An Investigation of an ESL Placement Test of Writing Using Many-facet Rasch Measurement

Taejoon Park[1]
Teachers College, Columbia University

ABSTRACT

Because performance assessment, such as a composition test, introduces a range of factors that may influence the chances of success for a candidate on the test, those in charge of monitoring quality control for performance assessment programs need to gather information that will help them determine whether all aspects of the programs are working as intended. In the present study, Many-facet Rasch measurement (Linacre, 1989) was employed to examine the effects of various sources of variability on students' performance on an ESL placement test of writing and also to investigate the validity of the assigned scores for students' essays.

INTRODUCTION

For the past two decades, most research on the evaluation of second language writing has focused on the issue of establishing the reliability of scoring among pools of raters (e.g., Shohamy, Gordon, & Kraemer, 1992; Stansfield & Ross, 1988; Weigle, 1998). In a test of writing, this has been of greatest concern because of the reliance on human interpretation in rating students' compositions. Writing assessment programs have tended to address this matter by carefully refining their scoring guides and their procedures for scoring, by training and maintaining pools of raters, and by establishing consistent agreement among these raters (i.e., inter-rater reliability). As Hamp-Lyons (1990) has pointed out, however, establishing and maintaining inter-rater agreement is only a minimum step toward a reliable and valid assessment of writing quality. Inter-rater reliability therefore needs to be complemented in testing practice by additional analyses, because performance assessment, such as a composition test, inevitably introduces a range of factors that may influence the chances of success for a candidate on the test. That is, a candidate's performance on a writing test can be affected by several factors, including variables related to the writing task itself (e.g., the topic, the expected discourse mode of the response, and the number of discrete writing samples a candidate is asked to provide) and by variables related to the scoring process (e.g., the background and experience of the raters, the nature of the rating scale, and the training given to raters).

[1] Taejoon Park is a doctoral student in Applied Linguistics at Teachers College, Columbia University. His current research interests are writing assessment, focusing in particular on rating scale development and validation, and the effects of method factors on observed ratings. Correspondence should be sent to Taejoon Park, 1230 Amsterdam Ave. #613, New York, NY 10027. E-mail: tp125@columbia.edu.

With such complex assessment challenges, Many-facet Rasch measurement (Linacre, 1989) has proven extremely useful in investigating the effects of sources of variability within the context of performance assessments. Many-facet Rasch measurement (MFRM), which represents an extension of the one parameter Rasch model, provides a framework for obtaining fair measurements of examinee ability that are statistically invariant over raters, tasks, and other aspects of performance assessment procedures.

Over the last several years, a number of researchers have used MFRM to examine and understand sources of variability in scores from second language performance assessments. Tyndall and Kenyon (1996) attempted to validate a newly developed holistic rating scale to be used in the placement test for Georgetown University's ESL program using a Rasch many-faceted approach. The results of their study indicated that there is a single construct of writing ability that is being measured with the scale in the operational procedure used. Milanovic, Saville, Pollitt, and Cook (1996) reported on the development of the Cambridge Assessment of Spoken English (CASE), with particular reference to the trialing and validation of the rating scales. In this study, the degree to which raters were able to differentiate between the points on the scale was investigated through Partial Credit analysis (Wright & Masters, 1982), which provides a means for the empirical validation of rating scales. This study provided evidence on the overall workability of their scales in terms of model-data fit, the quality of measurement as expressed in examinee misfit, and the sensitivity of the raters to particular sub-scales. Weigle (1998) investigated differences in rater severity and consistency among inexperienced and experienced raters both before and after training. The results provided support for the notion that rater training is more successful in helping raters give more predictable scores (i.e., intra-rater reliability) than in getting them to identical scores (i.e., inter-rater reliability). Myford and Wolfe (2000) examined four sources of variability in scores from the Test of Spoken English (TSE) assessment system to gain a better understanding of how the complex system operates. More recently, Kondo-Brown (2002) investigated how judgments of trained teacher raters were biased towards certain types of candidates and certain criteria in assessing Japanese second language (L2) writing. The results of the study showed that the raters scored certain candidates and criteria more leniently or harshly, and every rater's bias pattern was different. This study also showed that the modified version of the "ESL Composition Profile" (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981), a scoring procedure containing several clearly articulated scales for the scoring of different facets of writing, can be a reliable tool in assessing Japanese L2 writing in norm-referenced settings (for the major and minor changes made to the original version, see Kondo-Brown, 2002).

In the present study, building on the pioneering efforts of researchers who have employed Many-facet Rasch measurement within the context of second language performance assessments, I attempted to examine the validity of the composition component of the Community English Program (CEP) placement test battery developed at Teachers College, Columbia University. While most of the studies mentioned above have focused on only one or two aspects of complex assessment systems, the present study investigated all of the four sources of variability (i.e., examinees, raters, domains or performance criteria, and rating scales) within the CEP writing assessment system because the purpose of the study was to collect necessary information that will help determine whether all aspects of the CEP writing test are working as intended.

The study was designed to answer the following research questions about the sources of variability:

1. To what extent has the CEP writing test succeeded in separating examinees into distinct levels of proficiency?
2. Are there examinees that exhibit unusual profiles of ratings across the four domains of the CEP scoring rubric?
3. Do CEP raters differ in the severity with which they rate examinees?
4. Are there raters who rate examinee performance inconsistently?
5. Can a single summary measure capture the essence of examinee performance across the different domains of the CEP scoring rubric?
6. Are the CEP rating scales functioning appropriately? In other words, are the four 4-point rating scales appropriately ordered and clearly distinguishable?

It should be noted that a restricted definition of validity was used in this study, one that is common in Rasch analysis: if Rasch analysis shows little misfit, there is evidence for the construct validity of this measurement procedure (Wright & Masters, 1982; Wright & Stone, 1979).

METHOD

Participants

The participants in the present study consist of 99 ESL students with a wide range of English language proficiency. All of them were enrolled in the CEP at Teachers College, Columbia University at the time of the test administration. The CEP is an integral part of the TESOL program at Teachers College. It provides English instruction to adult learners of diverse nationalities and serves as a pedagogical laboratory for teacher preparation and materials development. Of the 99 participants, 49% were male and 51% were female.

Instrument

The test used for this study was the writing subtest of the CEP placement test battery that was designed for placing adult ESL learners enrolled in the CEP into a class that is appropriate for their level of language ability. The CEP placement test battery includes five sections: grammar, reading, listening, speaking, and writing. Of these five sections, the first three sections (i.e., grammar, reading, and listening) are scored dichotomously and the performance assessment sections (i.e., speaking and writing) are scored by trained raters using scoring rubrics (see Appendix A for the CEP writing scoring rubric). The CEP writing test consists of directions for test-takers and one prompt that is descriptive in nature (see Appendix B). The writing test can be characterized as a *timed impromptu essay test* because test-takers are required to write an essay on the spot in response to a given prompt within a relatively short period of time.

Procedures

The data for the present study were 99 essay samples collected from 99 students who took the CEP writing test in February 2003. The students were given 30 minutes to write an essay on the given topic.

Seventeen raters (who were all CEP teachers and graduate students in TESOL or Applied Linguistics) scored the students' essays. Most of them had not had experience with composition rating. Eleven of these raters were native speakers of English and six were non-native speakers of English. Immediately before the scoring session, the raters were all given a program of training (i.e., a norming session), consisting of an orientation to the test, a discussion of the scoring rubric, rating practice, and a discussion of several writing samples that represent the whole range of the CEP scoring rubric.

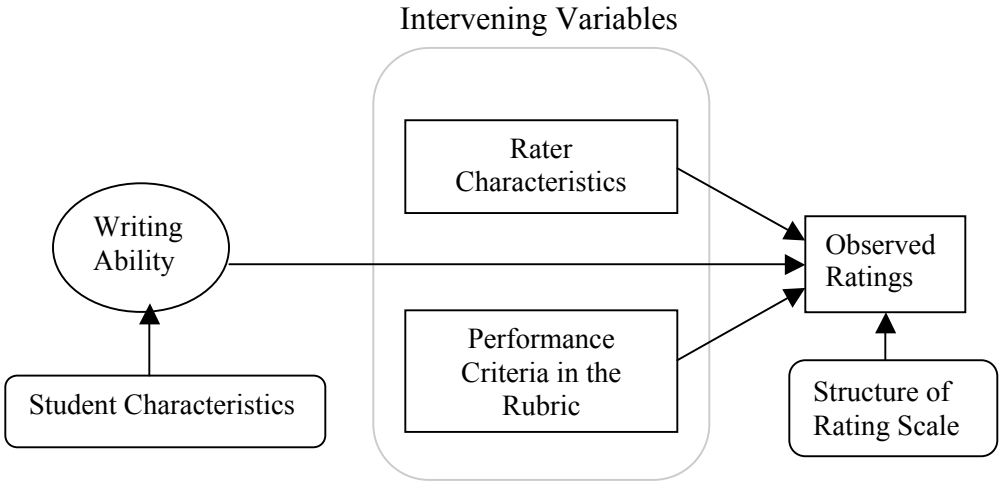
Each essay was rated by two independent raters using the CEP scoring rubric, which consists of the following four domains: overall task fulfillment, topic/content control, organization, and language control. A four-category rating scale (ranging from 1 to 4) was used for each domain. No adjudication was made during the scoring session to resolve discrepancies in ratings, and the final rating pattern used to estimate student writing ability consisted of eight ratings (2 raters x 4 domains).

Analyses

Measurement model for the assessment of writing ability

The measurement model underlying the CEP writing test is presented graphically in Figure 1.

FIGURE 1
Measurement Model Underlying the CEP Writing Test
(Adapted from Engelhard, Jr., 1992)



This model was originally put forth by Engelhard (1992) in an attempt to specify factors that influence observed ratings in the assessment of writing ability using an analytic scoring rubric. In Figure 1, the dependent variable is the observed rating. The three major factors that influence the rating are writing ability, rater characteristics, and performance criteria (i.e., domains) in the scoring rubric. Raters and domains can be viewed as intervening variables that are used to make the latent variable (writing ability) observable. In this model, the structure of the rating scale also affects the value of the rating obtained.

Computer equipment

SPSS version 10.0 was used for computing descriptive statistics, inter-rater reliability, and internal consistency reliability. The Many-facet Rasch measurement (Linacre, 1989) analysis was conducted using the computer program FACETS, version 2.62 for PC (Linacre, 1999a).

Statistical procedures

First, descriptive statistics were computed to check if the scores in each of the four domains (i.e., the performance criteria in the rubric) are normally distributed. Then inter-rater reliability was computed to estimate the degree of agreement between the two independent raters used to score each student's writing sample. Internal consistency reliability (alpha) was also computed to examine how the four domains of the scoring rubric performed as a group.

In addition, FACETS analysis was conducted to examine the overall rating patterns in terms of main effects for the examinee, rater, and domain facets. In FACETS analysis, individual rating patterns that were unusual in light of expected patterns were identified by examining fit statistics. Rating scale functionality was also investigated by examining the average examinee proficiency measure and the outfit mean-square index provided by FACETS.

In the many-facet Rasch model (Linacre, 1989), each element of each facet of the testing situation (e.g., rater, item, rating scale category) is represented by one parameter that represents the ability of examinees, the severity of raters, the difficulty of items, or the challenge of rating scale categories. The Partial Credit form of the many-facet Rasch model used for this study was:

$$\log (P_{nijk}/P_{nijk-1}) = B_n - C_j - D_i - F_{ik}$$

P_{nijk} = the probability of examinee *n* being awarded a rating of *k* when rated by rater *j* on item *i*
P_{nijk-1} = the probability of examinee *n* being awarded a rating of *k-1* when rated by rater *j* on item *i*
B_n = the ability of examinee *n*
C_j = the severity of rater *j*
D_i = the difficulty of item *i*
F_{ik} = the difficulty of achieving a score within a particular score category (*k*) on a particular item (*i*).

In the above model, the four domains of the CEP scoring rubric were treated as items and the step difficulty of the available scoring categories in each domain was calculated independently of the step difficulty of the other domains. This particular model was employed

for this study because the scoring criteria for the four domains were presumed to be qualitatively different, and thus it was assumed that each domain, or item, has its own step structure.

RESULTS

Descriptive Statistics

First, the descriptive statistics for the scores in each of the four domains are presented in Table 1. The means ranged from 2.47 to 2.62 and the standard deviations from 1.01 to 1.04. All values for skewness and kurtosis were within the accepted limits (i.e., +/- 2), indicating that the four domains appeared to be normally distributed.

TABLE 1
Descriptive Statistics of Scoring in Each Domain

	Overall Task Fulfillment	Content Control	Organization	Language Control
Mean	2.62	2.58	2.47	2.52
SD	1.01	1.04	1.03	1.03
Skewness	-0.19	-0.11	0.04	-0.06
Kurtosis	-1.04	-1.16	-1.15	-0.95

Table 2 presents the inter-rater reliability coefficients between two independent raters for each of the four domains. Because the two observed ratings were considered as ordinal data, inter-rater reliability was obtained by computing the Spearman rank-order correlation coefficients. These values were adjusted by using the Spearman-Brown Prophesy Formula, as suggested by Henning (1987). The values of inter-rater reliability ranged from 0.77 to 0.81 and they suggest that there existed some disagreement between the two independent raters, although there was a fair amount of consistency in assigning scores to the examinees' essays.

TABLE 2
Inter-rater Reliability

	Inter-rater reliability
Overall Task Fulfillment	0.81
Content Control	0.77
Organization	0.77
Language Control	0.78

The reliability estimate for internal consistency for the four variables (i.e., overall task fulfillment, topic/content control, organization, and language control) was relatively high (0.93), suggesting that the same abilities are being measured on each domains of the CEP scoring rubric.

The FACETS Analysis

As mentioned above, FACETS analysis was conducted to examine the overall rating patterns in terms of main effects for the examinee, rater, and domain facets.

Figure 2 shows graphically the measures for examinee ability, rater severity, and domain difficulty. The first column in the map displays the logit scale. The logit scale is a true interval scale, unlike raw test scores in which the distances between intervals may not be equal. The FACETS program calibrates the examinees, raters, domains, and rating scales so that all facets are positioned on the same equal interval scale, creating a single frame of reference for interpreting the results from the analysis.

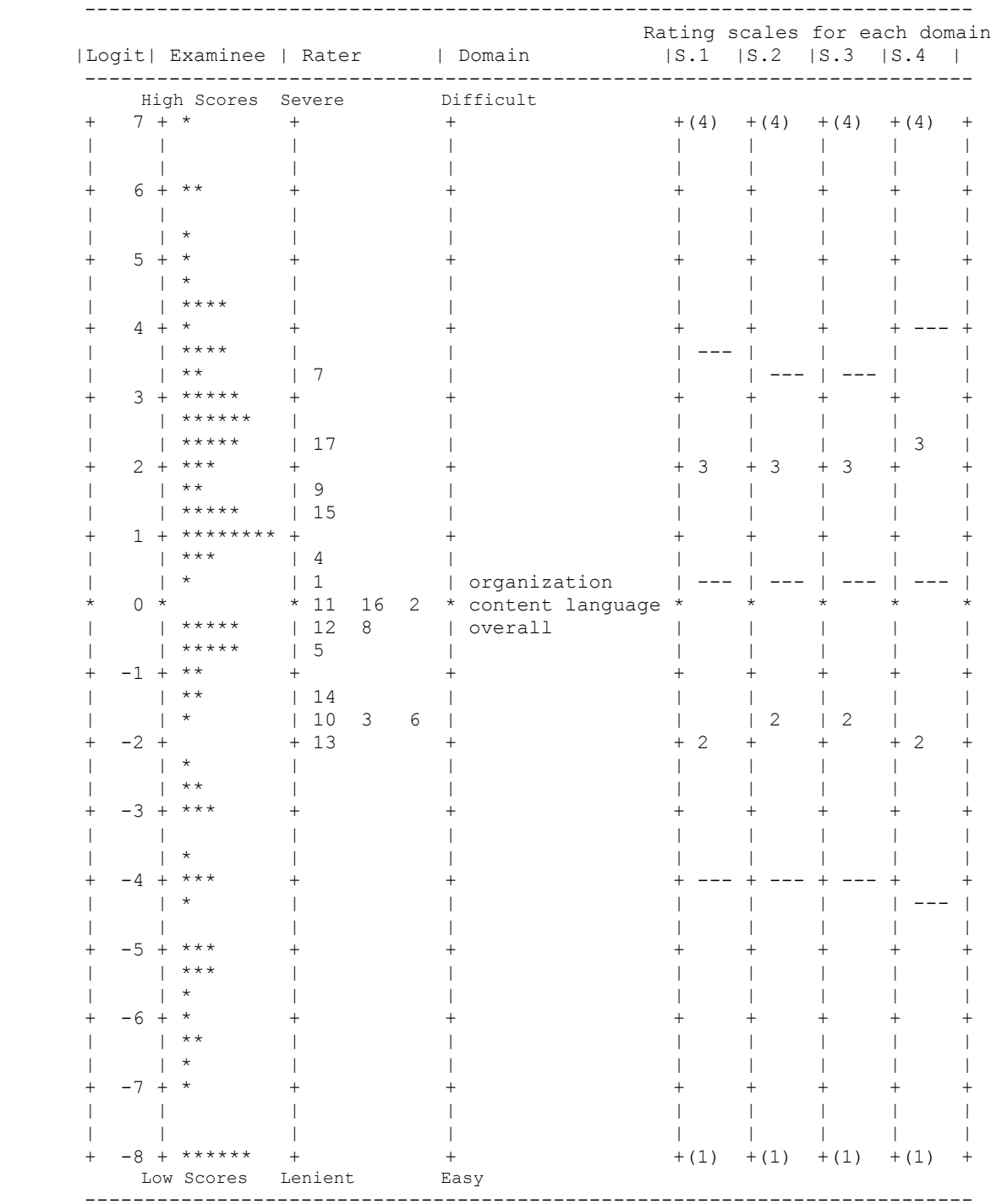
The second column displays estimates of examinee ability—single number summaries on the logit scale of each examinee's tendency to receive low or high ratings across raters and domains, given the scales. Higher scoring examinees appear at the top of the column, while lower scoring examinees appear at the bottom. The column for examinees shows that there is a wide range of variation in terms of examinee ability, with estimates ranging from a high of about 7 logits to a low of about -8 logits. This column shows that there are a much larger number of higher scoring examinees than lower scoring ones. In other words, the examinee ability measures appear as a negatively skewed distribution.

The third column shows the severity variations among raters. The most severe rater is at the top and the least severe at the bottom. Figure 2 shows that the harshest rater has a severity measure of about 3.2 logits and the most lenient rater has a severity measure of about –2.0 logits, indicating that the raters are not at the same level of severity.

The fourth column compares the four domains of the CEP scoring rubric in terms of their relative difficulties. Domains appearing higher in the column were more difficult for examinees to receive high ratings than on domains appearing lower in the column. Figure 2 shows that all of the four domains centered around zero. Zero is, by definition, set as the average domain difficulty on the logit scale. That the four domains centered around zero indicates that although the four domains cannot be considered equivalent, the difficulty span was relatively small.

Columns five through eight graphically describe the four-point rating scales used to score examinee responses. Each domain has its own scale. The horizontal lines across each column indicate the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating for a given domain (Myford, Marr, & Linacre, 1996, p. 21). For example, when we examine Figure 2, we see that examinees with ability measures from about –4.0 logits up through about 0.25 logits are more likely to receive a rating of 2 than any other rating on scale 1 (i.e., the overall task fulfillment scale); examinees with ability measures between about 0.25 logits and about 3.7 logits are more likely to receive a rating of 3 than any other rating on the overall task fulfillment scale. The issue of rating scale functionality will be discussed later in detail.

FIGURE 2
FACETS Summary (Examinee Ability, Rater Severity, Domain Difficulty)



Note. S.1 = Scale for overall task fulfillment
S.2 = Scale for topic/content control
S.3 = Scale for organization
S.4 = Scale for language control

Examinees

Table 3 provides a summary of selected statistics on the ability scale constructed by the analysis for 99 examinees. The mean ability of examinees was 0.33 logits, with a standard deviation of 3.64. The examinee ability measures ranged from –8.06 to 7.03 logits. The separation index and test reliability of examinee separation (the proportion of the observed variance in measurements of ability which is not due to measurement error) were 4.50 and 0.95 respectively. This reliability statistic indicates the degree to which the analysis reliably distinguishes between different levels of ability among examinees. This measure is termed the "Rasch analogue of the familiar KR20 index" by Pollitt and Hutchinson (1987). For examinees, the reliability coefficient of 0.95 indicates that the analysis is fairly reliably separating examinees into different levels of ability. The chi-square of 1836.90 ($df = 91$) was significant at $p = .00$ and, therefore, the null hypothesis that all examinees were equally able must be rejected.

TABLE 3
Summary of Statistics on Examinees (N=99)

Mean ability	0.33
Standard deviation	3.64
Mean Square measurement error	0.79
Separation index	4.50
Test reliability of examinee separation	0.95
Fixed (all same) chi-square	1836.90 (df=91, P = .00)

In order to identify examinees who exhibit unusual profiles of ratings across the four domains of the scoring rubric, fit statistics were examined. The FACETS analysis provides two measures of fit, or consistency: the infit and the outfit. The infit is the weighted mean-square residual that is sensitive to unexpected responses near the point where decisions are being made, whereas the outfit is the unweighted mean-square residuals and is sensitive to extreme scores. For the purposes of this study, only the infit statistics were examined because they are the ones usually considered the most informative, as they focus on the degree of fit in the most typical observations in the matrix (McNamara, 1996, p. 172). There are no hard-and-fast rules for setting upper- and lower-control limits for the infit statistics (i.e., infit mean-square index). In general, as Pollitt and Hutchinson (1987) suggest, any individual infit mean-square value needs to be interpreted against the mean and standard deviation of the set of infit-mean square values for the facet concerned. Using these criteria, a value lower than the mean minus twice the standard deviation would indicate too little variation, lack of independence, or *overfit*. A value greater than the mean plus twice the standard deviation would indicate too much unpredictability, or *misfit*.

For the examinee facet in this study, the infit mean-square mean was 1.0, with a standard deviation of 0.6, so a value greater than 2.2 (1.0 + [.6 x 2]) would be misfitting. There were four misfitting examinees, representing 4% of the examinees (N=99). The number of misfitting examinees (although small) is a problem, given that Pollitt and Hutchinson (1987) point out we

would normally expect around 2% of misfitting examinees. Table 4 presents the rating patterns and fit statistics for each of the four misfitting examinees.

TABLE 4
Rating Patterns and Fit Indices for "Misfitting" Examinees (N=4)

Ratings received by examinee #11
(Infit Mean-Square Index = 2.4, Ability Measure = 2.12, Standard Error = 0.64)

	Overall Task Fulfillment	Topic/Content Control	Organization	Language Control
Rater #5 (Severity = -0.54)	4	4	4	4
Rater #16 (Severity = -0.05)	2	3	2	2

Ratings received by examinee #42
(Infit Mean-Square Index = 2.8, Ability Measure = 4.34, Standard Error = 0.64)

	Overall Task Fulfillment	Topic/Content Control	Organization	Language Control
Rater #7 (Severity = 3.21)	3	4	4	3
Rater #9 (Severity = 1.51)	3	2	2	3

Ratings received by examinee #65
(Infit Mean-Square Index = 2.9, Ability Measure = 2.71, Standard Error = 0.80)

	Overall Task Fulfillment	Topic/Content Control	Organization	Language Control
Rater #3 (Severity = -1.51)	1	1	3	3
Rater #12 (Severity = -0.43)	2	2	2	2

Ratings received by examinee #93
(Infit Mean-Square index = 2.3, Ability Measure = -3.93, Standard Error = 0.75)

	Overall Task Fulfillment	Topic/Content Control	Organization	Language Control
Rater #3 (Severity = -1.51)	3	3	2	2
Rater #12 (Severity = -0.43)	1	1	1	1

Table 4 shows that examinee 11 received unexpectedly high ratings of 4 by rater 5 in all of the four domains. Examinee 93 received unexpectedly low ratings of 1 by rater 12 in all of the four domains. It should be noted that rater 3 and 12 were responsible for two out of four cases of misfitting examinees. This would suggest that adjudication of the scores of these misfitting examinees and retraining of these raters are called for.

Raters

Rater behavior can be analyzed in terms of relative severity, and also in terms of consistency within individual raters (i.e., intra-rater reliability). Table 5 provides a summary of selected statistics on the rater facet.

TABLE 5
Calibration of Rater Facet

Rater ID	Rater Severity Measure (in logits)	Standard Error	Infit Mean-Square Index
7	3.21	0.24	1.0
17	2.43	0.24	0.8
9	1.51	0.23	1.0
15	1.41	0.29	1.0
4	0.76	0.23	0.9
1	0.31	0.42	0.7
2	0.17	0.24	0.8
11	-0.02	0.26	1.0
16	-0.05	0.25	1.0
8	-0.36	0.27	1.0
12	-0.43	0.23	0.9
5	-0.54	0.25	1.1
3	-1.51	0.23	1.3
10	-1.83	0.25	1.0
6	-1.83	0.25	0.9
14	-1.20	0.29	0.8
13	-2.03	0.46	0.6
Mean	0.00	0.35	0.9
SD	0.08	0.01	0.2

Reliability of separation index = 0.97; fixed (all same) chi-square: 627.5, *df*: 16, significance: $p = .00$

Table 5 shows rater IDs, rater severity, error, and infit mean-square values. The second column shows that the severity span between the most lenient rater (Rater 13) and the most severe rater (Rater 7) was 5.24 logits. The reliability of separation index (which indicates the likelihood to which raters consistently differ from one another in overall severity) was high (0.97). For raters, a low reliability is desirable, since ideally the different raters would be equally

severe. In this case, however, the reliability is 0.97 for all raters, indicating that the analysis is reliably separating raters into different levels of severity. Also, the chi-square of 627.6 (*df* = 16) was significant at $p = .00$ and, therefore, the null hypothesis that all raters were equally severe must be rejected. These indicators of the magnitude of severity differences among raters indicate that significant variation in harshness did exist among the raters: Rater 7 was consistently harsher than other raters; conversely, Rater 13 was consistently more lenient than other raters. The third column shows that the level of error was small. The last column indicates that no raters were identified as misfitting: fit values for all raters were within the range of two standard deviations around the mean (0.9 +/- [0.2 x 2]). In other words, all raters were self-consistent in scoring.

Domains

Table 6 presents the results of the FACETS analysis for domains.

TABLE 6
Calibration of Domain Facet

Domain	Difficulty Measure (in logits)	Standard Error	Infit Mean-Square Index
Organization	0.36	0.12	1.1
Language control	0.04	0.13	1.0
Content	-0.09	0.12	1.1
Overall task fulfillment	-0.30	0.13	0.8
Mean	0.00	0.16	1.0
SD	0.26	0.00	0.1

Reliability of separation index = 0.64; fixed (all same) chi-square: 11.1, *df*: 3, significance: $p = .01$

Table 6 shows the domains, domain difficulty measures, error, and infit mean-square values. The most leniently scored domain was overall task fulfillment, the most harshly scored domain was organization, and the difficulty span between these two domains was relatively small (0.66), as were the separation index (1.34) and the reliability of domain separation (0.64), suggesting that the domains were relatively similar in difficulty. To further investigate the relationship among the four domains, the infit mean-square indices were examined. They are all within the acceptable limits of 0.80 to 1.2 (i.e., the range of two standard deviations around the mean: 1.0 +/- [2 x 0.1]). The fact that there is no overfitting domain (the infit mean-square index lower than 0.80) suggests that none of the domains function in a redundant fashion. That is, the four domains being scored in the rubric are not too similar. The fact that there is no misfitting item (the infit mean-square index greater than 1.2) indicates that there is little evidence of psychometric multidimensionality. The four domains on the rubric appear to work together; ratings on one domain correspond well to ratings on other domains. That is, a single pattern of proficiency emerges for these examinees across all domains. Therefore, ratings on the individual domains can be meaningfully combined; a single summary measure can appropriately capture the essence of examinee performance across the four domains of the scoring rubric.

Rating scale

To see if the four 4-point rating scales are appropriately ordered and clearly distinguishable (i.e., rating scale functionality), the *average examinee ability measure* and *outfit mean-square index* provided by FACETS for each rating category for each of the four domains were examined.

To compute the average examinee ability measure for a rating category, the examinee ability measures (in logits) for all examinees receiving a rating in that category on that domain are averaged. If the rating scale for the domain is functioning as intended, then average examinee ability measures will increase in magnitude as the rating scale categories increase. When this pattern is borne out in the data, the results suggest that examinees with higher ratings on the domain are indeed exhibiting more of the variable being measured than examinees with lower ratings on that domain, and therefore the intentions of those who designed the rating scale are being fulfilled (Linacre, 1999b).

Table 7 shows the average examinee ability measures along with outfit mean-square indices by rating scale category for each of the four domains.

TABLE 7
Average Examinee Ability Measures and Outfit Mean-square Indices from the FACETS Output

	Domain							
	Overall Task Fulfillment		Content Control		Organization		Language Control	
Category Label	Average Measures	Outfit MnSq	Average Measures	Outfit MnSq	Average Measures	Outfit MnSq	Average Measures	Outfit MnSq
1	-5.30	1.0	-5.35	0.9	-5.61	0.9	-5.67	1.1
2	-1.59	0.7	-1.23	0.9	-0.95	0.9	-1.79	1.1
3	2.16	0.8	1.94	1.0	1.64	1.4	2.15	0.9
4	5.29	0.8	4.93	1.0	4.65	0.9	5.38	0.9

In Table 7, the average examinee ability measures for all of the domains increase as the rating scale categories increase. For the overall task fulfillment domain, for example, the average examinee proficiency measures increase from –5.30 to 5.29 as the rating scale categories increase.

The *outfit mean-square index* is also a useful indicator of rating scale functionality. For each rating scale category for each domain, FACETS computes the average examinee ability measure (i.e., the observed measure) and an expected examinee ability measure (i.e., the examinee ability measure the model would predict for that rating category if the data were to fit the model). When the observed and expected examinee ability measures are close, then the outfit mean-square index for the rating category will be near the expected value of 1.0. The greater the discrepancy between the observed and expected measures, the larger the mean-square index will be. For a given rating category, an *outfit mean-square index* greater than 2.0 suggests that a rating in that category for one or more examinees may not be contributing to meaningful

measurement of the variable (Linacre, 1999b). As shown in Table 7, there was not a single domain that has outfit mean-square indices greater than 2.0 for any rating category, suggesting that the rating scales for the four domains seem to be functioning as intended.

Figures 3, 4, 5, and 6 present the Scale Category Probability Curves that enables one to see at a glance the structure of the scoring rubric and, particularly, whether raters are using all the categories on the rubric.

FIGURE 3
Scale Category Probability Curves for Overall Task Fulfillment

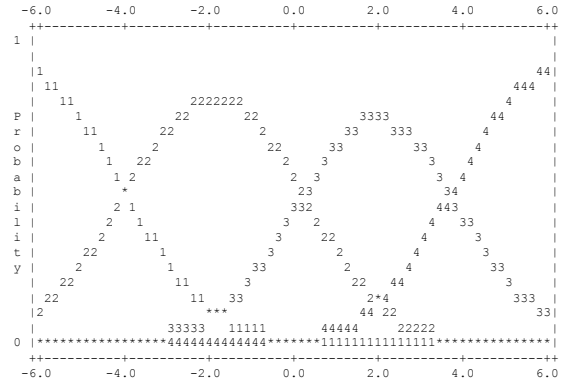


FIGURE 4
Scale Category Probability Curves for Content Control

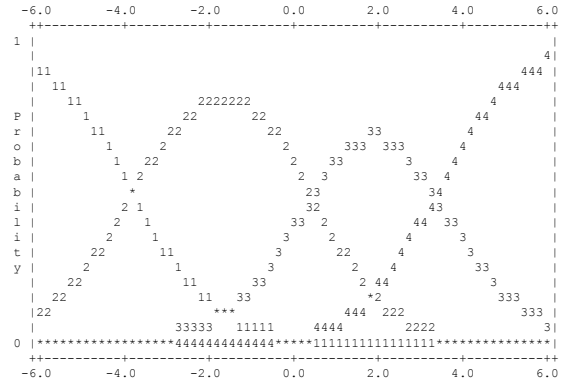


FIGURE 5
Scale Category Probability Curves for Organization

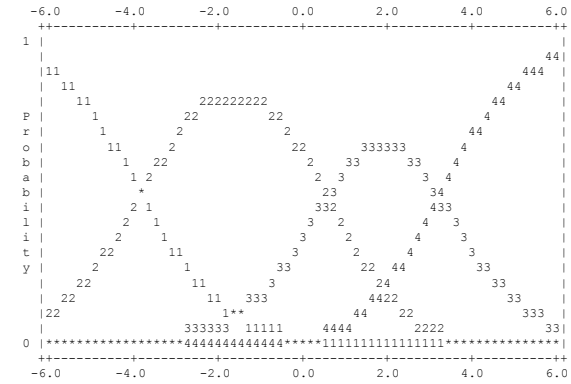
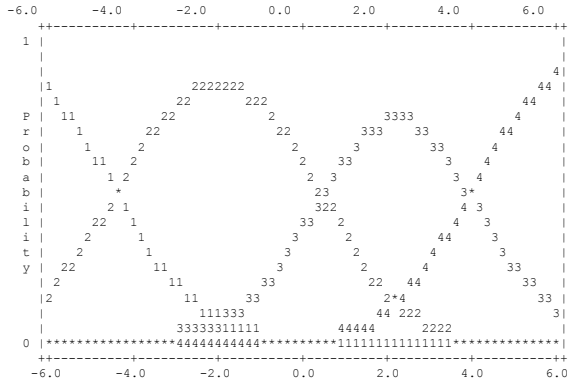


FIGURE 6
Scale Category Probability Curves for Language Control



The horizontal axis represents the examinee proficiency scale (in logits) and the vertical axis represents probability (from 0 to 1). There is a probability curve printed for each of the scale categories. Since the CEP scoring rubric uses 4-point scales, the scale category probability curves are labeled as 1, 2, 3, and 4.

When examining these graphs, the chief concern is whether there is a separate peak for each scale category probability curve, and whether the curves appear as an evenly spaced series of hills. If there is a separate peak for a scale category curve, then it denotes that for examinees in a specific portion of the examinee proficiency distribution, that category is the most probable rating their essays would receive. If there is not a separate peak for a scale category curve that rises above the peaks for adjacent category curves, then that would indicate that the category is never the most probable rating for any clearly designated portion of the examinee proficiency

distribution. As shown in Figures 3, 4, 5, and 6, the probability curves for the four scales appear as a fairly evenly spaced series of hills. For each scale category there is a clearly designated portion of the examinee proficiency distribution for which that category is the most probable rating given.

Bias analysis

In the context of writing performance assessments, there may be an interaction involving a rater and some other aspect of the rating situation. The identification of these systematic sub-patterns of behavior is achieved in MFRM in so-called *bias analysis*. In this study, a bias analysis was carried out on the interaction of raters with domains. This identifies raters who are responding consistently to a domain in a way that is both different from other raters, and different from their own behavior in relation to other domains.

There were nine instances (13.2% of the total interactions) of significant bias out of 68 possible interactions (17 raters x 4 domains). Table 8 presents all of the instances of significantly biased interactions.

TABLE 8
Significantly Biased Rater-Domain Interactions (N=9)

Rater #	Domain	Z-score
2	Organization	2.74
2	Language control	-2.31
4	Organization	-2.06
5	Language control	2.58
6	Organization	2.37
6	Language control	-2.23
7	Organization	-2.12
11	Organization	2.10
11	Language control	-2.62

If a z-score value in Table 8 is greater than +2, the domain is triggering a systematically more severe behavior than is normal for the rater in question. If a z-score value is smaller than -2, the domain is triggering a systematically more lenient behavior than is normal for the rater in question. In Table 8, for example, the interaction between rater #2 and the organization domain produced a statistically significant bias value ($z = 2.74$), suggesting that rater #2 is demonstrating a more severe than normal rating pattern with the organization domain. The same rater is demonstrating a more lenient than normal rating pattern with the language control domain, as indicated by a z-score value that is smaller than -2 ($z = -2.31$).

It should be noted that significant rater-domain interactions were found in the domains of organization and language control, but not in the two other domains (overall task fulfillment and topic/content control). This could mean that the descriptors for organization and language control were somehow more difficult to agree on than those for overall task fulfillment and topic/content

control. This finding suggests that clearer criteria and training for judging the performances of examinees especially on these domains might be required.

DISCUSSION AND CONCLUSION

In this study, four sources of variability (i.e., examinee, rater, domain, and rating scale) in scores from the CEP writing test were examined with the help of Many-facet Rasch measurement. The investigation of the *examinee* facet showed that the CEP writing test usefully separated test-takers into statistically distinct levels of proficiency. A few examinees exhibited unusual profiles of ratings across the four domains of the CEP scoring rubric. Indeed, about 4% of the examinees showed significant misfits. The rating patterns of these misfitting examinees should be reviewed before issuing score reports, particularly if an examinee's measure is near a critical decision-making point in the score distribution. From the decision-maker's viewpoint, the ability measures for individual examinees provided by MFRM are fairer than raw scores because they were corrected for differences in raters, domains, and rating scales. For example, adjustments for rater severity improve the objectivity and fairness of the measurement of writing ability because unadjusted scores can lead to under- or overestimates of writing ability when students are rated by different raters. MFRM thus provides a sound theoretical framework for obtaining objective and fair measurements of writing ability that generalize beyond the specific raters, domains, and rating scales.

The examination of the *rater* facet revealed that while the raters differed in the severity with which they rated examinees, all of them used the CEP scoring rubric in a consistent manner. That is, the raters appeared to be internally consistent but are not interchangeable, confirming the findings of Weigle (1998) that rater training is more successful in helping raters give more predictable scores (i.e., intra-rater reliability) than in getting them to give identical scores (i.e., inter-rater reliability).

The analysis of the *domain* facet showed that the domains work together; ratings on one domain correspond well to ratings on the other domains, indicating that a single pattern of proficiency emerges for these examinees across all domains on the scoring rubric. Therefore, ratings on the individual domains can be meaningfully combined; a single summary measure can appropriately capture the essence of examinee performance across the four domains. With regard to rating scale functionality, the average examinee proficiency measure and the outfit mean-square index indicated that the four 4-point subscales are appropriately ordered and clearly distinguishable.

The bias analysis carried out on the interactions between raters and domains revealed that the descriptors for organization and language control were somehow more difficult to agree upon than those for the other two domains. This may suggest that clearer criteria and rater training for scoring the examinee performance especially on these domains are required.

In the present study, the essay prompt was not considered as a facet in the FACETS analysis because only one prompt was used in the current CEP writing test. Ideally, students should be able to respond equally well to different types of writing tasks. However, several studies (e.g., Engelard, Gordon, & Gabrielson, 1991; Kegley, 1986; Prater, 1985; Quellmalz, Capell, & Chou, 1982) indicate that some topics elicit better writing than others, and that some topics are more difficult than others. Because of possible performance fluctuations from topic to topic and/or from one mode of discourse to another, perhaps more than one type of writing task

should be included in high-stakes assessments as a way of achieving a high level of reliability. If additional prompts are to be used in the CEP writing test, the essay prompt facet should be incorporated into the FACETS analysis. This modification would yield additional information on each student's writing ability, resulting in higher score reliability. As Lee, Kantor, and Mollaun (2002) have suggested, in order to maximize score reliability for writing assessments, it would perhaps be more cost-efficient to increase the number of tasks rather than the number of ratings per task.

To conclude, this study showed that the validity of an essay composition test could be investigated with the help of Many-facet Rasch measurement. As mentioned earlier, a restricted definition of validity was used in the present study: if Rasch analysis shows little misfit, there is evidence for the construct validity of this measurement procedure. Although this definition falls short of Messick's (1989) definition of validity based on the empirical and theoretical rationales that support the adequacy and appropriateness of inferences and actions based on the test scores, the FACETS analysis did provide evidence for the construct validity of the CEP writing test.

REFERENCES

Engelhard, G. (1992). The measurement of writing ability with a many-facet Rasch model. *Applied Measurement in Education, 5*, 171-191.

Engelhard, G., Gordon, B., & Gabrielson, S. (1991). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English, 26*, 315-336.

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

Henning, G. (1987). *A guide to language testing*. Boston, MA: Heinle & Heinle.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V.F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Kegley, P. H. (1986). The effect of mode of discourse on student writing performance: Implications for policy. *Educational Evaluation and Policy Analysis, 8*, 147-154.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, *19*, 1-29.

Lee, Y., Kantor, R., & Mollaun, P. (2002). *Score dependability of the writing and speaking section of New TOEFL.* Paper presented at the annual meeting of National Council on Measurement in Education (NCME), New Orleans, LA.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1999a). *FACETS*, Version 3.17 [Computer program]. Chicago: MESA Press.

Linacre, J. M. (1999b). Investigating rating scale category unity. *Journal of Outcome Measurement, 3*, 103-122.

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

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

Milanovic, M., Saville, N., Pollitt, A., & Cook, A. (1996). Developing rating scales for CASE: Theoretical concerns and analyses. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 15-38). Clevedon: Multilingual Matters.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the TWE* (TOEFL Research Report No. 95-40). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report No. 65). Princeton, NJ: Educational Testing Service.

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing, 4,* 72-92.

Prater, D. L. (1985). The effects of modes of discourse, sex of writer, and attitude toward task on writing performance in grade 10. *Educational and Psychological Research, 5*, 241-259.

Quellmalz, E., Capell, F. J., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19*, 241-258.

Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*, 27-33.

Stansfield, C., & Ross, J. (1988). A long-term research agenda for the Test of Written English. *Language Testing, 5*, 160-186.

Tyndall, B., & Kenyon, D. M. (1996) Validation of a new holistic rating scale using Rasch multi-faceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39-57). Clevedon: Multilingual Matters.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

APPENDIX A
Scoring Rubric for the CEP Writing Test (last updated on Jan. 22, 2003)

Category	Level	Criteria
1. Overall task fulfillment	4 Excellent to very good 3 Good to average 2 Fair to poor 1 Very poor	For this category, the rater reads an essay quickly and then assigns a score to the text based on "**an overall impression**." This category aims to rate "the overall proficiency level" reflected in a given sample of student writing.
2. Topic/Content control	4 Excellent to very good	knowledgeable; substantive; thorough development of argument; relevant to assigned topic
	3 Good to average	some knowledge of subject; adequate range; limited development of argument; mostly relevant to topic, but lacks detail
	2 Fair to poor	limited knowledge of subject; little substance; inadequate development of topic
	1 Very poor	does not show knowledge of subject; non-substantive; not pertinent; *or* not enough to evaluate
3. Organization	4 Excellent to very good	well-organized; logical sequencing; cohesive
	3 Good to average	loosely organized but main ideas stand out; limited support; logical but incomplete sequencing
	2 Fair to poor	ideas confused or disorganized; lacks logical sequencing and development
	1 Very poor	does not communicate; no organization; *or* not enough to evaluate
4. Language control (Grammar/Vocabulary)	4 Excellent to very good	effective complex constructions; few errors in grammar; sophisticated range of vocabulary
	3 Good to average	effective but simple constructions; minor problems in complex constructions; several errors in grammar; adequate range of vocabulary
	2 Fair to poor	Major problems in simple/complex constructions; frequent errors in grammar; meaning confused or obscured; limited range of vocabulary
	1 Very poor	virtually no mastery of sentence construction rules; dominated by errors; does not communicate; *or* not enough to evaluate

APPENDIX B

CEP Writing Test

DIRECTIONS:
You will have 30 minutes to write a well-organized essay on the following topic. Before you begin writing, consider carefully and plan what you will say. Make sure you proofread your essay before handing it in.

TOPIC:
Most people think that American schools encourage both cooperation and competition. What about education in your country? Which is considered more important, cooperation or competition? Use specific reasons and examples to support your answer.