Pertanika J. Soc. Sci. & Hum. 29 (S3): 1 - 25 (2021)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

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

ARTICLE INFO

Article history:
Received: 16 July 2021
Accepted: 04 October 2021
Published: 30 November 2021

DOI: https://doi.org/10.47836/pjssh.29.S3.01

E-mail addresses:
vahid@upm.edu.my, nimechie@gmail.com (Vahid Nimehchisalem)
jayakaranmukundan@yahoo.com (Jayakaran Mukundan)
shameemkhan@segi.edu.my (Shameem Rafik-Galea)
arshad@upm.edu.my (Arshad Abd Samad)
*Corresponding author

ISSN: 0128-7702
e-ISSN: 2231-8534

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 = 0.60-0.65) 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

We would like to thank the students and raters for their support in data collection. We also express our gratitude to the journal editors, guest editors, and blind reviewers who helped us improve the quality of our final manuscript.

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						