DOCUMENT RESUME

ED 334 259 TM 016 866

AUTHOR Engelhard, George, Jr.
TITLE The Measurement of Writing Ability with a Many-Faceted Rasch Model.
PUB DATE Apr 91
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991). Table 2 may not reproduce well.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Grade 8; Holistic Evaluation; *Interrater Reliability; *Item Response Theory; Junior High Schools; *Multivariate Analysis; Test Bias; *Writing Evaluation; Writing Tests
IDENTIFIERS *FACETS Model; Georgia Basic Skills Writing Test; *Rasch Model

ABSTRACT

A many-faceted Rasch model (FACETS) is presented 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. A FACETS model is described based on the current procedures of the Georgia Basic Skills Writing Test. These procedures can be seen as a prototype for other statewide assessments of writing. A small data set produced by 15 eighth grade students from this assessment is analyzed with the FACETS computer program of J. M. Linacre (1989). The FACETS model offers a promising approach for solving a variety of measurement problems in the statewide assessment of writing ability. It provides a framework for obtaining objective linear measurements of writing ability that generalize beyond specific raters and writing tasks. The FACETS model can also be applied to the assessment of writing ability based on holistic scoring and exploration of issues related to bias. A 37-item list of references, 5 tables, and 2 figures are included. (SLD)

THE MEASUREMENT OF WRITING ABILITY WITH A MANY-FACETED RASCH MODEL

George Engelhard, Jr.

Emory University

Address: Professor George Engelhard, Jr.
Emory University
Division of Educational Studies
210 Fishburne Building
Atlanta, GA 30322

(404) 727-0607 (w)
(404) 525-1115 (h)

Running head: Measurement of writing ability

[writing2 - Paper presented at the annual meeting of the American Educational Research Association in Chicago, April 1991]

April 26, 1991

Abstract

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 assessment of writing ability is presented here. A small data set from a statewide assessment of writing ability is used to illustrate the FACETS model.

THE MEASUREMENT OF WRITING ABILITY WITH A MANY-FACETED RASCH MODEL

Direct assessments of student writing ability are currently being conducted or planned in almost every state (Afflerbach, 1985). These statewide writing assessments are generally 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 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 non-dichotomous 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). Recent work 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 the 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, and Linacre, 1990) and inter-rater 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 statewide 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 (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 they 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 paper 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 models and the other on linear structural equation models. Examples of approaches to writing assessment based on analysis of variance models are the early work of Stanley (196_) 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 by several researchers (Bunch & Littlefair, 1988; Lane & Sabers, 1989). Blok (1985) and Ackerman & 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 non-linear 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 summarized as a raw score as the unit of analysis for both dichotomous and polytomous response data. Item response models can be developed to directly model the probability of a student obtaining a particular set of ratings based on an actual writing sample. Although an empirical comparison of different approaches to 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 modelling 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, Jaan & 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 are all essentially two facet models (writing ability and rater

severity), 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 problem encountered with other approaches to writing assessment. Rasch measurement models can provide a framework for obtaining objective and fair measurements of writing ability which 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 Basic Skills Writing Test (BSWT). Georgia's procedures can serve as a prototype for other statewide assessments of writing. Next, a small data set 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 statewide 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.

Insert Figure 1 about here

The dependent variable in the model is the observed rating which ranges from 0 to 3 (0 = inadequate, 1 = minimal, 2 = good, 3 = very good). The four major facets that influence this rating are writing ability, rater severity, the difficulty of the writing task and domain difficulty. The structure of the rating scale which defines the categories also affects the value of the rating obtained. Other statewide assessment 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.

Although not explicitly included in the measurement model, other student characteristics that reflect potential sources of bias may affect the observed rating of a student. Some examples of these student characteristics are gender, age, ethnicity, 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). 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 Many-Faceted Rasch Model

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

$$\log [Pnijmk/Pnijmk-1] = Bn - Ti - Rj - Dm - Fk$$

where

Pnijmk = probability of student n being rated k on writing task i by rater j for domain m

Pnijmk-1 = probability of student n being rated k-1 on writing task i by rater j for domain m

Bn = Writing ability of student n

Ti = Difficulty of writing task i

Rj = Severity of rater j

Dm = Difficulty of domain m

Fk = Difficulty of rating Step k relative to Step k-1

The student facet, Bn, provides a measure of writing ability on a linear logistic scale (logits) that ranges from +/- infinity. If the data fits 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, Ti, 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, Rj, are also obtained on the same linear logistic scale which are invariant over students and writing tasks. Finally, invariant calibrations of the domain facet, Dm, and rating scale step difficulties, Fk, are also obtained. The FACETS model is an additive linear model based on this logistic transformation to a logit scale.

Empirical Example

Subjects

Fifteen students were randomly selected from the Spring 1989 administration of the Basic Skills Writing Test (BSWT) that is administered to all of the eighth-grade students in Georgia. Seven of the students are female and eight are males; six of the students are black and nine are white.

Instrument

The BSWT is a criterion-referenced test designed to provide a direct assessment of student writing ability. Students are asked to write an essay of no more than two pages on an assigned writing task. The writing tasks are randomly assigned to the students. Each of the essays is rated by two raters on the following five domains: content/organization, style, sentence formation, usage and mechanics. A four category rating scale is used for each domain (0=inadequate, 1=minimal, 2=good and 3=very good). The final response pattern used to estimate student writing ability consists of ten ratings (two raters x five domains = ten ratings). Additional information on the BSWT is available in the Teacher's Guide (Georgia Department of Education, 1990).

The raters are highly trained and a variety of procedures are used to maintain the reliability and validity of the ratings. First, the raters must successfully complete an extensive training program; this program typically takes three 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 percent exact agreement with the standard on the ratings and 38 percent 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 re-scored by a third rater. Further details of the training procedures and the ongoing quality control processes are available in the Training Manual (Georgia Department of Education, 1989).

Although the full rhetorical specification of the writing tasks can not be revealed because this is a high-stakes test, the theme statements for the tasks examined here are "where you would go if you won an all expense paid trip" (Task 72) and "time you were successful" (Task 63). The mode of discourse for both of these tasks is narration.

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. The rating scale model with common step sizes across domains was used for the structure of the rating scale. The program calculates several fit statistics that provide evidence regarding the validity of the FACETS model. The standardized fit statistic is reported here which is based on a

transformation of the unweighted mean square residuals to an approximate t distribution (Wright & Masters, 1982). This standardized fit statistic is sometimes referred to as the outfit statistic because it is sensitive to outlying deviations from the expected values. The standardized fit statistics are rounded to the nearest integer by the FACETS program. For the purposes of this study, obtained values for standardized fit statistics that are less than 2 are interpreted as indicating acceptable fit to the FACETS model. In addition to the standardized fit statistic, a reliability coefficient which is similar to KR-20 (ratio of true score variance to observed score variance) is reported for each facet. Additional details regarding the computational and statistical aspects of the FACETS model are presented in Linacre (1989).

Results

The observed ratings for the 15 students are presented in Table 1. For this example, two writing tasks (63 & 72)

Insert Table 1 about here

---

appeared and were rated by three raters (117, 197 and 232).

The calibration of the raters, writing tasks and domains on the linear logistic scale are shown in Figure 2. Task 72 is harder

---

Insert Figure 2 about here

---

with a difficulty of .34 logits (SE = .27) as compared to Task 63 with a difficulty of -.34 logits (SE = .26). The reliability for the writing tasks is .42, p = .06, which suggests that the difference in the difficulties of these two writing tasks is close enough to the traditional critical value ($p < .05$) to be considered statistically significant. Both standardized fit statistics were less than 2; the obtained value for Task 72 is 0 and the value for Task 63 is -1.

Rater 197 (R197 = 1.58, SE = .53) is more severe than the other two raters (R117 = .-.57, SE = .30; R232 = -1.00, SE = .26). The reliability coefficient for the raters is .88, $p < .01$ which indicates that there is significant variation among the raters beyond the variation due to estimation error. This significant variation in the raters appears in spite of the extensive training and screening of the raters. The standardized fit statistics indicate that intra-rater consistency is acceptable with observed values of 1, 0 and -1 for raters 197, 117 and 232 respectively.

Turning to the five domains, the order of difficulty from hard to easy on the logistic scale is as follows: usage (D1 = .64, SE = .4°), style (D2 = .30, SE = .42), sentence formation (D3 = -.03, SE = .41), mechanics (D4 = -.37, SE = .41) and content/organization (D5 = -.54, SE = .42). The reliability is .08, p = .25 which suggests that there are not statistically significant differences in the relative difficulties of these five

domains. None of the standardized fit statistics is greater than 2, and the observed values for four of the domains are 0s, and -1 for sentence formation.

The calibration of the steps within the rating scales from 0 to 3 with standard errors in parentheses are as follows: -6.57 (.40), .32 (.23) and 6.25 (.48). The observed proportions for the four categories from 0 to 3 are .09, .48, .37 and .06. This indicates that categories 1 and 2 are the most frequently used by these raters.

Raw scores are calculated by summing the ten ratings for each student. The raw scores range from 0 to a maximum of 30 (two raters x five domains x maximum rating of 3 for each domain). The operational version of the BSWT includes differential weights for each domain, but this weighting is not used in the present example. Observed raw scores ranged from 5 to 25 (M = 13.9, SD = 5.9). The Rasch estimates of writing ability for these 15 students are presented in Table 1, and these values ranged from -6.26 to 6.32 logits (M = -1.08, SD = 3.68). The reliability coefficient is quite high for the student ability estimates (REL = .96, $p < .01$). The correlation between the raw scores and the Rasch estimates is high, $r(13) = .98$, $p < .01$. This high correlation does not, however, eliminate the possibility that some raw scores are biased by variation in rater severity and writing task difficulty.

In order to illustrate the consequences of not adjusting raw scores for rater and writing task effects, the ratings for two

students with the same raw scores of 8 are presented in Table 2.

Insert Table 2 about here

These students have identical rating patterns, and yet the writing ability of Student 12 (B12 = -4.12) is estimated to be 2.0 logits greater than Student 4 (B4 = -6.26). This difference in estimated writing ability is observed because Student 12 happened to be rated by Rater 197 who is more severe than Rater 117.

The fit statistics for the Rasch ability estimates are presented in Table 1. The observed values of the standardized fit statistic show acceptable fit of the data to the model for all of the students except Student 8. Student 8 has an observed fit statistic of 2 and a detailed residual analysis for this student is presented in Table 2. For comparison purposes, the rating patterns for Students 4 and 12 who both have consistent ratings with standardized fit statistics close to zero are also presented in Table 1. Fit statistics less than 2 indicate a close correspondence between the observed and expected ratings. For Student 4, only one of the standardized residuals is greater than twice its standard error, while none of the standardized residuals are significant for Student 12. Student 8 has three unexpectedly high ratings from Rater 197 in style, usage and mechanics. This essay should be examined in detail to determine whether or not

there is anything unusual about it, such as illegible handwriting, an off-topic essay or a controversial response.

In order to illustrate the consequences of not adjusting for differences in writing task difficulty, unadjusted estimates of writing ability were calculated. These are presented in Table 3.

Insert Table 3 about here

The data suggest that if students were asked to respond to Task 63, then their writing abilities would on the average be over estimated by .34 logits; if they were asked to respond to Task 72, then their writing abilities would be under estimated by -.38 logits. This is due to the differences in writing task difficulty with Task 72 being relatively more difficult than Task 63.

A similar analysis was conducted for the influence of raters, and these results are presented in Table 4. Students who were rated

---

Insert Table 4 about here

---

by Raters 117 and 232 tend to have their writing abilities over estimated (M = .70, SD = .37), while students who were rated by Raters 197 and 232 tend to have their writing abilities under estimated (M = -.40, SD = .37). This effect is due to the large differences in rater severity between Rater 197 (R197 = 1.58) who tends to be more severe than Rater 117 (R117 = -.57) who tends to be more lenient.

When adjustments for differences in both writing task difficulty and rater severity are not made, then the average differences between the adjusted and unadjusted estimates of writing ability is .45 logits (SD = .71). These results are shown

---

Insert Table 5 about here

---

in Table 5. Since the effects of writing task difficulty and rater severity are additive, some of the students have their writing abilities over estimated by more than 1.00 logit (Students 1 to 7) if the unadjusted estimates are used.

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 (1) rater severity (Lunz, Wright, & Linacre, 1990), (2) writing task difficulty (Ruth & Murphy, 1988), (3) domain difficulty when analytic scoring is used, (4) examinee characteristics other than ability (Brown, 1986) and (5) 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 Many-Faceted Rasch (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. 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 and writing tasks 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 modelled 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 scaling 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 based on analysis of variance 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 over estimates of writing ability when students are rated by different raters on different writing tasks.

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 which generalize beyond the specifics of the local writing assessment procedures. Tests of fit and residual analyses are available to examine whether or not the data fit the FACETS model, and these desirable invariance properties achieved.

5. The creation of rater and writing-task banks is straight forward, 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 straight forward. These banks, however, must be continually maintained and validated.

6. Incomplete research designs with missing cells and other forms of missing data 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. Differential facet functioning (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 statewide assessment of writing ability. The small 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 based on operational forms with large writing samples is needed to further examine the FACETS model within the context of the statewide 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 statewide assessment of writing ability is also needed. And 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 statewide writing assessments.

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, 12, 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., & Mullis, I. (1985). Writing: Trends across the decade, 1974-1984. Princeton, NJ: Educational Testing Service.

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.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. Journal of Educational Measurement, 22(1), 41-52.

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

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 number 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. Melbourne: Australian Council for Educational Research.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), Educational measurement, Second Edition, (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 at Buffalo.

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: John Wiley & Sons.

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(2), 172-179.

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

Georgia Department of Education. (1990). Georgia Basic Skills Writing Test: Teacher's Guide. Atlanta, GA: Author.

Georgia Department of Education. (1989). Georgia Basic Skills Writing Test: Training Manual. Atlanta, GA: Author.

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

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.

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

Kuhn, T. (1970). Structure of scientific revolutions. 2nd enlarged edition. 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(3), 195-205.

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

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

Linacre, J. M. (1988). FACETS: Computer Program for 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 at Buffalo.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 3(4), 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(1), 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 Publishing Company.

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., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press.

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

Table 1

Observed ratings of 15 students by three raters on two writing tasks with five domains

Student	Rater 117					Rater 197					Rater 232					Raw Score	Rasch Ability	SE	FIT
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5				
Task 63																			
1	2	2	2	1	2	.	.	.	.	.	2	1	2	1	1	16	-.40	.65	0
2	1	1	1	1	1	.	.	.	.	.	2	1	1	1	1	11	-2.99	.97	0
3	2	2	3	2	3	.	.	.	.	.	3	2	3	2	3	25	5.13	.65	-1
4	0	0	1	1	1	.	.	.	.	.	1	1	1	1	1	8	-6.26	.79	0
5	2	1	1	1	1	.	.	.	.	.	2	2	1	1	2	14	-1.24	.86	-1
6	2	1	1	2	1	.	.	.	.	.	2	2	2	2	2	17	.05	.69	0
7	2	2	1	1	1	.	.	.	.	.	2	2	1	1	1	14	-1.24	.66	0
Task 72																			
8	.	.	.	.	.	2	3	2	3	3	2	2	2	2	3	24	6.32	.75	2
9	2	2	2	1	2	.	.	.	.	.	2	2	2	1	2	18	1.25	.76	0
10	.	.	.	.	.	0	0	0	0	0	1	1	1	1	1	5	-5.94	.76	-1
11	2	1	2	2	2	.	.	.	.	.	2	1	1	1	2	16	.28	.65	0
12	.	.	.	.	.	0	0	1	1	1	1	1	1	1	1	8	-4.12	.82	0
13	1	2	1	2	2	.	.	.	.	.	2	1	1	1	1	14	-.56	.66	1
14	.	.	.	.	.	0	0	0	0	0	1	1	1	1	1	5	-5.94	.76	-1
15	1	1	2	1	1	.	.	.	.	.	1	2	2	2	1	14	-.56	.66	0

Note. Each student is rated by two raters using a four-category scale (0=inadequate, 1=minimal, 2=good, 3=very good) on each domain. The five domains are (1) content/organization, (2) style, (3) sentence formation, (4) usage and (5) mechanics. SE is the standard error of the Rasch estimate of writing ability, and FIT is the standardized fit statistic (F. ' values less than 2 indicate that the rating pattern fits the model.

Table 2

Observed and expected ratings for selected students

Student		Rater 117					Rater 197					Rater 232					Raw Score	Rasch Ability	FIT
		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5			
Consistent Ratings																			
4	Observed	0	0	1	1	1	.	.	.	.	.	1	1	1	1	1	8	-6.26	0
	Expected	.86	.72	.78	.64	.84	.	.	.	.	.	.91	.80	.85	.74	.89			
	Residual	-.86*	-.72	.22	.36	.16	.	.	.	.	.	.09	.20	.15	.26	.11			
12	Observed	.	.	.	.	.	0	0	1	1	1	1	1	1	1	1	8	-4.12	0
	Expected	.	.	.	.	.	.75	.56	.54	.47	.71	1.01	.96	.98	.93	1.00			
	Residual	.	.	.	.	.	-.75	-.56	.36	.53	.29	-.01	.04	.02	.07	.00			
Inconsistent Ratings																			
8	Observed	.	.	.	.	.	2	3	2	3	3	2	2	2	2	3	24	6.32	2
	Expected	.	.	.	.	.	2.30	2.08	2.12	2.05	2.17	2.78	2.60	2.68	2.52	2.75			
	Residual	.	.	.	.	.	-.20	.92*	-.12	.95*	.83*	-.78	-.60	-.68	-.52	.25			

Note. Residual is the difference between the observed and expected ratings. Asterisks indicate residuals that are more than twice their standard errors. The five domains are (1) content/organization, (2) style, (3) sentence formation, (4) usage and (5) mechanics.

Table 3

Comparison of writing ability estimates adjusted and unadjusted for differences in writing task difficulty

Student	Adjusted Estimate	Unadjusted Estimate	Difference Task 63	Difference Task 72
1	-.40	-.05	.35	
2	-2.99	-2.66	.33	
3	5.13	5.61	.48	
4	-6.26	-6.09	.17	
5	-1.24	-.90	.34	
6	.05	.40	.35	
7	-1.24	-.90	.34	
8	6.32	6.14		-.18
9	1.25	.93		-.32
10	-5.94	-6.46		-.52
11	.28	-.05		-.33
12	-4.12	-4.59		-.47
13	-.56	-.90		-.34
14	-5.94	-6.46		-.52
15	-.56	-.90		-.34
Mean	-1.08	-1.12	.34	-.38
SD	3.68	3.81	.09	.11

Note. The adjusted ability estimates are the same as the Rasch ability estimates of writing ability reported in Table 1. Differences are based on unadjusted minus adjusted estimates of writing ability. Negative values indicate under estimates, while positive values indicate over estimates of writing ability.

Table 4

Comparison of writing ability estimates adjusted and unadjusted for differences in rater difficulty

Student	Adjusted Estimate	Unadjusted Estimate	Difference 117,232	Difference 197,232
1	-.40	.27	.67	
2	-2.99	-2.30	.69	
3	5.13	6.15	1.02	
4	-6.26	-5.62	.64	
5	-1.24	-.56	.68	
6	.05	.71	.66	
7	-1.24	-.56	.68	
8	6.32	6.38		.06
9	1.25	1.92	.67	
10	-5.94	-6.33		-.39
11	.28	.94	.66	
12	-4.12	-4.97		-.85
13	-.56	.11	.67	
14	-5.94	-6.33		-.39
15	-.56	.11	.67	
Mean	-1.08	-.67	.70	-.40
SD	3.68	3.96	.37	.37

Note. The adjusted ability estimates are the same as the Rasch ability estimates of writing ability reported in Table 1. Differences are based on unadjusted minus adjusted estimates of writing ability. Negative values indicate under estimates, while positive values indicate over estimates of writing ability.

Table 5

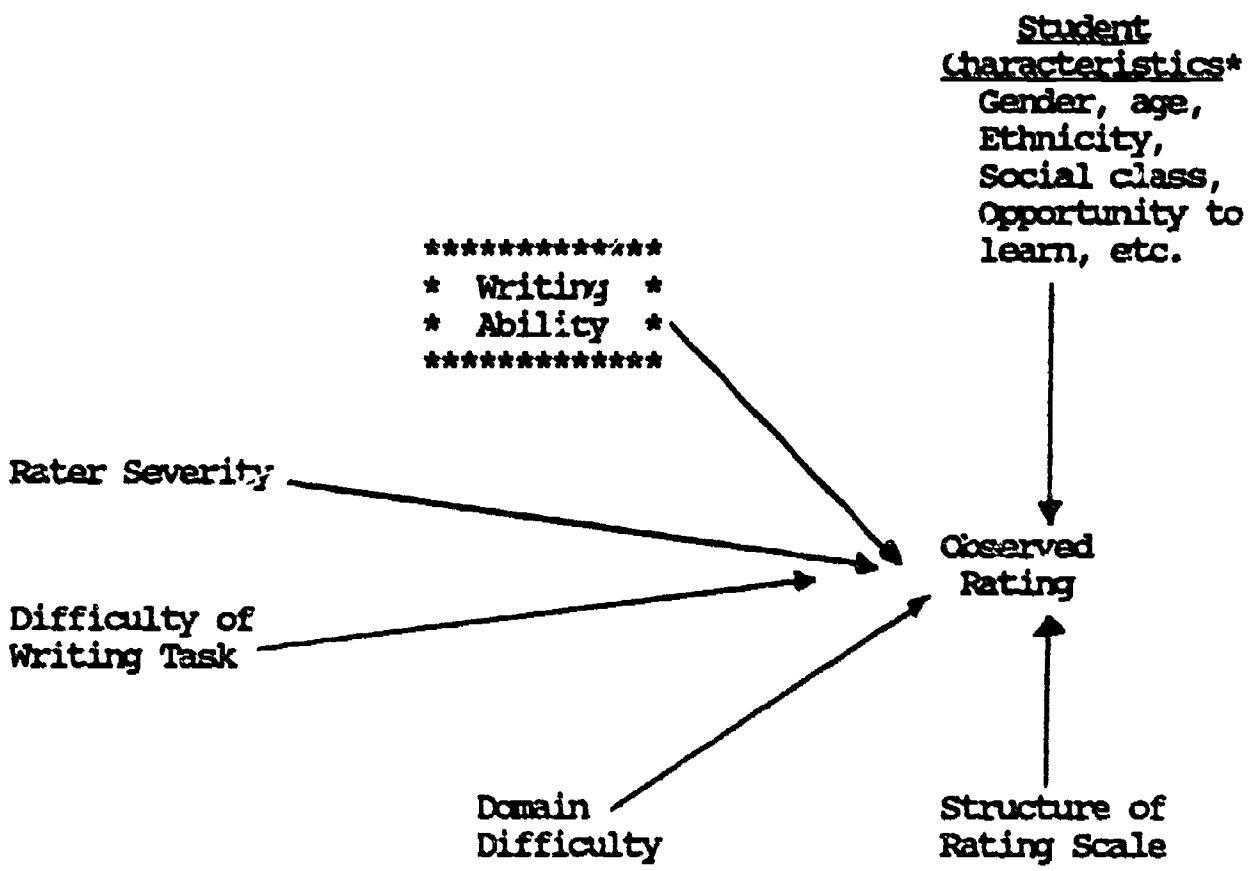
Comparison of writing ability estimates adjusted and unadjusted for differences in rater and writing task difficulty

Student	Adjusted Estimate	Unadjusted Estimate	Difference
1	-.40	.62	1.02
2	-2.99	-1.90	1.09
3	5.13	6.16	1.03
4	-6.26	-4.95	1.31
5	-1.24	-.20	1.04
6	.05	1.06	1.01
7	-1.24	-.20	1.04
8	6.32	5.74	-.58
9	1.25	1.58	.33
10	-5.94	-6.35	-.41
11	.28	.62	.34
12	-4.12	-4.95	-.83
13	-.56	-.20	.36
14	-5.94	-6.35	-.41
15	-.56	-.20	.36
Mean	-1.08	-.63	.45
SD	3.68	3.80	.71

Note. The adjusted ability estimates are the same as the Rasch ability estimates of writing ability reported in Table 1. Differences are based on unadjusted minus adjusted estimates of writing ability. Negative values indicate under estimates, while positive values indicate over estimates of writing ability.

Figure 1

Measurement model for the assessment of writing ability



* Potential bias factors that are not explicitly included in the measurement model.

Figure 2

Calibration of raters, writing tasks and domains on logistic scale