APPLIED MEASUREMENT IN EDUCATION, 5(3), 171-191

The Measurement of Writing Ability With a Many-Faceted Rasch Model

George Engelhard, Jr.
Emory University

The purpose of this study is to describe a Many-Faceted Rasch (FACETS) model for the measurement of writing ability. The FACETS model is a multivariate extension of Rasch measurement models that can be used to provide a framework for calibrating both raters and writing tasks within the context of writing assessment. The use of the FACETS model for solving measurement problems encountered in the large-scale assessment of writing ability is presented here. A random sample of 1,000 students from a statewide assessment of writing ability is used to illustrate the FACETS model. The data suggest that there are significant differences in rater severity, even after extensive training. Small, but statistically significant, differences in writing-task difficulty were also found. The FACETS model offers a promising approach for addressing measurement problems encountered in the large-scale assessment of writing ability through written compositions.

Direct assessments of student writing ability are currently being conducted or planned in almost every state (Afflerbach, 1985). These large-scale writing assessments are high-stakes tests for examinees with direct consequences for instructional placement, grade-to-grade promotion, and high school graduation. National assessments of writing ability (Applebee, Langer, & Mullis, 1985; Applebee, Langer, Jenkins, Mullis, & Foertsch, 1990), as well as international assessments (Gorman, Purves, & Degenhart, 1988), have also been conducted using essays written by students.

In spite of the increase in direct assessments of writing ability, relatively little is known about the validity of current measurement procedures for estimating writing ability. The objective assessment of writing ability based on student essays presents a variety of measurement problems that are difficult to address within the framework of current test theories that are

Requests for reprints should be sent to George Engelhard, Jr., Emory University, Division of Educational Studies, 210 Fishburne Building, Atlanta, GA 30322.

primarily designed to model dichotomous data from multiple-choice items. The first problem is that most of the common scoring procedures for essays are based on nondichotomous ratings, such as the traditional Likert-type scales; this is the case whether holistic (Cooper, 1977) or some form of analytic scoring is used (Lloyd-Jones, 1977). Research on psychometric models for this type of data has contributed to our understanding of rating scales (Wright & Masters, 1982), and some of these models have been used to analyze student essays (Pollitt & Hutchinson, 1987). A second problem is that the ratings of essays are made by raters who introduce a source of variation into the measurement process that is not found in multiple-choice tests. Several studies have suggested that, in spite of thorough training, raters still vary in severity (Lunz, Wright, & Linacre, 1990), and interrater reliability remains a significant problem (Braun, 1988; Cohen, 1960). As pointed out by Coffman (1971) in his review of the literature, one of the major problems with essay examinations is that when different raters are asked to rate the same essay they tend to disagree in their ratings. A third problem encountered within the context of large-scale assessments of writing ability is how to adjust for differences in writing-task difficulty when students respond to different writing tasks. There is substantial evidence that writing tasks do differ in difficulty (Engelhard, Gordon, & Gabrielson, 1992; Ruth & Murphy, 1988).

These measurement problems led earlier psychometricians to a Procrustean approach to writing assessment based on multiple-choice items. These indirect assessments led to reliable estimates of writing ability based on standard criteria used with traditional test theory for multiple-choice items. Although there is some evidence that different traits were being measured as a function of test format (Ackerman & Smith, 1988), indirect assessments of writing ability tend to be highly correlated with ratings based on actual writing samples. Indirect assessments of writing ability seem to work well when the major goal of the assessment is simply to rank order students, but these assessments do not encourage the teaching and learning of writing. This well-known connection between assessment procedures and teaching has provided the motivation for increased use of authentic and performance-based measurement of writing, as well as other competencies.

It is beyond the scope of this article to provide a detailed survey of other psychometric models that have been proposed for direct assessments of writing. Briefly, these models can be grouped into two major approaches, one based on analysis of variance (ANOVA) models and the other on linear structural equation models. Examples of approaches to writing assessment based on ANOVA models are the early work of Stanley (1962) and the research of Braun (1988) on the calibration of essay raters. Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) has also been used

to examine essay data (Bunch & Littlefair, 1988; Lane & Sabers, 1989). Blok (1985) and Ackerman and Smith (1988) present examples of how linear structural equation models using LISREL (Joreskog & Sorbom, 1979) can be used to address measurement problems related to writing assessment.

These two approaches are not adequate for a variety of reasons. First, they are based on raw scores that are nonlinear representations of a writing ability variable and do not directly lead to scales that have equal units. Second, the unit of analysis for these two approaches is the raw score rather than individual rating. Recent advances in item response theory highlight the advantages of using the item response directly rather than a raw score as the unit of analysis for both dichotomous and polytomous response data. Item response models can be developed to model directly the probability of a student obtaining a particular set of ratings based on an actual writing sample. Although an empirical comparison of different measurement models for the direct writing assessment would be interesting, it is difficult to develop fair criteria for comparing these models because these approaches possess many of the characteristics of different paradigms (Kuhn, 1970) or research traditions (Laudan, 1977).

Several Rasch-based approaches for modeling essay ratings have also been proposed. Andrich proposed a Poisson Process model based on the number of flaws observed in an essay (Andrich, 1973; Hake, 1986). The Partial Credit model (Masters, 1982) has been used to examine writing data (Ferrara & Walker-Bartnick, 1989; Harris, Laan, & Mossenson, 1988; Pollitt & Hutchinson, 1987). De Gruijter (1984) proposed two models (one additive and the other nonlinear) for rater effects; the nonlinear model is based on the pairwise Rasch model of Choppin (1982). Although each of these Rasch-based models offers significant advantages over earlier approaches to writing assessment, they all are essentially two-facet models (either writing ability and rater severity or writing ability and writing-task difficulty) and cannot adequately model assessment procedures that are designed to have multiple facets. A recent extension of the Rasch model proposed by Linacre (1989) and presented here provides for multiple facets that can be calibrated simultaneously, but examined separately. For example, the four facets defined in this study are writing ability, rater severity, writing-task difficulty, and domain difficulty.

In summary, an assessment framework based on extensions of item response theory seems to offer a promising approach to the measurement of writing ability. The Many-Faceted Rasch (FACETS) model addresses many of the measurement problems encountered with other approaches to writing assessment. Rasch measurement models can provide a framework for obtaining objective and fair measurements of writing ability that are statistically invariant over raters, writing tasks, and other aspects of the writing assessment process. A FACETS model for the direct assessment of

writing ability is described in the next section based on the current procedures used in Georgia for the Eighth Grade Writing Test. In addition, a random sample of 1,000 students is analyzed in order to illustrate the FACETS model. Finally, the implications of the FACETS model for theory, research, and practice within the context of the large-scale assessment of writing ability are summarized.

MEASUREMENT MODEL FOR THE ASSESSMENT OF WRITING ABILITY

The measurement model underlying the writing assessment program used in Georgia is presented graphically in Figure 1. The dependent variable in the model is the observed rating, which ranges from *inadequate* (1), *minimal* (2), *good* (3) to *very good* (4). The four major facets that influence this rating are writing ability, rater severity, difficulty of the writing task, and domain difficulty. Raters, writing tasks, and domains can be viewed as intervening variables that are used to make the latent variable (writing ability) observable. The structure of the rating scale that defines the categories also affects the value of the rating obtained. Other large-scale assessments of writing would require different forms of the FACETS model; for example, if holistic scoring is used, then the domain facet would not be necessary.

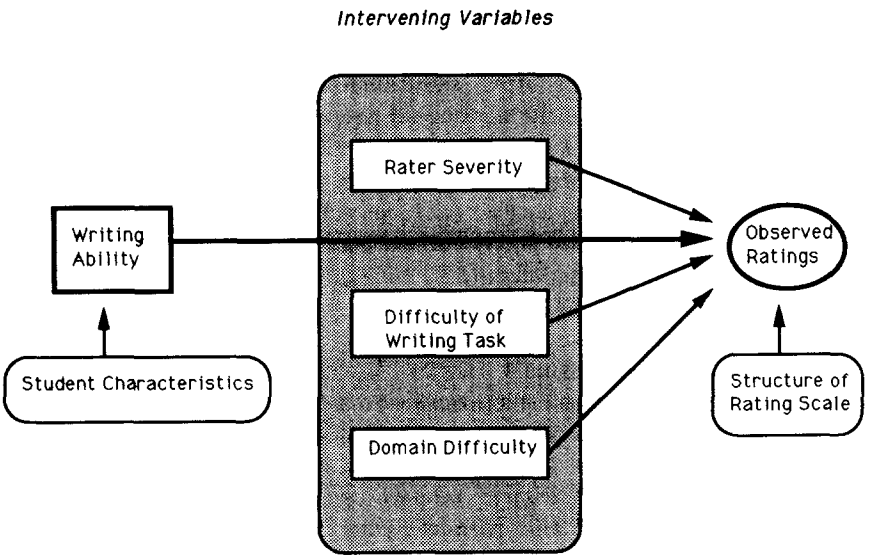

FIGURE 1 Measurement model for the assessment of writing ability.

Although not explicitly included in the measurement model, student characteristics influence writing ability and may reflect potential sources of bias that affect the observed rating of a student. Some examples of these student characteristics are gender, age, ethnicity, race, social class, and opportunity to learn. The biasing effects of these student characteristics can be examined after the facets are calibrated. Studies of differential facet functioning (DFF) can be conducted by a variety of procedures that are conceptually similar to current approaches for studying differential item functioning (Engelhard, Anderson, & Gabrielson, 1990; Holland & Thayer, 1988). For example, the individual facets of the model for the assessment of writing ability could be calibrated separately for females and males, and the correspondence between these estimates examined to detect DFF. Interactions between the facets can also be examined as a potential source of bias in the assessment of writing ability. The measurement model can also be elaborated in order to examine hypotheses about why raters differ in severity and also why writing tasks differ in difficulty.

THE FACETS MODEL

The FACETS model is an extension of Rasch measurement models (Rasch, 1980; Wright & Masters, 1982; Wright & Stone, 1979) that can be used for writing assessments that include multiple facets, such as raters and writing tasks. For the Georgia writing data analyzed here, the FACETS model can be written as follows:

$$\log[P_{nijmk}/P_{nijmk-1}] = B_n - T_i - R_j - D_m - F_k,$$

where

P_{nijmk} = probability of student n being rated k on writing task i by rater j for domain m

$P_{nijmk-1}$ = probability of student n being rated k − 1 on writing task i by rater j for domain m

B_n = Writing ability of student n

T_i = Difficulty of writing task i

R_j = Severity of rater j

D_m = Difficulty of domain m

F_k = Difficulty of rating Step k relative to Step k − 1.

The student facet, B_n, provides a measure of writing ability on a linear logistic scale (logits) that ranges from +/− infinity. If the data fit the FACETS model, then these estimates of writing ability are statistically

invariant over raters and writing tasks. These estimates of writing ability are invariant because adjustments have been made for differences in rater severity and the difficulty of the writing task. The writing-task facet, T_i, calibrates the writing tasks on the same linear logistic scale, and provides an estimate of the relative difficulty of each writing task that is invariant over students and raters. Estimates of rater severity, R_j, are also obtained on the same linear logistic scale that are invariant over students and writing tasks. Finally, invariant calibrations of the domain facet, D_m, and rating scale step difficulties, F_k, are also obtained. In essence, the FACETS model is an additive linear model based on this logistic transformation of the observed ratings to a logit scale.

Several indices that examine the fit of data to the FACETS model are available. These fit statistics provide evidence regarding the validity of the FACETS model. The OUTFIT statistic provides an index of the mean square residual differences between the observed and expected rating patterns (Wright & Masters, 1982; Wright & Stone, 1979). The OUTFIT statistic is particularly sensitive to outlying unexpected ratings. The INFIT statistic is a weighted mean square with weights proportional to the variance of the squared residuals. The INFIT statistic is less sensitive to outlying unexpected ratings. The expected value for both of these mean square statistics is 1.0. When the rating patterns are fairly consistent and there are no unexpectedly large outlying ratings, then the obtained values of the INFIT and OUTFIT will be quite similar. Both the OUTFIT and INFIT statistics can be summarized and reported separately for each facet in the model. Although these mean square fit statistics are continuous indices, Lunz, Wright, and Linacre (1990) have suggested that the region for acceptable fit ranges from 0.6 to 1.5. A reliability of separation index and a chi-square statistic are also available for examining the extent to which objects within a facet are sufficiently separated in order to define the facet reliably. These four fit statistics are similar to those that have been used with two-facet models and are described in detail in Wright and Masters (1982). Additional details regarding the computational and statistical aspects of these fit statistics for the FACETS model are presented in Linacre (1989).

EMPIRICAL EXAMPLE

Description of the Sample

One thousand students were randomly selected from the spring 1990 administration of the Eighth Grade Writing Test that is administered to all eighth-grade students in Georgia. Fifty-four percent of the students are

female and 46% are male. The racial/ethnic distribution is as follows: 60% white, 38% black, and 2% other. The student compositions were rated by 82 raters. The estimated interrater reliability is .82. This value is comparable with those reported by Breland, Camp, Jones, Morris, and Rock (1987).

Instrument

The Eighth Grade Writing Test is a criterion-referenced test designed to provide a direct assessment of student writing ability. Students are asked to write one essay of no more than two pages on an assigned writing task. The writing tasks are randomly assigned to the students.

Although the full rhetorical specification of the writing tasks cannot be revealed because this is a high-stakes test, the theme statements for the tasks examined here are presented in Table 3. Detailed descriptions of the writing tasks and examples of operational prompts that have been used in previous assessments are given in Engelhard et al. (1992). The essays are rated by two raters on each of the following five domains: content/organization, style, sentence formation, usage and mechanics. A four-category rating scale (ranging from *inadequate* [1] to *minimal* [2] to *good* [3] to *very good* [4]) is used for each domain. The final rating pattern used to estimate student writing ability consists of 10 ratings (2 raters $\times$ 5 domains $=$ 10 ratings).

These raters are highly trained, and a variety of procedures are used to maintain the validity and reliability of the ratings. First, the raters must successfully complete an extensive training program; this program typically takes 3 days. Next, the raters go through a qualifying process in order to become an operational rater. During the qualifying process, each rater rates 20 essays, and these ratings are compared with a set of standard ratings assigned by a validity committee of writing experts. Raters with at least 62% exact agreement with the standard on the ratings and 38% adjacent category agreement can become operational raters.

Finally, two ongoing quality control procedures are used to monitor the raters during the actual process of rating student essays. First, validity papers with a set of standard ratings are included in each packet of 24 essays, and rater agreement is examined continuously; the raters are not able to identify the validity paper. Second, each essay is rated by two raters, and if a large discrepancy is found, then the essay is rescored by a third rater.

Procedure

The FACETS computer program (Linacre, 1988) was used to analyze the data. A measurement model with four facets (writing ability, rater severity, writing-task difficulty, and domain difficulty) was estimated for the data.

Since the developers of the writing assessment program intended the rating scale to have a fixed structure, the rating scale model with common step sizes across raters, writing tasks, and domains was used for the structure of the rating scale.

Results

The calibrations for raters, writing tasks, and domains are shown graphically in Figure 2. The severities, standard errors, INFIT, and OUTFIT statistics for the raters are presented in detail in Table 1. The rater severities range from 1.78 logits for Rater 11 who is severe to −1.74 logits for Rater 19 who is lenient. Eighty-five percent of the raters are between −1.00 and +1.00 logits. The overall differences between the raters are significant, χ^2 $(81, N = 82) = 900.80, p < .01$ with a high reliability of separation index $(R = .87)$. The analysis suggests that there are significant differences in rater severity. This significant variation in the raters appears in spite of the extensive training and screening of the raters. Seven of the raters have misfitting rating patterns. Five of the raters (53, 85, 59, 87, and 25) have noisy rating patterns with OUTFIT statistics greater than 1.5, while Raters 20 and 75 have muted rating patterns with OUTFIT statistics less than .6. In order to illustrate the substantive interpretation of these fit statistics for raters, the rating patterns for Rater 59 are reported in Table 2. Rater 59 has a noisy rating pattern (INFIT = 1.7; OUTFIT = 1.6) with 15 unexpected ratings. Raters with muted rating patterns tend to score holistically rather than analytically. For example, Rater 20 has a muted rating pattern (INFIT = 0.6; OUTFIT = 0.5). Fifty percent of the 20 students rated by this rater had uniform rating patterns (5 students with "22222," 4 students with "33333," and 1 student with "44444"); this rater did not use the first rating category of 1 (inadequate). Similar interpretations of the INFIT and OUTFIT statistics apply to the other five raters.

The calibration of the writing tasks is presented in Table 3. Writing Task 77, in which students are asked to write about an "all-expense-paid trip," is the most difficult to write about for these students with a difficulty of .12 logits $(SE = .06)$, while Writing Task 75 ("experience that turned out better") is the easiest to write about with a difficulty of −.16 logits $(SE = .06)$. There is no evidence of misfit based on the INFIT or OUTFIT statistics with all of the fit mean squares between .9 and 1.1. Even though the reliability of separation index is small $(R = .46)$, the analysis suggests that the overall differences between the writing tasks are statistically significant, $\chi^2(7, N = 8) = 15.22, p = .03$. The writing tasks were designed by the developers of the writing assessment program to be equivalent, and the small differences between the difficulties of the writing tasks tend to reflect this intention. Although the differences are small, it appears that the

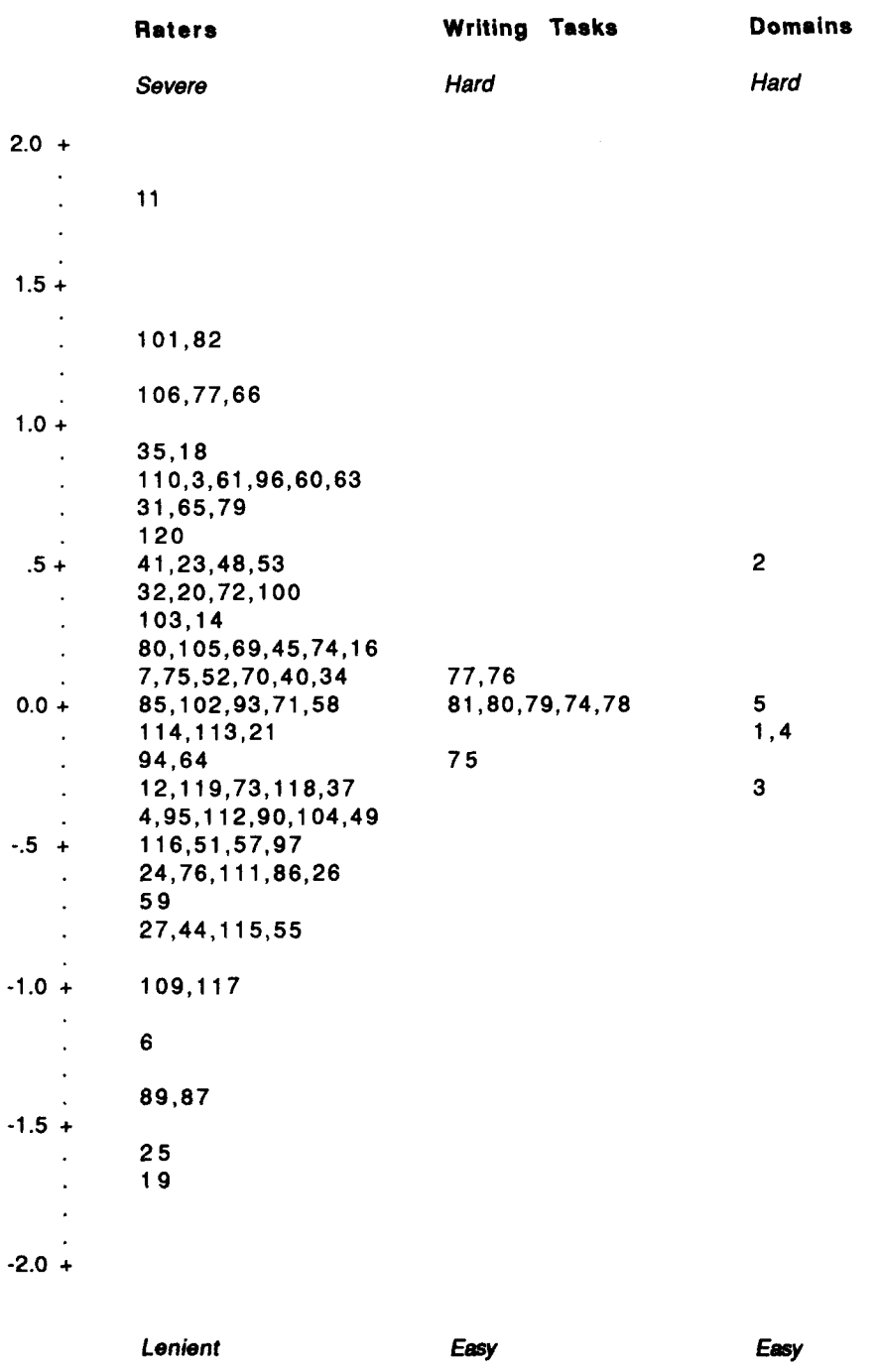

FIGURE 2 Calibrations of rater, writing task, and domain facets on logistic scale.

TABLE 1
Calibration of Rater Facet

Rater	*Severity*	*SE*	*INFIT MS*	*OUTFIT MS*	*Rater*	*Severity*	*SE*	*INFIT MS*	*OUTFIT MS*
11	1.78	.66	1.3	1.3	102	.01	.17	.9	.9
101	1.30	.21	.9	.8	93	−.02	.23	.9	1.0
82	1.29	.16	.8	.8	71	−.03	.26	1.3	1.3
106	1.13	.16	1.0	1.0	58	−.04	.16	1.3	1.3
77	1.11	.36	1.1	1.1	114	−.07	.18	.8	.8
66	1.06	.26	.9	1.0	113	−.08	.16	.8	.8
35	.94	.22	.8	.8	21	−.13	.24	.7	.6
18	.89	.19	1.0	1.0	94	−.15	.15	1.1	1.2
110	.85	.37	.8	.8	64	−.17	.28	.8	.8
3	.83	.42	.8	.9	12	−.28	.25	.9	.8
61	.81	.20	.8	.8	119	−.28	.50	1.1	1.1
96	.79	.19	.8	.7	73	−.30	.17	1.2	1.1
60	.76	.17	1.0	.9	118	−.31	.30	.9	.9
63	.76	.22	.9	.9	37	−.34	.16	1.0	1.0
31	.72	.20	1.2	1.3	4	−.36	.18	.9	.9
65	.72	.20	.9	.9	95	−.40	.25	1.0	1.0
79	.69	.19	.9	.9	112	−.40	.18	1.2	1.3
120	.59	.16	.9	.9	90	−.41	.32	1.1	1.1
41	.53	.13	1.1	1.0	104	−.41	.19	1.1	1.1
23	.50	.13	.8	.8	49	−.43	.16	1.1	1.1
48	.50	.15	1.1	1.1	116	−.46	.12	1.0	1.1
53	.49	.17	1.6	1.6*	51	−.49	.14	.9	.9
32	.45	.17	1.0	.9	57	−.49	.19	.9	.9
20	.42	.20	.6	.5*	97	−.52	.16	.9	.9
72	.37	.17	1.2	1.2	24	−.58	.29	1.0	.9
100	.37	.31	.8	.8	76	−.61	.20	.9	1.1
103	.34	.17	1.3	1.3	111	−.61	.16	1.4	1.4
14	.27	.16	1.0	1.1	86	−.63	.18	.8	.8
80	.23	.13	1.0	1.0	26	−.65	.14	.9	.9
105	.21	.22	.9	.9	59	−.68	.21	1.7	1.6*
69	.17	.17	1.0	1.0	27	−.76	.16	1.1	1.1
45	.16	.14	.8	.8	44	−.76	.22	1.4	1.3
74	.16	.14	.8	.8	115	−.83	.13	1.0	.8
16	.15	.26	.7	.7	55	−.85	.15	.9	1.3
7	.14	.33	.7	.7	109	−.97	.28	1.3	1.5
75	.11	.19	.5	.5*	117	−1.05	.30	1.0	1.0
52	.08	.22	1.0	1.0	6	−1.20	.60	.8	.7
70	.08	.17	.9	.9	89	−1.39	.13	1.0	1.2
40	.07	.16	.8	.8	87	−1.45	.22	1.5	1.6*
34	.06	.22	1.2	1.2	25	−1.61	.66	2.8	3.2*
85	.04	.21	1.6	1.6*	19	−1.74	.48	1.5	1.3
M	.00	.23	1.0	1.0					
SD	.70	.11	.3	.3					

Note. Asterisks indicate misfitting raters with OUTFIT mean squares *(MS)* less than .6 or greater than 1.5.

TABLE 2
Observed and Expected Ratings for Rater 59 with Noisy Rating Pattern (INFIT = 1.7, OUTFIT = 1.6)

Domain					*Domain*					*Domain*				
C/O	*S*	*SF*	*U*	*M*	*C/O*	*S*	*SF*	*U*	*M*	*C/O*	*S*	*SF*	*U*	*M*
Student 577					Student 852					Student 922				
2	2	2	2	3	2	2	4	3	3	3	3	3	2	3
2.45	2.30	2.51	2.47	2.45	2.50	2.35	2.57	2.52	2.50	2.93	2.80	2.98	2.95	2.93
−.45	−.30	−.51	−.47	.45	−.50	−.35	1.43*	.48	.50	.07	.20	.02	−.95	.07
Student 219					Student 296					Student 404				
3	3	3	2	2	4	3	3	4	4	3	3	4	4	4
2.26	2.14	2.32	2.28	2.26	3.28	3.14	3.34	3.30	3.27	3.21	3.08	3.27	3.23	3.21
.74	.86	.68	−.28	−.26	.72	−.14	−.34	.70	.73	−.21	−.08	.73	.77	.79
Student 850					Student 7					Student 314				
3	2	2	2	2	4	4	4	4	4	2	2	2	2	2
2.04	1.93	2.08	2.05	2.03	3.70	3.57	3.75	3.72	3.70	2.66	2.50	2.72	2.68	2.65
.96*	.07	−.08	−.05	−.03	.03	.43	.25	.28	.30	−.66	−.50	−.72	−.68	−.65
Student 619					Student 829					Student 916				
2	3	1	2	1	2	1	1	1	1	2	2	2	2	2
1.69	1.55	1.75	1.71	1.69	1.47	1.33	1.52	1.48	1.46	1.94	1.82	1.98	1.95	1.93
.31	1.45*	−.75	.29	−.69	.53	−.33	−.52	−.48	−.46	.06	.18	.02	.05	.07

(Continued)

TABLE 2 *(Continued)*

Domain					*Domain*					*Domain*				
C/O	*S*	*SF*	*U*	*M*	*C/O*	*S*	*SF*	*U*	*M*	*C/O*	*S*	*SF*	*U*	*M*
Student 572					Student 301					Student 920				
4	4	4	4	4	3	3	3	3	3	3	3	2	1	1
3.92	3.87	3.94	3.93	3.92	2.64	2.48	2.70	2.66	2.63	1.53	1.39	1.59	1.55	1.53
.08	.13	.06	.07	.08	.36	.52	.30	.34	.37	1.47*	1.61*	.41	−.55	−.53
Student 930					Student 975					Student 170				
4	4	4	2	3	3	4	2	3	4	2	1	4	4	3
2.94	2.81	2.99	2.96	2.94	2.92	2.78	2.97	2.93	2.91	2.68	2.52	2.73	2.69	2.67
1.06*	1.19*	1.01*	−.96*	.06	.08	1.22*	−.97*	.07	1.09*	−.68	−1.52*	1.27*	1.31*	.33
Student 750					Student 830									
2	1	2	2	1	3	3	3	2	3					
1.49	1.35	1.55	1.51	1.49	2.39	2.24	2.45	2.41	2.38					
.51	−.35	.45	.49	−.49	.61	.76	.55	−.41	.62					

Note. First row for each student is observed rating, second row is expected rating, and third row is the residual. Asterisks indicate residuals that are more than twice their standard errors. The five domains are content/organization (C/O), style (S), sentence formation (SF), usage (U), and mechanics (M).

TABLE 3
Calibration of Writing Task Facet

Writing Task	*Difficulty*	*SE*	*INFIT MS*	*OUTFIT MS*	*Theme Statement*
77	.12	.06	1.0	1.0	All-expense-paid trip
76	.06	.06	1.1	1.1	Your future
81	.05	.06	1.0	1.0	Your greatest hope for the future
80	.03	.06	1.0	1.0	Discovery or invention that makes life better
79	−.01	.06	1.0	1.0	Settle another planet
74	−.04	.06	.9	.9	Favorite holiday
78	−.04	.06	1.0	.9	Hero or heroine
75	−.16	.06	1.0	1.0	Experience that turned out better
M	.00	.06	1.00	1.00	
SD	.08	.00	.05	.06	

developers have not succeeded completely in constructing writing tasks of equal difficulty; adjustments are still necessary in the writing ability estimates to reflect differences in writing-task difficulty.

The calibration of the domains is presented in Table 4. It is hardest to get high ratings on Style with a difficulty of .50 logits (SE = .05), and easiest to get high ratings on Sentence Formation with a difficulty of −.28 logits (SE = .05). The overall differences between the domains are significant, $\chi^2(4, N = 5) = 162.05$, $p < .01$ with a high reliability of separation index (R = .97). There is no evidence of misfit for the domains with all of the fit mean squares between .9 and 1.1. The calibration of the steps within the rating scale from 1 to 4 with standard errors in parentheses is as follows: −4.12 (.06), .12 (.03), and 4.00 (.04). The observed proportions for the four categories from 1 to 4 are .05, .33, .47, and .16. This indicates that Categories 2 and 3 are the most frequently used by these raters.

TABLE 4
Calibration of Domain Facet

Domain	*Difficulty*	*SE*	*INFIT MS*	*OUTFIT MS*	*Domain Label*
2	.50	.05	1.0	1.0	Style (S)
5	−.04	.05	1.0	1.0	Mechanics (M)
1	−.06	.05	1.1	1.1	Content/Organization (C/O)
4	−.13	.05	.9	1.0	Usage (U)
3	−.28	.05	1.0	.9	Sentence Formation (SF)
M	.00	.05	1.00	1.00	
SD	.26	.00	.07	.07	

Raw scores are calculated by summing the 10 ratings for each student. The raw scores range from 10 to a maximum of 40 (2 raters × 5 domains × maximum rating of 4 for each domain). The operational version of the Eighth Grade Writing Test includes differential weights for each domain, but this weighting is not used in this example. Observed raw scores ranged from 10 to 40 ($M = 27.6$, $SD = 6.2$). Three of the students had minimum scores of 10, and 20 students had maximum scores of 40; two students had missing ratings for one or more domains. These 25 students were eliminated from the analysis, although procedures are available for assigning Rasch ability estimates for these students. On the logit scale, the Rasch estimates of writing ability ranged from 7.07 to −6.84 logits ($M = 1.02$, $SD = 2.46$). The overall differences between the students are significant, $\chi^2(974, N = 975) = 12188.58$, $p < .01$ with a high reliability of separation index ($R = .93$).

In order to illustrate the consequences of not adjusting the raw scores for rater effects, the ratings of four students are presented in the top panel of Table 5. Students 43 and 522 both have raw scores of 27 with identical rating patterns. Student 43 has a writing ability estimate of 1.16 logits ($SE = .63$), whereas Student 522 has an ability estimate of 1.99 logits ($SE = .61$). The difference of .83 logits appears because Student 522 was rated by two hard raters (R82 = 1.29; R106 = 1.13), whereas Student 43 was rated by one easy rater (R26 = −.65) and one hard rater (R106 = 1.13). If raw scores were used, then the writing ability of Student 522 would be underestimated. Students 621 and 305 also have the same raw score of 22 with identical rating patterns. Student 621 has a writing ability estimate of .08 logits ($SE = .66$), whereas Student 305 has an ability estimate of −1.18 logits ($SE = .65$). In this case the difference in writing ability estimates is 1.26 logits. Student 621 was rated by two hard raters (R66 = 1.06 and R82 = 1.29), whereas Student 305 was rated by an easy rater (R55 = −.85) and a hard rater (R61 = .81). If raw scores were used, then the writing ability of Student 621 would be underestimated.

The bottom panel in Table 5 presents three examples of misfitting rating patterns for students. Student 532 has a noisy rating pattern (INFIT = 2.7; OUTFIT = 2.7). This student received unexpectedly low ratings of 1 by both raters in the domain of usage. Student 735 also has a noisy rating pattern (INFIT = 2.4; OUTFIT = 2.4) with an unexpected high rating in usage by Rater 49, and an unexpected low rating in sentence formation by Rater 111. Student 861 has a muted rating pattern (INFIT = .1; OUTFIT = .1) with both raters assigning this student ratings of 2 in every domain. Misfitting student essays should be examined in detail to determine whether or not there is anything unusual about them, such as illegible handwriting, an off-topic essay, or a controversial response.

TABLE 5
Observed and Expected Ratings for Selected Students

Student	Domain C/O	S	SF	U	M	Domain C/O	S	SF	U	M	Raw Score	INFIT MS	OUTFIT MS	Rasch Ability
Consistent Rating Patterns														
	Rater 26					Rater 106								
43	2	3	3	3	3	3	3	3	2	2	27	.9	.9	1.16
	2.95	2.82	3.00	2.96	2.94	2.47	2.33	2.54	2.50	2.47				
	−.95	.18	.00	.04	.06	.53	.67	.46	−.50	−.47				
	Rater 82					Rater 106								
522	2	3	3	3	3	3	3	3	2	2	27	.9	.9	1.99
	2.70	2.55	2.76	2.72	2.69	2.74	2.59	2.80	2.76	2.73				
	−.70	.45	.24	.28	.31	.26	.41	.20	−.76	−.73				
	Rater 66					Rater 82								
621	2	2	3	2	3	2	2	2	2	2	22	.6	.6	.08
	2.23	2.11	2.29	2.55	2.23	2.18	2.07	2.24	2.20	2.18				
	−.23	−.11	.71	−.25	.77	−.18	−.07	−.24	−.20	−.18				
	Rater 55					Rater 61								
305	2	2	3	2	3	2	2	2	2	2	22	.5	.4	−1.18
	2.40	2.26	2.47	2.42	2.40	2.01	1.90	2.06	2.03	2.01				
	−.40	−.26	.53	−.42	.60	−.01	.10	−.06	−.03	−.01				

(Continued)

TABLE 5 *(Continued)*

Student	Domain C/O	S	SF	U	M	Domain C/O	S	SF	U	M	Raw Score	INFIT MS	OUTFIT MS	Rasch Ability
Misfitting Rating Patterns														
	Rater 58					Rater 106								
532	3	2	3	1	2	3	3	3	1	2	23	2.7	2.7	−.14
	2.46	2.31	2.53	2.48	2.46	2.16	2.05	2.22	2.18	2.16				
	.54	−.31	.47	−1.48*	−.46	.84	.95	.78	−1.18*	−.16				
	Rater 49					Rater 111								
735	2	3	2	4	3	3	2	1	3	2	25	2.4	2.4	−.58
	2.49	2.34	2.56	2.51	2.49	2.53	2.38	2.60	2.55	2.52				
	−.49	.66	−.56	1.49*	.51	.47	−.38	−1.60*	.45	−.52				
	Rater 21					Rater 106								
861	2	2	2	2	2	2	2	2	2	2	20	.1	.1	−1.38
	2.13	2.02	2.19	2.15	2.13	1.88	1.76	1.93	1.90	1.88				
	−.13	−.02	−.19	−.15	−.13	.12	.24	.07	.10	.12				

Note. First row for each student is observed rating, second row is expected rating, and third row is the residual. Asterisks indicate residuals that are more than twice their standard errors. The five domains are content/organization (C/O), style (S), sentence formation (SF), usage (U), and mechanics (M).

DISCUSSION

When the measurement of writing ability is based directly on student essays, there are many factors in addition to writing ability that can contribute to variability in the observed essay scores. Some of the major factors are differences in (a) rater severity (Lunz, Wright, & Linacre, 1990), (b) writing-task difficulty (Engelhard et al., 1992; Ruth & Murphy, 1988), (c) domain difficulty when analytic scoring is used, (d) examinee characteristics other than ability (Brown, 1986), and (e) the structure of the rating scale. Ideally, the estimate of an individual's writing ability should be independent of the particular raters, writing tasks, and domains that happen to be used. Further, examinee characteristics apart from writing ability (such as gender, race, ethnicity, and social class) should not influence the validity of the estimates of writing ability. The FACETS model described by Linacre (1989) provides a coherent framework for obtaining estimates of writing ability that are invariant over raters, writing tasks, and domains when the data fit the model. Issues related to bias can also be explored with the FACETS model. The FACETS model provides a framework for obtaining objective linear measurements of writing ability that generalize beyond the specific raters, writing tasks, and domains that happen to be used to obtain the observed rating. The FACETS model can also be applied to the assessment of writing ability based on holistic scoring procedures (Cooper, 1977). The structure of the rating scale can also be modeled using a partial credit model rather than the rating scale structure used here.

The FACETS model provides the following advantages over other measurement models that have been used within the context of writing assessments:

1. The FACETS model is a measurement model based on a linear logistic transformation of the observed scores. The estimates of writing ability are specified to be on an equal-interval scale, in contrast to the ordinal scale underlying the raw score-based approaches using ANOVA models or linear structural equation models.
2. The FACETS model provides an explicit approach for examining the multiple facets encountered in the design of most writing assessments. A sound theoretical framework is provided for adjusting for differences in raters and writing tasks. Adjustments for rater severity and writing-task difficulty improve the objectivity and fairness of the measurement of writing ability because unadjusted scores lead to under- or overestimates of writing ability when students are rated by different raters on different writing tasks. Further, an explicit set of procedures is provided for monitoring the quality of ratings obtained from raters.
3. The FACETS model is a Rasch measurement model, and possesses

desirable statistical and psychometric properties related to the separability of parameters with sufficient statistics available for estimating these parameters.

4. If the data fit the FACETS model, then invariant estimates of writing ability, rater severity, and writing-task difficulty can be obtained that generalize beyond the specifics of the local writing assessment procedures. Tests of fit and residual analyses are available to examine in detail whether or not the data fit the FACETS model, and these desirable invariance properties achieved.

5. The creation of rater and writing-task banks that are conceptually similar to item banks is straightforward, and can be viewed as simple extensions of current item banking procedures. When the data fit the FACETS model, the creation of rater and writing-task banks becomes simply a matter of adding and subtracting the appropriate linking constants. Once the banks are created, then the equating of the ratings for the influences of raters and writing tasks is straightforward. However, these banks must be continually maintained and validated.

6. Incomplete research designs with missing cells and other forms of missing data, as well as nested designs, can be handled routinely if attention is paid to the construction of a connected network of links within and between facets. Misfitting observations can be identified for diagnostic purposes and corrective actions taken when needed.

7. DFF can be examined within different groups (gender, race, and social class) in order to examine bias issues. This can be accomplished by calibrating the facets separately within relevant groups, and examining whether or not the relative difficulty of the components of the facet are invariant over groups. Interactions between facets can also be examined as a potential source of bias in the assessment of writing ability.

In summary, the FACETS model offers a promising approach for solving a variety of measurement problems encountered in the large-scale assessment of writing ability. Progress in addressing measurement problems within the context of direct assessment of writing ability is reflected in the movement from ad hoc to model-based methods, such as the FACETS model, for examining key aspects of the assessment process and providing an approach for minimizing these measurement problems. The empirical example presented here was intended to illustrate the FACETS model, and not intended to provide a definitive examination of its usefulness for solving these measurement problems. Additional research is needed to examine further the FACETS model within the context of the large-scale assessment of writing ability. This research should address in detail the problems encountered in the development of calibrated rater banks using the FACETS model. Further research on the use of the FACETS model to

address measurement problems encountered in the development of operational writing-task banks for a large-scale assessment of writing ability is also needed. Finally, research is needed on differential facet functioning related to gender, race, and social class; this research will contribute to our knowledge regarding the use of the FACETS model to examine potential sources of bias in large-scale writing assessments.

ACKNOWLEDGMENTS

Earlier versions of this article were presented at the annual meeting of the American Educational Research Association in Chicago (April 1991) and the Midwest Objective Measurement Seminar at the University of Chicago (December 1991).

Richard M. Jaeger, Mike Linacre, Judith A. Monsaas, and Vanessa Siddle Walker provided helpful comments on earlier drafts of this article.

REFERENCES

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice and free-response writing tests. *Applied Psychological Measurement, 2,* 117–128.

Afflerbach, P. (1985). *The statewide assessment of writing.* Princeton, NJ: Educational Testing Service.

Andrich, D. (1973). *Latent trait psychometric theory in the measurement and evaluation of essay writing ability.* Unpublished doctoral dissertation, The University of Chicago.

Applebee, A. N., Langer, J. A., Jenkins, L. B., Mullis, I., & Foertsch, M. A. (1990). *Learning to write in our nation's schools: Instruction and achievement in 1988 at grades 4, 8 and 12.* Princeton, NJ: Educational Testing Service.

Applebee, A. N., Langer, J. A., & Mullis, I. (1985). *Writing: Trends across the decade, 1974–1984.* Princeton, NJ: Educational Testing Service.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22,* 41–52.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13,* 1–18.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill.* New York: College Entrance Examination Board.

Brown, R. C. (1986). Testing black student writers. In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 98–108). New York: Longman.

Bunch, M. B., & Littlefair, W. (1988). *Total score reliability in large-scale writing assessment.* Paper presented at the conference of the Education Commission of the States, Boulder, CO. (ERIC Document Reproduction Service No. ED 310 149)

Choppin, B. H. (1982). The use of latent trait models in the measurement of cognitive abilities and skills. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology* (pp. 41–63). Melbourne: Australian Council for Educational Research.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 271–302). Washington, DC: American Council on Education.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging.* Buffalo: State University of New York.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

De Gruijter, D. N. M. (1984). Two simple models for rater effects. *Applied Psychological Measurement, 8,* 213–218.

Engelhard, G., Anderson, D., & Gabrielson, S. (1990). An empirical comparison of Mantel-Haenszel and Rasch procedures for studying differential item functioning on teacher certification tests. *Journal of Research and Development in Education, 23,* 172–179.

Engelhard, G., Gordon, B., & Gabrielson, S. (1992). The influences of mode of discourse, experiential demand and gender on the quality of student writing. *Research in the Teaching of English, 26*(3), 315–336.

Ferrara, S., & Walker-Bartnick, L. (1989, April). *Constructing an essay prompt bank using the Partial Credit model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Gorman, T. P., Purves, A. C., & Degenhart, R. E. (Eds.). (1988). *The IEA study of written composition I: Writing tasks and scoring scales.* Oxford, England: Pergamon.

Hake, R. (1986). How do we judge what they write? In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 153–167). New York: Longman.

Harris, J., Laan, S., & Mossenson, L. (1988). Applying partial credit analysis to the construction of narrative writing tests. *Applied Measurement in Education, 1,* 335–346.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Joreskog, K. G., & Sorbom, D. (Eds.). (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Kuhn, T. (1970). *Structure of scientific revolutions* (2nd ed.). Chicago: The University of Chicago Press.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied Measurement in Education, 2,* 195–205.

Laudan, L. (1977). *Progress and its problems.* Berkeley: University of California Press.

Linacre, J. M. (1988). *FACETS: Computer Program for Many-Faceted Rasch Measurement.* Chicago: MESA Press.

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

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging.* Buffalo: State University of New York.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3,* 331–345.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing.* Norwood, NJ: Ablex.

Stanley, J. C. (1962). Analysis-of-variance principles applied to the grading of essay tests. *Journal of Experimental Education, 30,* 279–283.

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: Rasch measurement.* Chicago: MESA Press.