati Zakaria[2], Siaw Yan-Li[3] and Noorlila Ahmad[1]

[1]Department of Foundations of Education, Faculty of Educational, Studies, Universiti Putra Malaysia, 43400 Serdang, Malaysia
[2]Politeknik Banting, 42700 Banting, Selangor, Malaysia
[3]Department of Educational Psychology and Counseling, Faculty of Education, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

E-mail addresses:
samsilah@upm.edu.my (Samsilah Roslan)
znorhayati@polibanting.edu.my (Noorhayati Zakaria)
yanli@um.edu.my (Siaw Yan-Li)
noorlila_ahmad@yahoo.com (Noorlila Ahmad)
*Corresponding author

List of Table/Figure: Table 1.
Figure 1.

Page 3: Abstract

This page should **repeat** the **full title** of your paper with only the ***Abstract***, usually in one paragraph and ***Keywords.***

Keywords: **Not more than 8 keywords in alphabetical order** *must be provided to describe the content of the manuscript.*

For articles in Bahasa Malaysia, **the title, abstract and keywords should be written in both English and Bahasa Malaysia.**

Page 4: Text

A regular paper should be prepared with the headings *Introduction, Materials and Methods, Results and Discussions, Conclusions, Acknowledgements, References,* and *Supplementary data* (if any) in this order. The literature review may be part of or separated from the *Introduction*.

2. Levels of Heading

Level of heading	Format
1st	**LEFT, BOLD, UPPERCASE**
2nd	**Flush left, Bold, Capitalise each word**
3rd	**Bold, Capitalise each word, ending with .**
4th	***Bold italic, Capitalise each word, ending with .***

3. Equations and Formulae

These must be set up clearly and should be typed double-spaced. Numbers identifying equations should be in square brackets and placed on the right margin of the text.

4. Tables

- All tables should be prepared in a form consistent with recent issues of *Pertanika* and should be numbered consecutively with Roman numerals (Table 1, Table 2).
- A brief title should be provided, which should be shown at the top of each table (APA format):

  Example:

  Table 1
  Reliability of subscales of the job satisfaction questionnaire
- Explanatory material should be given in the table legends and footnotes.
- Each table should be prepared on a new page, embedded in the manuscript.
- Authors are advised to keep backup files of all tables.

** Please submit all tables in Microsoft word format only, because tables submitted as image data cannot be edited for publication and are usually in low-resolution.

5. Figures & Photographs

- Submit an original figure or photograph.
- Line drawings must be clear, with a high black and white contrast.
- Each figure or photograph should be prepared on a new page, embedded in the manuscript for reviewing to keep the file of the manuscript under 5 MB.
- These should be numbered consecutively with Roman numerals (Figure 1, Figure 2).
- Provide a brief title, which should be shown at the bottom of each table **(APA format):**

  Example: *Figure 1*. Mean willingness to pay (WTP) for a clean environment

- If a figure has been previously published, acknowledge the original source, and submit written permission from the copyright holder to reproduce the material.

- Authors are advised to keep backup files of all figures.

** Figures or photographs must also be submitted separately as TIFF or JPEG, because figures or photographs submitted in low-resolution embedded in the manuscript cannot be accepted for publication. For electronic figures, create your figures using applications that are capable of preparing high-resolution TIFF files.

6. Acknowledgement

Any individuals and entities who have contributed to the research should be acknowledged appropriately.

7. References

References begin on their own page and are listed in alphabetical order by the first author's last name. Only references cited within the text should be included. All references should be in 12-point font and double-spaced. If a Digital Object Identifier (DOI) is listed on a print or electronic source, **it is required to include the DOI in the reference list.** Use Crossref to find a DOI using author and title information.

NOTE: When formatting your references, please follow the **APA-reference style** (7th edition) (refer to the examples). Ensure that the references are strictly in the journal's prescribed style, failing which your article will **not be accepted for peer-review**. You may refer to the *Publication Manual of the American Psychological Association* (https://apastyle.apa.org/) for further details.

Examples of reference style are given below:

Books		
	Insertion in text	**In reference list**
Book/E-Book with 1-2 authors	**Information prominent' (the author's name is within parentheses):** … (Bales, 2017) … (Mahat & Ali, 2020) … Or **'Author prominent' (the author's name is outside the parentheses):** Bales (2017)... Mahat and Ali (2020) …	Bales, S. (2017). *Social justice and library work: A guide to theory and practice.* Chandos Publishing. https://doi.org/10.1016/C2016-0-00181-X Mahat, F., & Ali, N. A. (2020). *Fundamental of Islamic finance.* UPM Press.
Book/E-Book with 3 or more authors	*For all in-text references, list only the first author's family name and followed by* 'et al.' **Information prominent' (the author's name is within parentheses):** … (Meera et al., 2012) … Or **'Author prominent' (the author's name is outside the parentheses):** Meera et al. (2012) …	Meera. N., Ampofo-Boateng, K., & Abd Latif, R. (2012). *Coaching athletes with disabilities*. UPM Press.
Book/E-Book with more than 20 authors		For books with more than 20 authors, please follow the guidelines for journal articles with more than 20 authors.
Chapter in an edited Book/E-Book	**Inform ation pr ominent' (the author 's name is within parentheses):** … (Nagamine et al., 2018) … … (Van der Port, 2015) … Or **'Author prominent' (the author's name is outside the parentheses):** Nagamine et al. (2018) … Van der Port (2015) …	Nagamine, T., Fujieda, Y., & Iida, A. (2018) The role of emotions in reflective teaching in second language classrooms: Felt sense, emotionality, and practical knowledge acquisition. In J. Martínez Agudo (Ed.), *Emotions in second language teaching* (pp. 145-163). Springer, Cham. https://doi.org/10.1007/978-3-319-75438-3_9 Van de Port, M. (2015). Reading Bruno Latour in Bahia. In M. Jackson & A. Piette (Eds.), *What is existential anthropology?* Berghahn Books.

	Insertion in text	In reference list
Editor	**Information prominent' (the author's name is within parentheses):** … (Machado & Davim, 2014) … … (Sheldon & Turner-Vorbeck, 2019) … Or **'Author prominent' (the author's name is outside the parentheses):** Machado and Davim (2014) … Sheldon and Turner-Vorbeck (2019) ...	Machado, C., & Davim, J. P. (Eds). (2014). *Work organization and human resource management.* Springer, Cham. https://doi.org/10.1007/978-3-319-06376-8 Sheldon, S. B., & Turner-Vorbeck, T. A. (Eds.). (2019). The Wiley handbook of family, school, and community relationships in education. New Jersey, USA: John Wiley & Sons.
Several works by the same author in the same year	**Information prominent' (the author's name is within parentheses):** … (Camilleri, 2018a, 2018b) … Or **'Author prominent' (the author's name is outside the parentheses):** Camilleri (2018a, 2018b) …	Camilleri, M. A. (2018a). *Travel marketing, tourism economics and the airline product: An introduction to theory and practice.* Springer, Cham. https://doi.org/10.1007/978-3-319-49849-2 Camilleri, M. A. (2018b). Understanding customer needs and wants. In *Travel marketing, tourism economics and the airline product: An introduction to theory and practice* (pp. 29-50). Springer, Cham. https://doi.org/10.1007/978-3-319-49849-2_2
	Journals	
Journal article with 1-2 authors	**Information prominent' (the author's name is within parentheses):** … (Al-Shboul & Maros, 2020) … Or **'Author prominent' (the author's name is outside the parentheses):** Al-Shboul and Maros (2020) …	Al-Shboul, Y., & Maros, M. (2020). The high and low-context communication styles in refusal strategies by Jordanian Arabic and American English speakers. *Pertanika Journal of Social Sciences and Humanities, 28*(3), 2063-2080.
Journal article with 3 or more authors	*For all in-text references, list only the first author's family name and followed by 'et al.'* **Information prominent' (the author's name is within parentheses):** … (Khajouei et al., 2018) … … (Yusop et al., 2020) … Or **'Author prominent' (the author's name is outside the parentheses):** Khajouei et al. (2018) … Yusop et al. (2020) …	Khajouei, R., Abbasi, R., & Mirzaee, M. (2018). Errors and causes of communication failures from hospital information systems to electronic health record: A record-review study. *International Journal of Medical Informatics, 119*(January), 47-53. https://doi.org/10.1016/j.ijmedinf.2018.09.004 Yusop, F. D., Ab Ghaffar, F., Danaee, M., Firdaus, A., Hamzaid, M. A., Abu Hassan, Z. F., Senom, F., Ebrahim, N. A., Bonn, B. Y., & Chen, Y. M. (2020). Two decades of research on early career faculties (ECFs): A bibliometric analysis of trends across regions. *Pertanika Journal of Social Sciences and Humanities, 28*(1), 325-342.
Journal article with more than 20	**Information prominent' (the author's name is within parentheses):** … (Tobler et al., 2017) … Or **'Author prominent' (the author's name is outside the parentheses):** Tobler et al. (2017) …	Tobler, R., Rohrlach, A., Soubrier, J., Bover, P., Llamas, B., Tuke, J., Bean, N., Abdullah-Highfold, A., Agius, S., O'Donoghue, A., O'Loughlin, I., Sutton, P., Zilio, F., Walshe, K., Williams, A. N., Turney, C. S. M., Williams, M., Richards, S. M., Mitchell, N. ... Cooper, A. (2017). *Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. Nature, 544*(7649), 180-184. https://doi.org/10.1038/nature21416
Journal article with an article number	**Information prominent' (the author's name is within parentheses):** … (De Rubeis et al., 2017) … Or **'Author prominent' (the author's name is outside the parentheses):** De Rubeis et al. (2017) …	De Rubeis, J., Lugo, R. G., Witthöft, M., Sütterlin, S., Pawelzik, M. R., & Vögele, C. (2017). Rejection sensitivity as a vulnerability marker for depressive symptom deterioration in men. *PloS One, 12*(10), Article e0185802. doi: 10.1371/journal.pone.0185802
Journal article with missing information	**Information prominent' (the author's name is within parentheses):** … (Bajaj et al., 2014) … … (Jdaitawi, 2015) … … (Nastasa & Farcas, 2015) …	**Missing volume number** Bajaj, G., Deepa, N., Bhat, J. S., D'Souza, D., & Sheth, P. (2014). Self-efficacy and verbal fluency — does age play a role? *Healthy Aging & Clinical Care in the Elderly,* (6), 17-24. http://dx.doi.org/10.4137/HACCE.S14292

	Insertion in text	In reference list
Journal article with missing information	Or **'Author prominent' (the author's name is outside the parentheses):** Bajaj et al. (2014) … Jdaitawi (2015) … Nastasa and Farcas (2015) …	**Missing issue number** Nastasa, L. E., & Farcas, A. D. (2015). The effect of emotional intelligence on burnout in healthcare professionals. *Procedia - Social and Behavioral Sciences, 187*, 78-82. **Missing page or article number** Jdaitawi, M. (2015). Social connectedness, academic, non-academic behaviors related to self-regulation among university students in Saudi Arabia. *International Education Studies, 8*(2). https://doi.org/10.5539/ies.v8n2p84
Several works by the same author in the same year	**Information prominent' (the author's name is within parentheses):** … (Ibrahim, 2019a, 2019b) … Or **'Author prominent' (the author's name is outside the parentheses):** Ibrahim (2019a, 2019b) …	Ibrahim, M. H. (2019a). Capital regulation and Islamic banking performance: panel evidence. *Bulletin of Monetary Economics and Banking, 22*(1), 47-68. Ibrahim, M. H. (2019b). Oil and macro-financial linkages: Evidence from the GCC countries. *The Quarterly Review of Economics and Finance, 72*(May) 1-13. https://doi.org/10.1016/j.qref.2019.01.014
Newspaper		
Newspaper article – with an author	… (Davidson, 2018) ... Or ... Davidson (2018) ...	Davidson, J. (2018, January 9). CES 2018: Samsung vows to add artificial intelligence to everything it does. *Australian Financial Review.* https://www.afr.com/technology/ces-2018-samsung-vows-to-add-artificial-intelligence-to-everything-it-does-20180109-h0fdtd
Newspaper article – without an author	("Economics nudging," 2017). OR "Economics nudging" (2017) … Use a shortened title (or full title if it is short) in Headline Case enclosed in double quotation marks.	Economics nudging people away from war. (2017, December 16). *The Age*, 33.
Dissertation/Thesis		
Published Dissertation or Thesis References	… (Solomon, 2016) ... Or … Solomon (2016) ...	Solomon, M. (2016). *Social media and self-evaluation: The examination of social media use on identity, social comparison, and self-esteem in young female adults* [Doctoral dissertation, William James College]. ProQuest Dissertations Publishing. https://search.proquest.com/ openview/7d66a63f277a84a64907db68fff991ba/1?pq-origsite=gscholar&cbl=18750&diss=y
Unpublished Dissertation or Thesis References	… (Curry, 2016) ... Or … Curry (2016) ...	Curry, J. (2016). *A guide to educating single mothers about early gang intervention and prevention* (Unpublished Master's thesis). Pacific Oaks College.
Conference/Seminar Papers		
Conference proceedings published in a journal	… (Chaudhuri et al., 2017) … Or ... Chaudhuri et al. (2017) …	Chaudhuri, S., & Biswas, A. (2017). External terms-of-trade and labor market imperfections in developing countries: Theory and evidence. *Proceedings of the Academy of Economics and Economic Education, 20*(1), 11-16. https://search-proquest-com.elibrary.jcu.edu.au/docview/1928612180?accountid=16285
Conference proceedings published as a book chapter	… (Morgan et al., 2017) … Or ... Morgan et al. (2017) …	Morgan, R., Meldrum, K., Bryan, S., Mathiesen, B., Yakob, N., Esa, N., & Ziden, A. A. (2017). Embedding digital literacies in curricula: Australian and Malaysian experiences. In G. B. Teh & S. C. Choy (Eds.), *Empowering 21st century learners through holistic and enterprising learning: Selected papers from Tunku Abdul Rahman University College International Conference 2016* (pp. 11-19). Springer. https://doi.org/10.1007/978-981-10-4241-6_2

	Insertion in text	In reference list
Online	… (McDonald et al., 2019) … Or ... McDonald et al. (2019) …	McDonald, E., Manessis, R., & Blanksby, T. (2019, July 7-10). *Peer mentoring in nursing - Improving retention, enhancing education* [Poster presentation]. STARS 2019 Conference, Melbourne, Australia. https://unistars.org/ papers/STARS2019/P30-POSTER.pdf
Government Publications		
Government as author	**First in-text reference: Spell out the full name with the abbreviation of the body.** … U.S. Department of Housing and Urban Development (2020) … Or … (U.S. Department of Housing and Urban Development, 2020) …	U.S. Department of Housing and Urban Development. (2020). *National comprehensive housing market analysis.* https://www.huduser.gov/portal//publications/pdf/National-CHMA-20.pdf

8. General Guidelines

Abbreviations: Define alphabetically, other than abbreviations that can be used without definition. Words or phrases that are abbreviated in the *Introduction* and following text should be written out in full the first time that they appear in the text, with each abbreviated form in parenthesis. Include the common name or scientific name, or both, of animal and plant materials.

Authors' Affiliation: The primary affiliation for each author should be the institution where the majority of their work was done. If an author has subsequently moved to another institution, the current address may also be stated in the footer.

Co-Authors: The commonly accepted guideline for authorship is that one must have substantially contributed to the development of the paper and share accountability for the results. Researchers should decide who will be an author and what order they will be listed depending upon their order of importance to the study. Other contributions should be cited in the manuscript's *Acknowledgements*.

Copyright Permissions: Authors should seek necessary permissions for quotations, artwork, boxes or tables taken from other publications or other freely available sources on the Internet before submission to *Pertanika*. The *Acknowledgement* must be given to the original source in the illustration legend, in a table footnote, or at the end of the quotation.

Footnotes: Current addresses of authors if different from heading may be inserted here.

Page Numbering: Every page of the manuscript, including the title page, references, and tables should be numbered.

Spelling: The journal uses American or British spelling and authors may follow the latest edition of the Oxford Advanced Learner's Dictionary for British spellings. Each manuscript should follow one type of spelling only.

SUBMISSION OF MANUSCRIPTS

1. MANUSCRIPT:

Ensure your manuscript has followed the *Pertanika* style particularly the first-4-pages as explained earlier. The article should be written in a good academic style and provide an accurate and succinct description of the contents ensuring that grammar and spelling errors have been corrected before submission. It should also not exceed the suggested length.

2. DECLARATION FORM:

The corresponding author has to sign a declaration form. In signing the form, authors declare that the work submitted for publication is original, previously unpublished, and not under consideration for any publication elsewhere.

Note:
COPYRIGHT FORM: Author will be asked to sign a copyright form when the paper is accepted. In signing the form, it is assumed that authors have obtained permission to use any copyrighted or previously published material. All authors must read and agree to the conditions outlined in the form and must sign the form or agree that the corresponding author can sign on their behalf. Articles cannot be published until a signed form (original pen-to-paper signature) has been received.

Visit our Journal's website for more details at http://www.pertanika.upm.edu.my/

ACCESS TO PUBLISHED MATERIALS

Under the journal's open access initiative, authors can choose to download free material (via PDF link) from any of the journal issues from *Pertanika*'s website. Under "**Browse Journals**" of each journal, you will see a link, "*Regular Issue*", "*Special Issue"* or "*Archives*". Here you will get access to all current and back-issues from 1978 onwards. No hard copy of journals or offprints are printed.

Any queries may be directed to the **Chief Executive Editor's** office via email to
executive_editor.pertanika@upm.edu.my

Contents