The purpose of this study was to describe the Rasch model for measurement and apply the model to examine the relationship between raters, domains of written compositions, and student writing ability. Twenty raters were randomly selected from a group of 87 operational raters contracted to rate essays as part of the 1993 field test of the Georgia High School Writing Test. Each rater rated the complete set of 366 essays written by high school students and each essay was scored on 4 domains: (1) content and organization; (2) style; (3) conventions; and (4) sentence formation. The Rasch model-based FACETS computer program (J. Linacre and B. Wright, 1994) was used to examine whether there were statistically significant differences in rater severity and domain difficulty, and to explore the rater by domain interaction effect. Results indicate significant differences between raters, between domains, and a significant interaction effect between raters and domains. (Contains 4 tables, 8 figures, and 20 references.) (Author/SLD)
APPLYING THE RASCH MODEL TO EXPLORE RATER INFLUENCES ON THE ASSESSED QUALITY OF STUDENTS' WRITING ABILITY

Ismail S. Gyagenda
and
George Engelhard, Jr.
Emory University

Division of Educational Studies
Emory University
North Decatur Building
Atlanta, GA 30322

Abstract

The purpose of this study was to describe the Rasch Model for measurement and apply the model to examine the relationship between raters, domains of written compositions, and student writing ability. Twenty raters were randomly selected from a group of 87 operational raters contracted to rate essays as part of the 1993 field test of the Georgia High School Writing Test. Each rater rated the complete set of three hundred sixty six essays written by high school students and each essay was scored on four domains: a) content and organization, b) style, c) conventions, and d) sentence formation. The Rasch model based FACETS computer program (Linacre & Wright, 1984) was used to examine whether there were statistically significant differences in rater severity and domain difficulty, and to explore the rater by domain interaction effect. Results indicated significant differences between raters, between domains, and a significant interaction effect between raters and domains.
Applying the Rasch model to explore rater influences on the assessed quality of students' writing ability.

Ability and achievement measurement is problematic. The meaning and estimated quality of the traditional scales of measurement "depend upon the specific set of items actually standardized and the particular ability distribution of the children who happened to appear in the standardizing sample" (Wright, 1967, p. 86). This sample and item dependence challenges the requisite objectivity of ability measuring instruments. If height or weight measuring instruments depended for their accuracy on the persons measured or their texture, color, etc., their objectivity would be similarly challenged. It is critical, therefore, to develop objective measures in the social sciences. This paper describes the Rasch model of measurement, which aims at meeting this requirement, and illustrates its application by exploring the relationship between raters, domains of written compositions, and writing ability.

Student writing ability is increasingly being assessed nationally and in most states (Afflerbach, 1985; Applebee, Langer, & Mullis, 1985; Linn, Baker, & Dunbar, 1991; Zwick, Donoghue, & Grimes, 1993). Given the high-stakes nature of many of these large-scale performance assessments (Welch & Miller, 1995), researchers have been concerned not only with various factors that may influence performance assessments, but with the unintended consequences of these tests (Bond, 1995). The major factors influencing the assessment of
writing ability may be related to: a) the assessment process (rater bias or severity, rating method); b) student characteristics like "gender, age, ethnicity, race, social class, or opportunity to learn" (Engelhard, 1992, p. 75); c) factors linked to the writing task itself (prompt or domain) or d) a combination of the above factors.

**Raters**

Raters have been shown to influence the assessed quality of student writing (Engelhard, 1994, 1996; Lunz, Wright, & Linacre, 1990; Du & Wright, 1997). Engelhard (1994) examined rater differences in assessing essays. From the 1990 administration of the Eighth Grade Writing Test in Georgia, a sample of 264 randomly selected compositions assessed by 15 raters formed the data for the study. Fifty-one percent of the compositions were written by female students and forty-nine percent by males. Each essay was scored by two operational raters on five domains (content/organization, style, sentence-formation, usage, and mechanics). The ratings of the validity committee were used to anchor the calibrations of the 15 operational raters. Three facets (writing competence, rater severity, and domain difficulty) were utilized to calibrate the raters using the FACETS computer program. Results indicated significant differences between the raters, $\chi^2 (15) = 170.7, p< .01$ with a high reliability of separation index ($R = .87$). Unless adjustments were made, the writing competence of the students judged by severe raters would be underestimated.
Du and Wright (1997), using data from the 1993 direct writing assessment of the Illinois State Goal Programs (IGAP), explored rater, rating scale, and writing task effects and used the many-faceted Rasch model to adjust student measures based on those effects. The data for the study comprised 1734 randomly selected essays by 867 students in grades 6 (27%), 8 (24%), and 10 (49%), and scored by 89 trained raters. Results from the study indicated that there were significant differences in rater severity, and that the scales and topics were further sources of significant differences in student scores. Du and Wright concluded that to ensure objective measurement of student writing ability, rater, rating scale, and task variations had to be identified and student scores had to be adjusted accordingly.

**Domain**

Domains refer to aspects or characteristics of essay quality that are analyzed and separately scored. Such aspects may include, for example, language mechanics, style, sentence formation, spelling, or content/organization. Research on domain influences seems scanty. However, Gabrielson, Gordon, and Engelhard (1995) in their examination of the effects of task choice on the writing quality of 11th grade students used as the dependent variables the following four domains in each essay: a) content and organization, b) style, c) conventions, and d) sentence formation. Female students scored significantly higher than the males on all domains, with the largest gender differences on the
conventions domain, which referred to the appropriate usage of the mechanics of standard American English. White students performed significantly higher than Black students on all four domains, with the largest differences on the conventions domain, followed by sentence formation, the style, and the content and organization domains. These results tallied with earlier research by Engelhard, Gordon, Walker, and Gabrielson (1994).

This study extends the above research, with a focus on the rater by domain interaction effect. The study is also unique because of the data set it utilizes: each one of the twenty raters rated the 366 essays. The purpose of this study is to apply the Rasch model to examine the rater and domain influences on the assessed quality of student writing. The Rasch model is used to estimate the student abilities, the rater severities, the domain difficulty, and the rater/domain interaction effect.

**The Rasch Model**

In measuring the performance of persons on items, the Rasch model (Wright, 1977) has one ability parameter ($\beta_v$) for each person $v$ and one difficulty parameter $\delta_i$ for each item $i$. These parameters are used in the model to determine the probability of person $v$ succeeding on item $i$. That probability is governed by the difference between the person ability and item difficulty parameters ($\beta_v - \delta_i$). Because the probability must remain between zero and one, the difference
(β_v-δ_i) is expressed as an exponent of a base e^{(β_v-δ_i)} (e = 2.718) and the Rasch probability for a right answer is 
\[ \frac{e^{(β_v-δ_i)}}{1 + e^{(β_v-δ_i)}}. \]

According to the Rasch model, the probability of a person’s success on an item depends on the difference between that person’s ability and the difficulty of the item attempted. When that person has more ability than the item demands, β_v is greater than δ_i, their difference is positive, and the person’s ability to answer that item correctly is greater than 0.5. Conversely, when an item requires more ability than a person has, the difference β_v-δ_i will be negative, and the probability of success on the item will be less than 0.5.

To measure person ability on a given latent trait and to calibrate items used in estimating that ability, the following Rasch model is used:

\[ \Pr(x_{vi} \mid β_v, δ_i) = \frac{e^{x_{vi}(β_v-δ_i)}}{1 + e^{(β_v-δ_i)}}. \]

where if person v answers item i correctly it is expressed as \( x_{vi} = 1 \) and if wrongly, \( x_{vi} = 0 \).

The Rasch model uses a mathematical unit of measurement known as a “logit”. A person’s ability in logits is their natural log odds for succeeding on items of the kind used to define the scale origin. The probability \( P \) for succeeding on an item with difficulty \( δ = 0 \) is \( e^{β}/(1+e^{β}) \) from which their
success are $P/(1-P) = e^\beta$, the natural log of which is $\beta$. An item's difficulty in logits is the natural log odds for failure on that item by persons with abilities at the scale origin. The probability $P$ of these persons with abilities at $\beta=0$ of succeeding on an item with difficulty $\delta$ is $e^{\delta}/(1 + e^{\delta})$ from which their odds for failure are $(1 - P)/ P = e^{\delta}$, the natural log of which is $\delta$.

Because reporting of measured person abilities in logits may be cumbersome for teachers or parents, it is possible to transform the logits into other units that can be expressed conveniently in positive integers. A linear transformation can be used to set the location and scale to the desired values.

**Estimating and Analyzing person ability and item difficulty**

The Rasch model provides mathematical procedures to estimate person abilities and item difficulties. The simple manual estimation of parameters is called PROX. However, computer programs like FACETS (Linacre & Wright, 1994) are available to make the necessary analyses and estimates. The analyses provide "fit" statistics to enable an evaluation of how the data fit the Rasch model. Using squared residuals and chi-square statistics, it is possible to flag items or persons that are behaving unexpectedly based on the model. For example, if a person with low ability succeeds on a very
Applying the Rasch model

In this study, the Rasch model was used to explore the relationship between raters, domains of written compositions, and student writing ability.

The data for this application comprised 366 English language compositions written by high school students. The students' demographic characteristics were as follows: 46.7% female and 53.3% male; 77.7% White, 17.3% Black, and 5.0% Other. These essays were scored on four domains by 20 raters randomly selected from a group of 87 operational raters contracted to rate essays as part of a 1993 field test of the Georgia High School Writing Test. The four domains were:

1. Content/Organization: measuring student competence in the development of a controlling idea.
2. Style: measuring student competence in the control of language to establish individuality.
3. Conventions: measuring student competence in the use of the appropriate conventions of standard written English.
4. Sentence formation: measures student competence in the formulation of correct sentences.

In this application, the Rasch model was used to estimate the student abilities, the rater abilities, the domain difficulty, and rater/domain interaction effect. The following research questions were addressed:
1. Are there significant rater differences using the students' domain scores?

2. Are there significant differences in the level of domain difficulty?

3. Are raters equally severe or lenient on all domains?

The major focus of the study was on identifying and describing rater by domain interaction effects.

The FACETS computer program (Linacre & Wright, 1994) was used to run the Rasch model analyses. In this case, three facets were in the model: students, raters, and domains.

Results

Facet Map

Fig. 1 maps the elements of the three facets on a vertical "ruler". Column 1 shows the scale ("Measure") in logit units from -8 (corresponding to lowest student ability, most lenient rater, or easiest domain) to 9 (corresponding to highest student ability, most severe rater, or most difficult domain). Column 2 shows the spread of student writing ability with upper portion being the high ability and the lower portion being the lower ability. The rater facet in column 3 is centered at zero. Raters whose severity level is below zero were more lenient than those above zero, who were more severe. The domain facet in column 4 is also centered at zero, with Domain 2 (Style) being the most difficult, and Domain 4 (sentence formation) being the easiest.

Student Measures
Student ability ranged from -7.45 to 8.06 and the fit statistics showed a high reliability of separation index $R = .99$. There was also significant differences in student writing ability $\chi^2 (N =366)=45754.3$, df = 365, $p < .001$ (see Table 1).

**Domain Measures**

Table 2 shows the Domains measurement report. The style domain was more difficult (0.66 logits), followed by Content/Organization (.32 logits), then Conventions (-0.08 logits), and the easiest was Sentence formation (-0.91 logits). The fit statistics indicate that the fit of the domain difficulty ratings to the model are acceptable (Msq. ranged between .9 to 1.1). However, there were significant differences in the level of difficulty between the domains ($\chi^2 (N = 4) = 2531.5$, df = 3, $p < .001$). Figure 2, based on an enlarged facet map scale, highlights the differences in the levels of domain difficulty.

**Rater Severity**

The severity of the raters ranged from -0.57 logits for rater 15 who was most lenient to .74 logits for rater 9 who was most severe. Overall, rater severity falls between +1.00 and -1.00 logits which may indicate a healthy range of severity. However, there were still overall significant differences between raters ($\chi^2 (N = 20) = 850.1$, d.f. = 19, $p< .001$ (Table 3). The outfit Msq. ranges from 0.7 to 1.1, indicating that the fit of the ratings to the model was
acceptable. Figure 3, which is based on an enlargement of the facet map scale, clearly highlights the rater differences.

**Rater/Domain Interaction**

Z-scores above 2.0 or below -2.0 would indicate an interaction effect. According to the Bias/Interaction report, there were several raters who seemed to be too lenient or too severe on certain domains. Z-scores ranged from -7.8 to 7.2. For example, rater 15 with an expected score of 1052.0 had an observed score of 983 on domain 3, translating into a Z-score of 7.2. There was an overall statistically significant rater by domain interaction effect $\chi^2 (N = 80) = 901.8$, df. = 80, $p < 0.001$ (Table 4).

Figures 4 to 7 graphically illustrate the rater by domain interaction effect as indicated by the different Z-scores on the four domains respectively: Content/Organization (#1 in the output), Style (#2), Conventions (#3), and Sentence Formation (#4). For example, in Fig. 4 regarding the Content/Organization domain, it can clearly be seen that raters #5, #13, and #18 have much lower (interaction) Z-Scores than raters #17, #19, and #21. Fig. 8 which illustrates the rater by domain interaction effect for all domains, indicates some rater patterns. For example, rater #4’s Z-scores center around 0, ranging from -0.90 on Content/Organization to 0.7 on Style compared to rater #13 whose Z-scores range from -6.2 (too lenient) on the Conventions domain to 6.7 (too severe) on the Style domain.
Discussion and Conclusion

The finding of significant differences between raters means that for the individual student, it matters who rates their essay. This should not be the case. Other studies have reported this seemingly persistent rater effect (Du & Wright, 1997; Engelhard, 1994) and it prevails in spite of intensive, expensive, and rigorous rater training. More effort is needed to try to minimize these differences. Perhaps, qualitative surveys exploring rater insights about the rating process and dynamics may provide some meaningful explanations of the rater differences. Given the increasing nationwide acceptance of large-scale performance assessments, many of which are high-stakes (Welch & Miller, 1995), the need for continuous exploration of rater differences cannot be overemphasized.

The finding of an overall significant rater by domain interaction effect is interesting and warrants further investigation. Although the outfit mean squares for the rater by domain interaction (ranging from 0.7 to 1.4) is within acceptable limits (Engelhard, 1996; Lunz, Wright, Linacre, 1990), Smith, Schmacker, and Bush (1998, p. 77) suggest that "mean squares are more sensitive to sample size and reliance on a single critical value for the mean square can result in an under detection of misfit." Based on the Z-scores for the interaction effects, raters seem to be too lenient or too severe depending on the domain being scored, as Figures 4 to 8 illustrate. For the individual student, therefore, this
creates another obstacle to a fair score: the student's score should not depend on rater "bias" towards a certain domain. The interaction also complicates the situation because it makes it inappropriate to adjust for overall rater severity. Therefore, the rater by domain interaction effect points to another area of concern for administrators and test developers.

By focusing on the individual person, rater, and domain parameters, the Rasch model seems to be a very useful tool for test developers and administrators who can use the results not only to calibrate raters and domains and measure persons, but also to identify misfitting raters and domains. Such information is indispensable if the goal is objective measurement.
References


Table 1. Raters Measurement Report (arranged by mN).

<table>
<thead>
<tr>
<th>Obsvd Score</th>
<th>Obsvd Count</th>
<th>Average Avrge</th>
<th>Measure Model 1</th>
<th>Infit MnSq Std</th>
<th>Outfit MnSq Std</th>
<th>Nu raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>3685</td>
<td>1464</td>
<td>2.5 2.3</td>
<td>0.74 0.05</td>
<td>0.8 -6 0.7</td>
<td>6 6</td>
<td>9 9</td>
</tr>
<tr>
<td>3783</td>
<td>1464</td>
<td>2.6 2.3</td>
<td>0.48 0.05</td>
<td>1.1 2 1.3</td>
<td>5 18</td>
<td></td>
</tr>
<tr>
<td>3809</td>
<td>1464</td>
<td>2.6 2.4</td>
<td>0.41 0.05</td>
<td>1.0 0 1.1</td>
<td>1 16</td>
<td></td>
</tr>
<tr>
<td>3865</td>
<td>1464</td>
<td>2.6 2.4</td>
<td>0.26 0.05</td>
<td>0.9 -1 0.9</td>
<td>-2 8</td>
<td></td>
</tr>
<tr>
<td>3897</td>
<td>1464</td>
<td>2.7 2.4</td>
<td>0.18 0.05</td>
<td>1.0 0 1.1</td>
<td>1 14</td>
<td></td>
</tr>
<tr>
<td>3901</td>
<td>1464</td>
<td>2.7 2.4</td>
<td>0.17 0.05</td>
<td>1.0 0 1.0</td>
<td>0 13</td>
<td></td>
</tr>
<tr>
<td>3906</td>
<td>1464</td>
<td>2.7 2.4</td>
<td>0.15 0.05</td>
<td>1.0 0 1.0</td>
<td>0 20</td>
<td></td>
</tr>
<tr>
<td>3922</td>
<td>1464</td>
<td>2.7 2.4</td>
<td>0.11 0.05</td>
<td>0.8 -5 0.8</td>
<td>-4 17</td>
<td></td>
</tr>
<tr>
<td>3931</td>
<td>1464</td>
<td>2.7 2.5</td>
<td>0.09 0.05</td>
<td>0.8 -5 0.8</td>
<td>-4 4</td>
<td></td>
</tr>
<tr>
<td>3940</td>
<td>1464</td>
<td>2.7 2.5</td>
<td>0.06 0.05</td>
<td>1.3 6 1.2</td>
<td>4 19</td>
<td></td>
</tr>
<tr>
<td>3973</td>
<td>1464</td>
<td>2.7 2.5</td>
<td>-0.02 0.05</td>
<td>0.8 -5 0.8</td>
<td>-4 2</td>
<td></td>
</tr>
<tr>
<td>3977</td>
<td>1464</td>
<td>2.7 2.5</td>
<td>-0.03 0.05</td>
<td>1.0 0 1.0</td>
<td>0 3</td>
<td></td>
</tr>
<tr>
<td>3976</td>
<td>1464</td>
<td>2.7 2.5</td>
<td>-0.03 0.05</td>
<td>1.3 6 1.2</td>
<td>4 21</td>
<td></td>
</tr>
<tr>
<td>3999</td>
<td>1464</td>
<td>2.7 2.5</td>
<td>-0.09 0.05</td>
<td>0.8 -5 0.8</td>
<td>-4 11</td>
<td></td>
</tr>
<tr>
<td>4031</td>
<td>1464</td>
<td>2.8 2.5</td>
<td>-0.18 0.05</td>
<td>0.9 -3 0.9</td>
<td>-2 12</td>
<td></td>
</tr>
<tr>
<td>4064</td>
<td>1464</td>
<td>2.8 2.6</td>
<td>-0.27 0.05</td>
<td>1.2 4 1.2</td>
<td>3 5</td>
<td></td>
</tr>
<tr>
<td>4133</td>
<td>1464</td>
<td>2.8 2.6</td>
<td>-0.45 0.05</td>
<td>1.0 0 1.0</td>
<td>0 6</td>
<td></td>
</tr>
<tr>
<td>4151</td>
<td>1464</td>
<td>2.8 2.6</td>
<td>-0.50 0.05</td>
<td>1.2 5 1.2</td>
<td>4 7</td>
<td></td>
</tr>
<tr>
<td>4152</td>
<td>1464</td>
<td>2.8 2.6</td>
<td>-0.51 0.05</td>
<td>1.1 1 1.1</td>
<td>1 10</td>
<td></td>
</tr>
<tr>
<td>4175</td>
<td>1464</td>
<td>2.9 2.7</td>
<td>-0.57 0.05</td>
<td>1.0 0 1.1</td>
<td>1 15</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obsvd Score</th>
<th>Obsvd Count</th>
<th>Average Avrge</th>
<th>Measure Model 1</th>
<th>Infit MnSq Std</th>
<th>Outfit MnSq Std</th>
<th>Nu raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>3963.5</td>
<td>1464.0</td>
<td>2.7 2.5</td>
<td>0.00 0.05</td>
<td>1.0 -0.3 1.0</td>
<td>-0.0</td>
<td>Mean (Count: 20)</td>
</tr>
<tr>
<td>126.3</td>
<td>0.0</td>
<td>0.1 0.1</td>
<td>0.34 0.00</td>
<td>0.1 4.1 0.2</td>
<td>3.4</td>
<td>S.D.</td>
</tr>
</tbody>
</table>

RMSE 0.05  Adj S.D. 0.33  Separation 6.46  Reliability 0.98
Fixed (all same) chi-square: 850.1  d.f.: 19  significance: .00
Random (normal) chi-square: 19.0  d.f.: 18  significance: .39
Table 2. Domains Measurement Report (arranged by mN).

<table>
<thead>
<tr>
<th>Obsvd</th>
<th>Obsvd</th>
<th>Obsvd</th>
<th>Fair</th>
<th>Measure Model</th>
<th>Infit</th>
<th>Outfit</th>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Avrge</th>
<th>Logit Error</th>
<th>MnSq Std</th>
<th>MnSq Std</th>
<th>N domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>18569</td>
<td>7320</td>
<td>2.5</td>
<td>2.3</td>
<td>0.66</td>
<td>0.02</td>
<td>1.0</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2.2</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>19217</td>
<td>7320</td>
<td>2.6</td>
<td>2.4</td>
<td>0.32</td>
<td>0.02</td>
<td>1.1</td>
<td>3</td>
<td>1.1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>19975</td>
<td>7320</td>
<td>2.7</td>
<td>2.5</td>
<td>-0.08</td>
<td>0.02</td>
<td>0.9</td>
<td>-4</td>
<td>0.9</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>33</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>21509</td>
<td>7320</td>
<td>2.9</td>
<td>2.8</td>
<td>-0.91</td>
<td>0.02</td>
<td>1.0</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>44</td>
<td>4</td>
<td>44</td>
</tr>
<tr>
<td>19817.5</td>
<td>7320.0</td>
<td>2.7</td>
<td>2.5</td>
<td>0.00</td>
<td>0.02</td>
<td>1.0</td>
<td>-0.3</td>
<td>1.0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>Mean (Count: 4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1096.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.59</td>
<td>0.00</td>
<td>0.0</td>
<td>2.8</td>
<td>0.0</td>
<td>2.3</td>
<td>0</td>
<td>2.3</td>
<td>S.D.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RMSE 0.02  Adj S.D.  0.59  Separation 25.35  Reliability 1.00
Fixed (all same) chi-square: 2531.5  d.f.: 3  significance: .00
Random (normal) chi-square: 3.0  d.f.: 2  significance: .22
Table 3. Student Summary Report

<table>
<thead>
<tr>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Avrge</th>
<th>Logit Error</th>
<th>MnSq Std</th>
<th>MnSq Std</th>
<th>Num students</th>
</tr>
</thead>
<tbody>
<tr>
<td>216.6</td>
<td>80.0</td>
<td>2.7</td>
<td>2.7</td>
<td>0.79</td>
<td>0.23</td>
<td>1.0</td>
<td>-0.1</td>
</tr>
<tr>
<td>59.3</td>
<td>0.0</td>
<td>0.7</td>
<td>0.7</td>
<td>2.84</td>
<td>0.07</td>
<td>0.2</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Mean (Count: 366)

RMSE 0.24  Adj S.D. 2.83  Separation 11.58  Reliability 0.99
Fixed (all same) chi-square: 45754.3  d.f.: 365  significance: .00
Random (normal) chi-square: 362.5  d.f.: 364  significance: .51
Table 4. Rater by Domain Interaction Summary Report

<table>
<thead>
<tr>
<th>Score</th>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Logit Error Z-Score</th>
<th>MnSq</th>
<th>MnSq</th>
<th>Sq Nu rat logit N do logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>990.9</td>
<td>990.9</td>
<td>366.0</td>
<td>0.00</td>
<td>-0.00</td>
<td>0.10</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>71.1</td>
<td>63.2</td>
<td>0.0</td>
<td>0.09</td>
<td>0.35</td>
<td>0.00</td>
<td>3.4</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Fixed (all = 0) chi-square: 901.8 d.f.: 80 significance: .00
Figure 1. FACET Map for Students, Raters, and Domains.

<table>
<thead>
<tr>
<th>Measure</th>
<th>High Ability</th>
<th>Severe</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>+9</td>
<td>+</td>
<td>+</td>
<td>+(4)</td>
</tr>
<tr>
<td>+8</td>
<td>.</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+7</td>
<td>.</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+6</td>
<td>.</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+5</td>
<td>*</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+4</td>
<td>****</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+3</td>
<td>***********</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+2</td>
<td>**********</td>
<td>+</td>
<td>+3</td>
</tr>
<tr>
<td>+1</td>
<td>*********</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>****</td>
<td>+18</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>***</td>
<td>+12</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>****</td>
<td>+11</td>
<td>+</td>
</tr>
<tr>
<td>-3</td>
<td>***</td>
<td>+13</td>
<td>+</td>
</tr>
<tr>
<td>-4</td>
<td>*</td>
<td>+14</td>
<td>+</td>
</tr>
<tr>
<td>-5</td>
<td>*</td>
<td>+17</td>
<td>+</td>
</tr>
<tr>
<td>-6</td>
<td></td>
<td>+19</td>
<td>+</td>
</tr>
<tr>
<td>-7</td>
<td></td>
<td>+20</td>
<td>+</td>
</tr>
<tr>
<td>-8</td>
<td></td>
<td>+21</td>
<td>+</td>
</tr>
<tr>
<td>+5</td>
<td></td>
<td>+22</td>
<td>+(1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>Low Ability</th>
<th>Lenient</th>
<th>Easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>= 4</td>
<td>-raters</td>
<td>-domains</td>
</tr>
</tbody>
</table>
Figure 2
Domain differences

Domains
1 = Content/Organization
2 = Style
3 = Conventions
4 = Sentence Formation
Figure 3

Rater Differences

Rater Severity in Logits vs. Raters
Figure 4
Rater by Domain Interaction for Content/Organization
Figure 5
Rater by Domain Interaction for Style
Figure 6
Rater by Domain Interaction for Conventions
Figure 7
Rater by Domain Interaction for Sentence Formation
Figure 8
Rater by Domain Interaction for all Domains
I. DOCUMENT IDENTIFICATION:

Title: APPLYING THE RASCH MODEL TO EXPLORE RATER INFLUENCES ON THE ASSESSED QUALITY OF STUDENTS' WRITING ABILITY

Author(s): ISMAIL S. GYAGENDA and GEORGE ENGELHARD, Jr.

Corporate Source: EMORY UNIVERSITY

Publication Date: APRIL 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: ISMAIL S. GYAGENDA

Printed Name/Position/Title: ISMAIL S. GYAGENDA

Organization/Address: EMORY UNIVERSITY, DIVISION OF EDUCATION STUDIES, 1784 NORTH BEACON ROAD, ATLANTA, GA 30322

Telephone: (770) 727-1999 FAX: (770) 727-1999

E-Mail Address: igyagen@emory.edu

Date: MAY 11, 1998