Pertanika J. Soc. Sci. & Hum. 21 (3): 1149 - 1162 (2013)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

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

Examining Factors Affecting Language Performance: A Comparison of Three Measurement Approaches

Kadeessa Abdul Kadir

Language Programme, Centre for Research and Consultancy Management, National Institute of Public Administration, Public Service Department, Malaysia, Jalan Ilmu, 59100 Kuala Lumpur, Malaysia

ABSTRACT

This article presents the findings of three approaches, namely, Classical Test Theory, Generalizability Theory (Brennan, 2001; Shavelson & Webb, 1991) and Multi-facet Rasch Analysis (Linacre, 2004; McNamara, 1996) in examining the effects of different factors on language test performance. Through these approaches, the investigator sought to determine the extent test takers, raters, and test tasks contribute to the source of variance in performance on writing. Additionally, the investigator also examined whether raters' severity and task difficulty affect the reliability and dependability of the observed language performance scores. The purpose of using these measurement approaches was to determine whether findings could be integrated so as to strengthen validity arguments for test score dependability and generalizability.

Keywords: Classical test theory, generalizability theory, multi-facet rasch analysis, score reliability and dependability, writing assessment

INTRODUCTION

Measurement Approaches for Reliability and Score Dependability

The use of measurement approaches such as Classical Test Theory (CTT), Generalizability Theory (G-Theory) (Brennan, 2011; Shavelson & Webb, 1991) and Multi-facet Rasch Analysis (Linacre, 2004; McNamara, 1996) in estimating the effects of different factors on test performance has been explored in many in language testing situations. This is evident in the numerous studies reported in previous literature. In performance-based language testing, the reliability of ratings is a major issue in ensuring the dependability of scores. This dependability of scores can be attributed to many factors

ARTICLE INFO

Article history:
Received: 30 October 2011
Accepted: 28 August 2012

E-mail address:
kadeessa@gmail.com (Kadeessa Abdul Kadir)

ISSN: 0128-7702

such as the ability of the test-takers, the difficulty of the test-tasks, the reliability or severity of the raters and others. Looking at the complementary roles of the different measurement approaches, the study opted for the use of the different approaches to strengthen the validation argument put forward for test score reliability and dependability.

Classical Test Theory

Classical Test Theory (or CTT), a body of theory that rests its foundation that variation in performance is decomposed only in terms of true and observed scores and any other source of variation, is simply labelled as "error." The theoretical definition of reliability (which refers to the proportion of true score variance to that of observed scores) is further operationalised as the correlation between two sets of parallel tests providing the basis for all estimates of reliability (Lord & Novick, 1968). Hence, the overriding concern of CTT is to cope effectively with random error portion (E) of the raw score. The less random error in the measure, the more the raw score reflects the true score and hence, increases reliability and score dependability.

In CTT, several types of reliability estimates can be used depending on whether the tests are norm or criterion referenced tests. The first, which is the Cronbach's Alpha, is more suitable for norm-referenced tests for dichotomous test data. The other is more appropriate for criterion-referenced tests (Bachman, 2004; Lynch, 2003). In a criterion-referenced performance-based testing, this is achieved through ensuring consistency across tasks, raters, and occasions as these variables can affect the validity of inference. More importantly, rater characteristics such as consistency or severity in ratings as well as task variability have been researched as the major sources of measurement errors in the context of performance-based assessment (Bachman, Lynch & Mason, 1995; Fulcher, 2003; Lynch & McNamara, 1998).

Various methods have been proposed to identify and quantify the extent of disagreement between raters and to reduce it to acceptable levels through proper rater training and monitoring procedures (McNamara, 1996). McNamara further suggests that fairness in rating practices can be enhanced by rating scales which are defined and described carefully for different levels of the scales, having raters who have been trained and can demonstrate the required level of agreement, and, finally, by the practice of double ratings to check for inter-rater consistency as well as agreement between raters.

Generalizability Theory (G-theory)

G-theory extends CTT by allowing researchers to further decompose and estimate the variance associated with the errors, and therefore, address the issues of score generalizability and dependability. Multiple sources of errors in a measurement can be estimated separately in a single analysis and allow decision makers to determine the number of occasions, test forms, and raters required in order to

maintain a dependable score through G and D studies. Findings from the generalizability analysis or GENOVA will provide evidence in the form of the magnitude that the different facets contribute to the source of variance in a test.

Rasch Multi-Facet measurement (FACETS)

Rasch Multi-Facet measurement is an extension of the Rasch-model, a one-parameter Item Response theory (IRT) model which was traditionally used for the analysis of multiple-choice examinations where the parameters involved are the difficulty of the test items and the ability of the examinees. Meanwhile, estimates of each examinee's ability and each item's difficulty are reported on a common log-linear scale. The probability of a correct response to an item is simply a function of the difference between examinee's ability and item's difficulty. Now, Multi-Facet Rasch analysis provides the capability to model additional facets of interest making it particularly useful for analysis of subjectively rated performance tasks. Using this method, the chances of success on a performance task are related to a number of aspects of the performance setting itself. These aspects (i.e., facets) include the test taker's ability, the difficulty of the performance task, and the characteristics of the raters themselves (i.e., rater severity/leniency). These facets are related to each other as either increasing or decreasing the likelihood of a test taker of given ability achieving a given score on a particular task.

RESEARCH QUESTIONS

The purpose of this paper is to present the findings of using three different approaches in examining the relative effects of persons, items, raters and test tasks on the writing component of the English Language Proficiency Assessment (ELPA) assessment. To that end, the following three research questions were formulated:

1. What are the inter-rater reliability estimates according to CTT across two writing tasks?
2. To what extent do test takers, raters, and test tasks contribute to the source of variance in writing performance?
3. To what extent do raters' severity and task difficulty affect the reliability and dependability of the observed scores on the writing tasks?

METHODS AND MATERIALS

The Test

The English Language Proficiency Assessment (ELPA) was developed for the Malaysian Public Service. This assessment, which was introduced in 1998, has been used to assess the proficiency of officers in the Public Service from various schemes. Test scores were used to determine follow-up training as well as job placement within the service.

The assessment consists of reading, writing and speaking components. For the purpose of the current study, only the writing component was addressed. The writing component consists of two tasks. For task 1, the officers are required to write a formal

letter and for task 2 a formal report. Both tasks are rated on a six criteria analytical rating scale comprising task fulfilment, organization, grammar, vocabulary, style, and mechanics. A composite of their writing ability score, ranging from bands 1 to 5, is obtained by averaging the scores on both the tasks.

Test Data

The data for the analysis included selection of random writing scripts from previously administered ELPA. Eighty sets of the writing scripts for both the writing tasks were used for the analysis. From a pool of 7 raters, 3 raters per group were randomly assigned to re-score each of the writing tasks. The rationale for having a minimum of 3 raters was to ensure that the data matrix conformed to the Generalizability as well as Rasch models, while better estimates could also be derived from the analyses (McNamara, 1996). To maintain consistency, the original ratings for both the writing tasks were excluded from the analyses as they were rated by different individuals.

Analysis

As explained in the earlier part of the paper, three different approaches were used to analyse the test data. For the CTT approach, due the criterion-referenced nature of the writing tasks and therefore subjectively scored, Pearson's correlation coefficient (r) was used to determine inter-rater reliability. Since the writing tasks used a criterion-referenced scale, the estimates of inter-rater reliability are not as straightforward as determining the internal consistency of test items. The procedures included getting the Pearson's correlation between the pairs of raters and transforming them using Fisher's z. The values obtained from the Fisher's z were then converted back to Pearson's r. This was done to correct the analytical and holistic rating scale, which reflected a more ordinal rather than continuous nature of the data. The formula for z transformation is produced as follows:

$$r_{tt} = n\, r_{AB} / 1 + (n - 1)\, R_{AB}$$

Where:

n = number of raters

r_{AB} = the average of the corrected (Fisher's z transformation) correlations between the raters

r_{tt} = the inter-rater reliability estimate, after it is transformed back to Pearson.

For the Generalizability analysis, several different analyses were carried out to determine the magnitude of variance for each of the facets considered in the design. These facets included the following: (a) test tasks (i), (b) raters (r), (c) and test takers (p).

The analyses examined the source of variance when raters, test takers, and tasks or items were included. The results of the G-studies can be used to optimize the number of conditions for each facet as in this case the number of raters so as to obtain the desired index of generalizability (reliability). The two indexes of reliability used in G-theory are: (a) G or generalizability

coefficient (ρ) and (b) dependability or phi coefficient (φ). The G coefficient is used for decisions based on relative standing or ranking of individuals whereas the phi coefficient is based on the absolute level of their scores. For this study, since the writing scores were used for both relative and absolute decisions, both coefficients would therefore be reported. For the G and D-studies, the researcher used the GENOVA programme to conduct all the analyses.

Further investigation of the effects of test-takers abilities, task difficulties, and rater-severity using Multi-Facet Rasch analysis was carried out. All the analyses were carried out using FACETS (2007; Linacre, 2004) computer software.

RESULTS AND DISCUSSION

The results of the three different approaches are organized to address the research questions posed at the beginning of the paper, as follows:

1. The first section discusses the classical theory reliability estimates for the writings tasks;
2. The second section presents the G and D-studies to examine the relative contribution to test variance of persons, raters and test tasks and their interactions; and
3. The third section will show whether rater severity and task difficulty affects reliability and test score dependability.

Classical Theory Reliability Estimates

Table 1 reports the inter-rater reliability estimates denoted by the Pearson's r for the writing tasks, before and after transformation for two separate groups of raters. It can be noted that the average inter-rater reliability estimates for group 1 raters was slightly lower than for group 2 (r_{tt1} = .77, r_{tt2} = .80); this means the groups had 59.0 percent and 64.0 percent respectively due to true score variability. These are moderate estimates, especially for a complex performance-based assessment, as reflected in the analytical rating of the writing tasks. Within group 1, some differences existed among the pairs of raters, with R2R3 (r_{23} = .81) showing the highest estimates, while paired rater R1R3 (r_{13} = .72) the lowest. Similar differences in

TABLE 1
Inter-Rater Reliability Estimates for the Writing Tasks

Task 1 (Letter)			Task 2 (Report)		
Paired raters (Group1)	Pearson's *r* (r_{tt1})	[a] Pearson's *r*	Paired raters (Group 2)	Pearson's *r* (r_{tt2})	[b] Pearson's *r*
R1R2	.78		R4R5	.74	
R1R3	.72	.77	R4R6	.78	.80
R2R3	.81		R5R6	.87	

Note. n = 80. [a] Average Inter-Rater reliability for group 1 after z transformation. [b] Average Inter-Rater reliability for group 2 after z transformation.

the paired raters can also be noted for group 2, whereby R5R6 showed the highest inter-rater reliability (r_{56} = .87), and R4R5 had the lowest (r_{45} = .74).

As for the writing tasks, on the whole the estimates of inter- rater reliability (which ranged between .72 and .87) were considered moderate with raters overlapping or agreeing between 52% to 76% percent of the time. The average estimate for the inter-rater reliability for the paired raters was higher on Task 2 (report) than Task 1 (letter). This indicated that on average, the raters agreed at least 61% on their ratings. Another interesting point is that the estimates of reliability were higher on the longer task (i.e. task 2), suggesting that raters performed better or were more accurate in their judgment if they were provided with longer test-taker responses. Although this assertion needs further study, the findings thus far do suggest this.

Such is the only conclusion one can make when using the CTT. The use of correlations as estimates of reliability has come under scrutiny for it is sample-dependent. Weir (2005) and Fulcher (2003) warned that certain rater behaviour such as the extent to which the raters actually use the entire range of the rating scale as well as distortions in the correlation when there exists either very high or low scores cannot be entirely captured by correlations. In addition, even if raters were found to be consistent in their ratings and the inter-rater correlations were high, this still did not reveal information about rater severity or other source of variance. It is also difficult to ascertain as to what is acceptable reliability although in the literature, estimates ranging from .60 to .90 have been described as acceptable (Lynch, 2003). The use of multiple methods, including generalizability analyses as well as multifaceted Rasch analysis, was to factor in the limitations of using only correlations as a measure of rater reliability. These limitations include the inability for CTT to address different sources of error and to distinguish systematic measurement error from random measurement error (Bachman, 2004).

Generalizabilty Theory Analysis (GENOVA)

Findings from the GENOVA reveal a very different kind of information from that of the CTT approach. The generalizability analysis carried out provided a deeper understanding as to what is really at play when assessing language performance. The first finding when average scores of the writing tasks were used as a facet in the *(p x i)*, the GENOVA output as in Table 2 suggests that the largest source of variance was from test takers (90.61%), whereas test tasks had only 0.45%. This small variance component suggests that the two writing tasks remained stable across the test takers and were of the same difficulty level for this particular group. However, the moderately large residual effects suggest a large persons-by-items interaction, unmeasured sources of variation, or both. As in any complex performance based assessment, averaging scores on the writing tasks also provided a reliable estimate of ability for operational

purposes, as reflected in the large test takers variance (90.61%).

When raters are introduced as facets for the analyses, the results of p x r analyses are presented in Tables 3 and 4. The magnitude of variance contributed by raters for both Tasks 1 and 2 is 6.77% and 4.16%, respectively. This suggests that on the whole, the variability in test scores is largely due to test-takers' abilities rather than rater's behaviour. However, the interaction between test-taker and raters seemed quite high, with 21.50% and 20.18% for Task 1 and Task 2, respectively. If average ratings were used for each of the three raters for Task 1 and Task 2 writing variability due to test takers (71.73% and 75.65%) contributed more than rater variance (6.77% and 4.16%) once again. In other words, the groups of raters were quite consistent in their ratings of the writing tasks.

Table 5 presents the results of the D-study for the same number of conditions for the different facets used in the analyses. This was done to compare with the estimate of inter-rater reliability obtained using the CTT approach. As can be noted for the observed data in the table, both the *G* and *phi* coefficients were substantially high for all groups of raters and test tasks.

Given that the writing component is reported using the average score of both tasks, both the *G* and *phi* coefficients for Task 1 were .91 and .88 respectively when three raters were considered. The coefficients dropped to .86 and .83 respectively for two

TABLE 2
Variance Components for p x i Design for the Two Writing Tasks

Source of variance	Variance	Percentage
Persons (*p*)	49.90	90.61
Items (*i*)	0.25	0.45
pi, e	4.92	8.93

TABLE 3
Variance Components for p x r Design for Writing Task 1

Sources of variance	Variance	Percentage
Persons (*p*)	54.05	71.73
Raters (*r*)	5.10	6.77
pr, e	16.20	21.50

TABLE 4
Variance Components for p x r Design for Writing Task 2

Sources of variance	Variance	Percentage
Persons (*p*)	45.40	75.65
Raters (*r*)	2.50	4.16
pr, e	12.11	20.18

raters. The *G* coefficient and *phi* coefficients for the operational single rater setting dropped to .76 and .71, respectively. These values are still considered within the range of acceptability. Similar trends can be noted for Task 2, where the values of the *G* and *phi* coefficient dropped from as high as .92 and .91 with three raters to a moderate .78 and .74 for only one rater. From these results, it is best that double raters are used as opposed to the operational single rater setting that is currently practiced for a moderately high generalizability and dependability of the writing scores.

Multi-facet Rasch Analysis

As in the G-study, raters, test-tasks, and test takers can be viewed as a source of method variance. In Rasch analysis, the effects of the different facets can be further explained at a much deeper level. Fig.1 and Fig.2 provide the relative abilities of the test takers, the severity of the raters, and the relative difficulty of the different criteria (1 = task fulfillment, 2 = organization, 3 = grammar, 4 = vocabulary, 5 = style, and 6 = mechanics) on the two writing tasks. For Task 1, as noted in Fig.1, there was a considerable variation in the abilities of test takers (i.e. ranging approximately between – 6.00 and + 4.00 on the logit scale). The majority of the test takers, however, were placed between 0.00 and + 4.00 on the scale, suggesting that most of them demonstrated average and high performances on Task 1. As for Task 2, the facet map (see Fig.2) shows that for the same group of test takers, the writing ability ranged from -6.00 to + 5.00 on the logit scale, indicating a slightly wider range of abilities. Unlike Task 1, there was more variability in the abilities for the majority of the test takers, as indicated by the clustering between -1.00 and + 5.00. Unlike the wide variation in test-takers' abilities, Fig.1 and Fig.2 show that the estimates for the group of three raters for each of the task seemed to cluster around the mean on the logit scale, with slightly more variability among the raters for Task 2. As for the estimates for the six criteria used for assessing the writing tasks, similar trends can be noted for both tasks.

A more detailed record of rater severity and difficulty estimates of the criteria as well as comparisons between the groups of raters on the two writing tasks are reported in Table 6. As can be noted, the two groups of raters assigned for the two tasks differed

Table 5: Generalizability and Phi Coefficients for Writing

		Number of raters					
		3 Raters		2 Raters		1 Rater	
Type of analysis	Design	ρ	φ	ρ	φ	ρ	φ
Average scores of each rater on Task 1	*p* x *r*	.91	.88	.86	.83	.76	.71
Average scores of each rater on Task 2	*p* x *r*	.92	.91	.88	.88	.78	.74

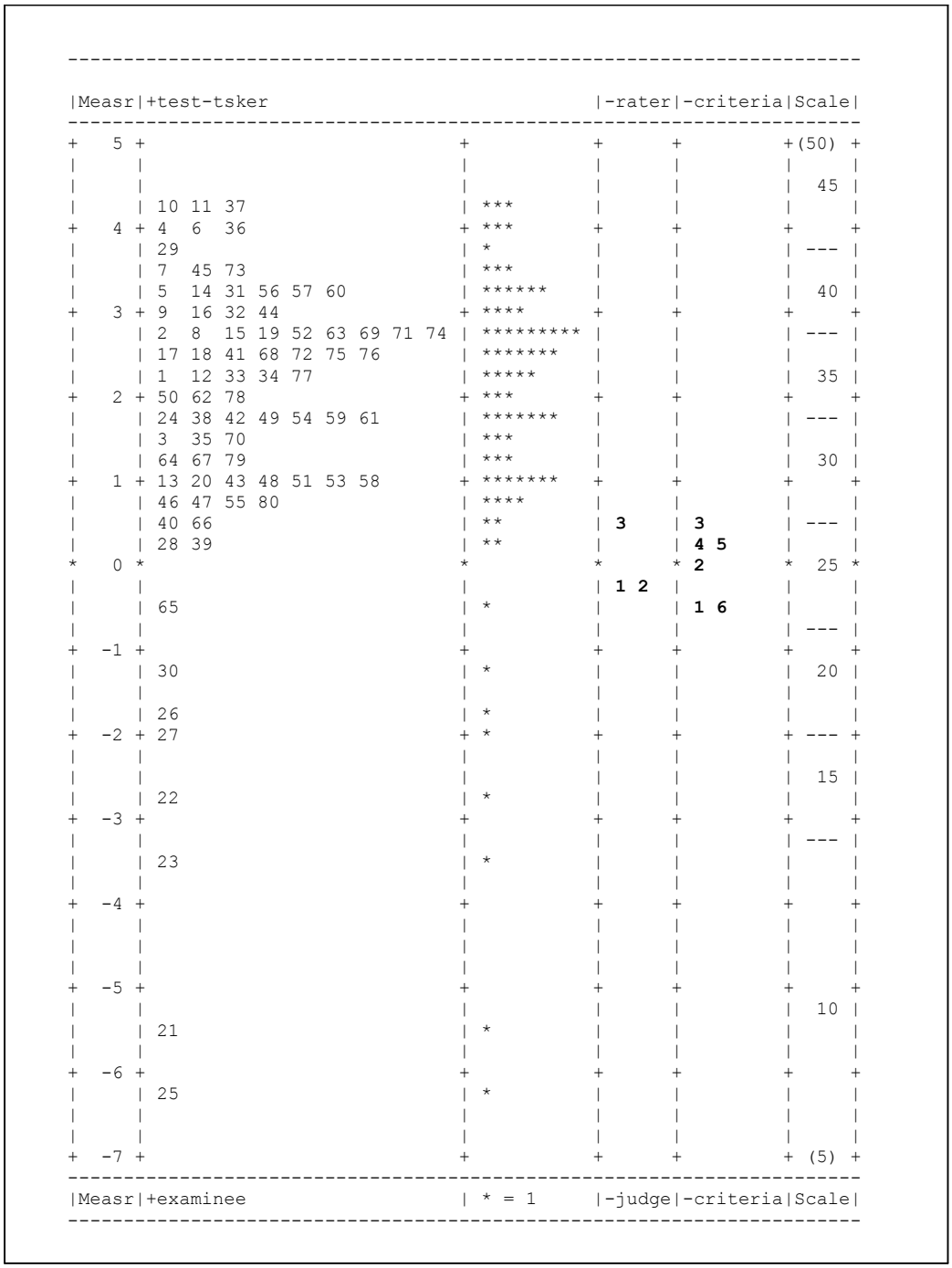

Fig.1: Facet map for writing task 1

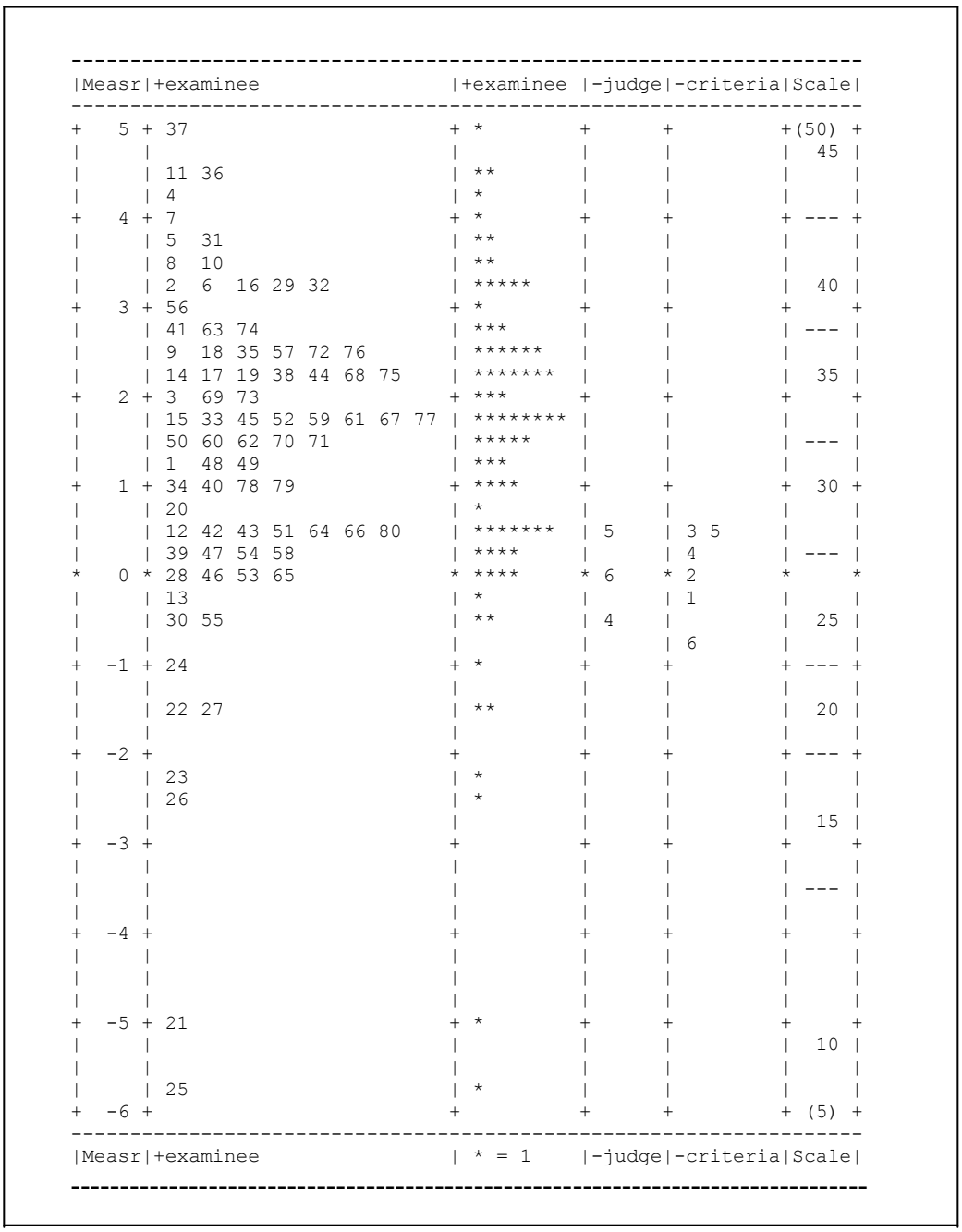

Fig.2: Facet map for writing task 2

very little in their ratings. Both groups had means of .00 and the standard deviation was between.17 and .23, suggesting very little variability in terms of rater severity across the two groups. Comparing both groups of raters across the two tasks, the most moderate raters were Rater 1 and Rater 6, with estimates of -.23 and -.03 on the logit scale, respectively. This changed, however, when the estimates within each group of raters for each task were examined. For Task 1, the estimates of severity ranged between -.37 to .60 on the logit scale. Meanwhile, the reliability of the separation index was .97, indicating that the raters in this group consistently differed from one another. Meaningful variation in harshness did exist among the raters, with the most lenient rater estimating -.37 on the logit scale and the most severe .60. The infit values of the raters were between .85 and .94, suggesting that Rater 2 for Task 1 seemed to be the most consistent rater. Rater 1, on the other hand, seemed to have more variations in her ratings with the infit values of 1.19, whilst raters 2 and 3 varied less in their ratings. However, no raters for this group were identified as misfitting as the infit values were within range of two standard deviations around the mean [.99 ± (.17 x 2 = .34)].

For Task 2, the severity gap was between -.42 to and .44 on the logit scale, which also suggested that for this group, there existed differences with a reliability of the separation index of .98. Rater 5 seemed to be the harshest with an estimate of .44 on the logit scale with Rater 4 being the most lenient. For Task 2, rater 4 seemed to be less consistent as compared to raters 5 and 6 who were closer to the mean infit value of 1.00. Like the raters for Task 1, none was identified as misfitting as the infit values were within the range of two standard deviations around the mean [1.00 ± (.23 x 2 = .46)].

Table 7 reports the difficulty estimates of the six criteria on both tasks. It indicated that for both the tasks, the most leniently scored was mechanics, whereas for Task 1 and Task 2, grammar and style respectively were the most harshly scored. The difficulty span for Task 1 between the most leniently and the most harshly scored criteria was 1.01

TABLE 6
Estimates of Rater Severity for the Writing Tasks

Task 1				Task 2			
Rater	Severity estimate	*SE*	Infit	Rater	Severity estimate	*SE*	Infit
2	-.37	.05	.94	4	-.42	.05	1.25
1	-.23	.05	1.19	6	-.03	.05	.81
3	.60	.05	.85	5	.44	.05	.94
M	.00	.00	.99	*M*	.00	.00	1.00
SD	.53	.00	.17	*SD*	.43	.00	.23

Note. Reliability (not inter-rater) for Task 1=.99. Reliability (not inter-rater) for Task 2= .99. Inter-Rater agreement for Task 1 = 30.5%; Inter-Rater agreement for Task 2 = 30.6%.

on the logit scale, suggesting that there was a considerable difference in scoring of these two criteria. As for Task 2, mechanics was -.83 whereas style was .51 on the logit scale. Both groups of raters, however, showed considerable differences in the way they used the rating scale. Clearly, the groups of raters assigned varied considerably within and across both tasks, with the rater group assigned to rescore Task 2 demonstrating more variation. Mechanics was found to be misfitting on Task 1 as the infit values exceeded the acceptable range of two standard deviations [.99 ± (.15x 2 = .30)], whereas for Task 2, none of the criteria was found to be misfitting [1.00 ± (.13x 2 = .26)].

In summary, the multi-faceted Rasch provided estimates of rater severity on a linear scale as well as fit statistics, which are indicators of rater consistency. Aside from rater severity, two different sets of estimates of difficulty were also provided. The first set was for the six criteria (task fulfilment, organization, grammar, vocabulary, style, and mechanics) of the analytical rating scale and the second were estimates of difficulty for the two speaking tasks. Overall, the findings revealed that both groups had raters who were either considered lenient or severe although the estimates on the logit scale were not that extreme. The second group of raters, who scored in Task 2, was slightly more lenient than those who scored in Task 1. Both exhibited fairly consistent ratings with each group having at least one rater with slightly high infit values. Rater 1 of Task 1 and Rater 4 of Task 2 had the infit values of 1.19 and 1.25 respectively. Rater 4 was also considered the most lenient amongst all the six raters. The most severe rater was Rater 3 who scored Task 1.

The findings also revealed that the raters who scored the writing tasks used the criteria or dimension of the analytical scale differently. Overall, raters were more lenient on the mechanics and task fulfilment than on the other dimensions. The infit values of 1.29 and 1.19 for tasks 1 and 2 respectively were within the acceptable range. As for consistency, it was difficult to see any kind of pattern to suggest whether raters were more or less consistent in their ratings of

TABLE 7
Estimates of Criteria Difficulty for the Writing Tasks

	Task 1			Task 2		
Criteria	Estimate	*SE*	Infit	Estimate	*SE*	Infit
Task fulfillment	-.39	.07	.88	-.13	.07	1.10
Organization	-.07	.07	.91	-.08	.07	1.00
Grammar	.62	.07	.92	.41	.07	.95
Vocabulary	.19	.07	1.01	.22	.07	.83
Style	.14	.07	.96	.51	.07	.93
Mechanics	-.49	.07	1.29	-.83	.07	1.19
M	.00	.07	.99	.00	.07	1.00
SD	.41	.07	.15	.47	.07	.13

the different criteria although the infit values showed that they were more consistent on grammar and style.

CONCLUSION

From the use of the three measurement approaches, one can ascertain the different kinds of information that would provide test developers and administrators in ensuring that the reporting of test scores is valid for the intended purpose. Using CTT, two types of information can be ascertained from the analysis. The first are the estimates of inter-rater reliability as suggested in the findings. The second is the percentage of the agreement between the raters. For the subsequent approaches using both GENOVA and FACETS, the analyses revealed different kinds of information. The difference in GENOVA and FACETS appears to be akin to using a microscope - as the level of magnification increases, the details can be seen much clearer (McNamara, 1996). The GENOVA output provided information on the source and magnitude of variance, as well as an estimate of score dependability when changes were made to the number of raters and test tasks. The FACETS analysis, on the other hand, revealed information that could be useful for the test revision process and in the training and certification of raters. From the FACETS analysis, raters' severity and item difficulty may affect the rating of test tasks which are subjectively rated using criterion-referenced marking and such information can be used to prompt test administrators to provide further training. As such, the evidence suggested that on the whole, the ELPA raters were experienced and therefore resulted in fewer occurrences of inconsistency and severity. On the other hand, evidence from using FACETS points to the need for the test administrators to relook at how the ELPA raters use and interpret the rating scale, especially for the analytical scale. Such counter-evidence can affect the validity argument for test score dependability and generalizability. Finally when used appropriately, these three approaches can strengthen the validity argument for score dependability and generalizability.

REFERENCES

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and raterjudgments in a performance test of foreign language speaking. *Language Testing, 12*, 238–58.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer.

FACETS Rasch Measurement [Computer software]. (2007). Chicago: Winsteps.com.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Professional Education.

Linacre, J. M. (2004). *A user's guide to FACETS: Rasch-model computer programs* [Software manual]. Chicago: Winsteps.com.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Boston: Addison-Wesley Publications.

Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh, Scotland: Edinburgh University Press.

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*, 158–80.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park, CA: Sage.

SPSS (Version 16.0 for Windows) [Computer Software]. (2007). Chicago: SPSS Inc.

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