DOCUMENT RESUME

ED 476 466 TM 034 937

AUTHOR	Monahan, Michael P.; Schumacker, Randall E.
TITLE	An Analysis of Correctional Education GED Essays.
PUB DATE	2003-02-15
NOTE	25p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, February 13-15, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Adolescents; Correctional Institutions; *Essays; High School Equivalency Programs; *Item Response Theory; Judges; *Prisoners; *Scoring; Statistical Bias; Training
IDENTIFIERS	*General Educational Development Tests; Rasch Model

ABSTRACT

Three judges rated General Educational Development (GED) student essays that had been written by nine incarcerated youths. If judge's ratings were not consistent, students would receive a biased average rating. Traditional classical measurement theory computes an intraclass reliability coefficient to determine if judges' ratings are reliable. In contrast, a latent trait measurement theory approach can determine a student's fair average rating adjusted for any bias in the judges ratings. A comparison of these two approaches with this sample indicated that the latent trait approach yielded more useful information for training judges to be more consistent and less biased in rating GED student essays. An appendix contains the scoring guide, computer screen illustrations, and the Rasch measurement programs. (Contains 18 references.) (Author/SLD)

TM TMTR 0133

An Analysis of Correctional Education GED Essays

Michael P. Monahan, Ph.D.
Texas Youth Commission

&

Randall E. Schumacker, Ph.D.
University of North Texas

A paper presented at the annual meeting
of the Southwest Educational Research Association
San Antonio, Texas
February 15, 2003

An Analysis of Correctional Education GED Essays

Three judges rated GED student essays that were written by incarcerated youth in a maximum-security facility. If judge ratings are not consistent then students will receive a bias average rating. Tradition classical measurement theory computes an intraclass reliability coefficient to determine if the judges' ratings are reliable. In contrast, a latent trait measurement theory can determine a student's fair average rating adjusted for any bias in the judges' ratings. A comparison of these two approaches indicated that the latent trait approach yielded more useful information toward training judges to be more consistent and less biased in rating student GED essays.

An Analysis of Correctional Education GED Essays

There are currently more than 134,000 youth in 3,700 public and private correctional facilities (Sickmund, 2002) with approximately 86% of the population being male (Gallagher, 1999). In addition, the percent of the youth in these facilities who are in special education programs is very high. Wolford (2000) found that out of the 20 states in his study, an average of 40% of the populations had Individual Education Plans. The rates varied from as low as 12 % or as high as 70 %. An Individual Education Plan includes curriculum and modification instructions that direct the learning of special education students.

Educational skills are critical to juvenile offenders to reduce recidivism rates. Susswein (2000) reports that incarcerated youth were 37% less likely to recidivate if taught to read. Moreover, the U.S. Department of Justice (2003) reported that 77% of state prison inmates who did not complete high school or a Test of General Education Development (GED) were recidivists. Correctional educators are therefore very interested in GED completion by incarcerated youth before their release from the facility. States mandate that incarcerated youth receive a basic education, so it becomes essential that basic education courses be offered. However, with such a high incidence of special education student populations and a low high school graduation rate, one of the ways these youth can complete their basic education is through a GED program.

The Test of General Educational Development (GED Tests) has been in existence since 1942. Over the past 60 years the test has evolved four times. The original GED test was released in 1942. This test was developed to reflect the industrial era and to help World War II veterans complete their education and return to work. This first edition of the GED series continued until the 1978 series was published.

The second edition of the GED was introduced in 1978. This test was developed to reflect the end of the industrial age and a new way of thinking about educational needs. The GED 1978 series reflected this by moving away from just the recollection of facts and towards the application of knowledge within the tests. The second edition was continued until 1988.

The third edition of the GED was introduced in 1988. The GED Testing service updated the series with input from professionals from all areas of adult education. The input from these professionals resulted in five recommended changes. On the top of the list of the five recommendations was the addition of a writing sample or essay. This writing sample was rated on a five-point scale from zero to five. However, the candidate could receive a zero on the essay section, but still pass the language arts test, if the candidate scored high enough on the multiple choice section.

The fourth edition of the GED was introduced in 2002. A panel of experts in adult education was brought together to recommend further changes. From the panels deliberations came four enhancements to the GED test. Once again the GED language arts test would be revised. Although the language arts test would continue to have a multiple-choice section and an essay section, the essay section is now rated on a four-point scale from one to four. The candidate would now have to pass the multiple-choice section and receive a score of two or higher to pass the essay section, or take both sections again (ACE, 2003a; ACE, 2003b; ACE, 2003c).

METHODOLOGY

Three distinctly different judges rated the GED Essays of nine (9) adjudicated youth in a Texas Correctional Facility. The many-faceted Rasch measurement software was used to analyze the three different ratings of judges on the 9 different adjudicated youth. The many-faceted Rasch measurement software yields an index of inter-rater reliability, judge bias, and corrects the students GED essay rating score for any judge bias. The three distinctly different judges, an experienced GED teacher, a Special Education teacher, and a new GED teacher, may give different GED essay ratings on the students. This forms the basis for the analysis and research question. Do the three different teachers rate the GED student essays differently? and Is judge bias present?

Subjects

Three different judges rated 9 adjudicated youth on their GED essays, which yielded 27 GED essay ratings on a scale of 1 (inadequate) to 4 (effective). The students were males, between the ages of 15 and 20, learning disabled, and were incarcerated in a maximum-security facility. Each student was given two twenty minutes sessions on how to outline and write a five-paragraph essay. The students were given 45 minutes to write the GED essay on topics that were provided to them.

Design

The dependent variable was the GED essay ratings using the Official GED essay-scoring rubric (see appendix). The independent variable was the three judges. A determination of how well students do on the GED essay is a function of whether the judges consistently rate the students the same. If a judge is too harsh, too lenient, or too in-experienced, then ratings they give a particular student may differ. A crossed facet design was utilized, rather than a nested or mixed facet design (Schumacker, 1999).

Analysis

A traditional classical theory intra-class correlation index of inter-rater reliability was computed using a mixed model (judges – fixed facet; students – random facet). A latent trait index of reliability reported in the Rasch Facet analysis software program was also computed. In addition, student observed average ratings, judge bias scores, and a corrected student logit score (fair average score adjusted for judge bias) was reported. A comparison of the two approaches is therefore possible (Lunz & Schumacker, 1997).

RESULTS

The GED essay ratings of nine (9) students by three (3) judges are in Table 1. The three judges differed in their experience with GED essay ratings, which formed the basis for the experimental design (Winer, 1971). One judge, Judge A, was very experienced, while Judge B was new and Judge C was new, but from the field of special education. The three judges did not use the Official GED Scale (1 to 4) correctly. The first and second judge used 1.5 and 2.5 values believing that was okay. This indicated an immediate need to train the judges on the necessity of using the Official GED Essay scoring rubric correctly. Consequently, the five misapplied ratings in Table 1 were rounded up and analyzed.

Table 1. GED Essay Ratings

	Judge A	Judge B	Judge C
Student	(Experienced)	(New)	(Special Education)
1	3	2	2
2	1	2	1
3	1.5	2.5	3
4	1.5	2	3
5	1	2	1
6	1	2	1
7	1	2.5	2
8	1	3	1
9	1.5	2	1

Intra-Class Correlation (Classical Theory)

Shrout and Fleiss (1979) indicated 6 different types of intraclass reliability coefficients. These 6 different intraclass reliability coefficients are based on 3 types of models and two versions of the model, i.e., one based on a single individual rating and the other based on the average of all the ratings (McGraw & Wong, 1996). The first model reflects judges who are randomly selected from a population of judges (random) and a specified number of students (fixed), i.e., differences in the individual judge ratings from the average rating of the judge for each individual student. The second model reflects judges who are randomly selected from a population of judges (random) and students who are randomly selected from a large pool of students (random). The third model reflects a specified number of judges (fixed) who rate a randomly selected number of students from a larger pool of students (random). This study involved using the third method for computing the intraclass reliability coefficient because the three judges indicated a fixed facet that rated a differing number of students that represented a random facet. In G-theory, this is referred to as a mixed design.

From a two-way repeated measures analysis of variance, the mean squares can be used to calculate this desired intraclass correlation coefficient when judges are fixed and subjects are random (mixed model). This approach to calculating intraclass reliability coefficients is referred to as G-Theory (Crocker & Algina, 1986; Shavelson & Webb, 1991; Winer, 1971). The analysis of variance summary table was created using the general linear model repeated measures analysis in the *Statistical Package for the Social Sciences*, version 10.05 (see dialog boxes in the appendix). The analysis of variance summary table results that indicate the partitioning of the variance is in Table 2. Variance is partitioned in two facets: student and judge. The mean square error indicates relative measurement error, while the mean square error plus the mean square judge effect indicates absolute measurement error.

The two different types of reliability estimates indicate differences in score interpretation (Shavelson & Webb, 1991). Relative interpretations pertain to the rank ordering of students from poor GED essays to good GED essays. Absolute interpretations refer to the students' performance in writing the GED essay without regard to how well the other students performed. An absolute interpretation is relevant because a student must be rated a minimum of a 2.0 by the judges to pass the GED essay portion of the test. The two different types of decisions impact our definition of measurement error and the magnitude of error variance used to compute the two separate reliability coefficients. For relative decisions, all variance related to the judges interaction with the students GED essay is used. The interaction between judges and students GED essays clearly influences the relative standing of the students' performance. For absolute decisions, all variance that reflects differences in the relative ordering of students plus differences in judges across the students contribute to measurement error. Absolute decision measurement error therefore includes all of the interaction variance plus the variance from the judges to indicate how stringent the judges' ratings.

Table 2. G-Theory Analysis of Variance Summary Table

Source	SS	df	MS
Student (i)	4.35	8	.544
Judges (k)	3.24	2	1.62
Error	6.09	16	.381
Total	13.68	26	

The relative and absolute ICC reliability coefficients are computed and reported for both a single judge estimate and an average of the three judge estimates for both relative and absolute score interpretation in Table 3. The intraclass reliability coefficients reflect a mixed model because the judges define a fixed facet while the students define a random facet. In other words, the same three judges will be rating the GED essays of other students who become eligible for the GED classes. The generalizability formula for calculating the *relative* single judge ICC reliability coefficient is:

$$\hat{p}^2_{(rel:1)} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{rel}} = \frac{(MS_p - MS_e)k}{(MS_p - MS_e / k) + MS_e} = \frac{(.544 - .381)/3}{.054 + .381} = \frac{.054}{.435} = .125$$

The generalizability formula for calculating the *absolute* single judge ICC reliability coefficient is:

$$\hat{p}^2_{(abs:1)} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{abs}} = \frac{(MS_p - MS_e)k}{(MS_p - MS_e / k) + [MS_e / k + (MS_k - MS_e / k)]} = \frac{(.544 - .381)/3}{.054 + .127 + .413} = \frac{.054}{.594} = .090$$

The generalizability formula for calculating the *relative* average of three judges ICC rater reliability coefficient is:

$$\hat{p}^2_{(rel:3)} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{rel} / k} = \frac{(MS_p - MS_e) / k}{(MS_p - MS_e) / k + (MS_e / k)} = \frac{(.544 - .381)/3}{.054 + .127} = \frac{.054}{.181} = .298$$

The generalizability formula for calculating the *absolute* average of three judges ICC rater reliability coefficient is:

$$\hat{p}^2_{(abs:3)} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{abs}^2} = \frac{(MS_p - MS_e)/k}{(MS_p - MS_e)/k + [(MS_e)/k + MS_k - MS_e/n]} = \frac{(.544 - .381)/3}{.054 + .127 + .045} = \frac{.054}{.226} = .238$$

Table 3 indicates that the intraclass reliability coefficients, consistency (relative) and absolute agreement (absolute), that were computed in the *Statistical Package for the Social Sciences*, version 10.05 (see dialog boxes in appendix). The intraclass reliability coefficients are well below an acceptable value of .90 for classical theory inter-rater reliability. The mean square error for the judges indicates that the judges were different in their ratings of students. This is reflected in the low intraclass reliability coefficients. The classical true score approach however doesn't indicate the amount of bias inherent in the judges' ratings, nor adjusts the student GED essay rating for any bias.

Table 3. Intraclass Reliability Coefficients – Mixed Model

Judge Reliability	Relative	Absolute
Single Judge	.125	.090
(95% Confidence Interval)	(-.22, .62)	(-.15, .53)
Average Judge	.298	.238
(95% Confidence Interval)	(-1.2, .83)	(-.65, .78)

Rasch Measurement (Latent Trait Theory)

Latent trait theory permits an analysis of ratings by judges that indicates inter-rater reliability, the amount of judge bias, and produces a "fair score" that has been adjusted for bias (Linacre, 1994). Rasch measurement produces these objective measures in rating scale analysis by using the Facets software program (Linacre, 1994; Wright & Masters, 1982). The many-facet Rasch model implies that student measures are obtained under measurement conditions involving different judges and tasks. A complete set of measures involving all judges rating all students on one or more tasks can be obtained, or different judges can rate different persons on one or more tasks (Allen & Schumacker, 1998). The many-facet Rasch analysis model is: log $(P_{nikj} / P_{nik(j-1)}) = B_n - D_i - C_k - F_j$, where:

P_{nijk} = probability of student n on essay i by judge k being given a rating of j.

$P_{nij(k-1)}$ = probability of student n on task i by judge k being given a rating of j – 1.

B_n = ability of student n.

D_i = difficulty of essay i.

C_k = severity of judge k.

F_j = difficulty of threshold across rating scale categories from j – 1 to j, e.g., 1 to 4.

Each facet, i.e., judge and essay, is assumed to be independent from the other facets. The facets combine to give the expected probability for a student's rating by a particular judge. The facets are comparable on the same linear scale, i.e., logits scale. A judges' bias is indicated for each student by the interaction between the judge's rating of the student and the expected probability rating for the student (observed minus expected average).

The Facets program was specifically created to analyze ratings on persons by judges (Linacre, 1994). The analysis of ratings in Rasch measurement involves several analysis steps that are necessary to create programs (FACFORM) and run data files (FACETS) programs (see appendix for the necessary files). The steps taken to create the files are described next. First, enter raw data into an ASCII file, i.e., school.asc, noting the column locations. Next, write a FACFORM program (school.key) to read the ASCII data file (school.asc) and output a comma separated data file (school.fac) and a Facets program specification file (school.spe). Once this has been accomplished, the FACFORM software program is run (facform.exe) with the FACFORM file (school.key), i.e., c:\facform school.key. If the FACFORM program reads the ASCII data correctly it should create the comma separated file (school.fac) and facets program specification file (school.spe). Check the "school.fac" file to verify that the ASCII data was read correctly. The data should indicate the student id number followed by a range for the number of judges (1-3), and then the ratings of the three judges. It is possible to edit the Facets program specification file (school.spe) and add labels to the facets and label the rating scale, i.e., place variable labels in the program. Now you are ready to execute the FACETS program (facets.exe) with the Facets program specification file (school.spe), i.e., c:\facets school.spe. If the FACETS program runs correctly it should read the comma separated data file (school.fac) and create a computer output file (school.out) and a score file (school). The output file will contain information on the inter-rater reliability and give a "fair" average rating for each student. The score file will indicate the judge bias. The many-faceted Rasch software is available at http://www.rasch.org/.

The three judges were asked to use an Official GED scoring rubric to rate the students GED essays from 1 = inadequate, 2 = marginal, 3 = adequate, and 4 = effective (see appendix). Previous GED requirements only involved passing the multiple choice part of the test, however, current standards require students to pass both the multiple choice section and the essay written section. Consequently, inter-rater reliability is critical to knowing that the judges are similar in their ratings of individuals, i.e., in agreement.

Table 4 indicates Rasch measurement output for the nine students. The first column presents the sum of the ratings for each student, the second column indicates 3 ratings, and the third column indicates the average rating. The fourth column is of particular interest because it indicates a "fair" average or the average rating of the judges corrected for bias. Students 1,3, and 4 had observed averages and fair averages that differed slightly, i.e., 2.3 vs. 2.4; 2.7 vs. 2.8; and 2.3 vs. 2.4. Chi-square = 18.6, df = 8, and p=.02 indicates that the student ratings are statistically significantly different (Schumacker & Lunz, 1997). The Rasch reliability = .56 indicates that judges produced a moderate separation in the student ratings. The negative logit measures indicate those students who were rated low on the GED essays overall by the judges, e.g., student 2 logit = -2.66. The positive logit measures indicate those students who were rated high on the GED essays overall by the judges, e.g., student 3 logit = 2.68.

Table 4. Student Ratings and Calibration (n = 9)

```
---------------------------------------------------------------------------
 Obsvd   Obsvd  Obsvd   Fair | Logit Model | Infit      Outfit   |
 Score   Count Average Avrge |Measure  S.E.|MnSq Std   MnSq Std  | student
---------------------------------------------------------------------------
     7       3     2.3   2.4 |   1.23  1.16 | 2.3   1    2.7   1  |    1
     4       3     1.3   1.3 |  -2.66  1.40 | 0.2  -1    0.1  -1  |    2
     8       3     2.7   2.8 |   2.68  1.29 | 0.6   0    0.5   0  |    3
     7       3     2.3   2.4 |   1.23  1.16 | 1.9   0    2.4   1  |    4
     4       3     1.3   1.3 |  -2.66  1.40 | 0.2  -1    0.1  -1  |    5
     4       3     1.3   1.3 |  -2.66  1.40 | 0.2  -1    0.1  -1  |    6
     6       3     2.0   2.0 |   0.01  1.07 | 0.7   0    0.7   0  |    7
     5       3     1.7   1.7 |  -1.13  1.11 | 1.1   0    1.0   0  |    8
     5       3     1.7   1.7 |  -1.13  1.11 | 0.9   0    1.0   0  |    9
---------------------------------------------------------------------------
Mean 5.6     3.0     1.9   1.8 |  -0.56  1.23 | 0.9  -0.4  1.0  -0.2|
S.D. 1.4     0.0     0.5   0.5 |   1.86  0.13 | 0.7   0.8  0.9   0.8|
---------------------------------------------------------------------------
RMSE  1.24  Adj S.D.  1.39  Separation  1.12  Reliability 0.56
Fixed (all same) chi-square: 18.6  d.f.: 8  significance: .02
---------------------------------------------------------------------------
```

Table 5 indicates the Rasch measurement output for the three judges. The first column indicates the sum of the judges' ratings across the nine students. The second column indicates that each judge gave 9 ratings. The third column is the observed average rating of each judge. The first and third judge had similar average ratings, i.e., 1.6 and 1.7, while the second judge had a much higher average rating, i.e., 2.3. The second judge was therefore more lenient than the other two judges. This is also reflected in the Logit Measure where the first and third judge had logit values of 1.2 and .71, respectively, compared to a logit of –1.91 for the second judge. The chi-square = 11.6, df = 2, p = .001 indicates that the judges' ratings were statistically significantly different. The Rasch reliability = .74 indicates that the judges ratings were consistently different.

Table 5. Judges ratings and calibrations (n=3)

```
-------------------------------------------------------------------------------
 Obsvd     Obsvd  Obsvd    Fair |Logit    Model | Infit        Outfit   |
 Score     Count Average Avrge |Measure  S.E.  |MnSq Std    MnSq Std   |  judge
-------------------------------------------------------------------------------
   14        9     1.6     1.6  |  1.20  0.72  | 1.1    0     0.9    0  | 1
   21        9     2.3     2.6  | -1.91  0.68  | 1.1    0     1.4    0  | 2
   15        9     1.7     1.8  |  0.71  0.69  | 0.7    0     0.6    0  | 3
-------------------------------------------------------------------------------
Mean 16.7      9.0    1.9    2.0 |   0.00  0.70 | 1.0  -0.1  1.0 -0.2 |
S.D.  3.1      0.0    0.3    0.4 |   1.37  0.02 | 0.2   0.5  0.4  0.6 |
-------------------------------------------------------------------------------
RMSE  0.70  Adj S.D.  1.17  Separation  1.68  Reliability 0.74
Fixed (all same) chi-square: 11.6  d.f.: 2  significance: .001
```

Table 6 indicates whether any of the three judges were biased in their individual ratings of the nine students. The amount of bias in the table for each student indicated that students 1, 3, and 4 had different bias values by the three judges (boldfaced values). This confirms why the observed average and fair average differed for these three students in Table 4. For example, Judge 1 was the most stringent (logit = 1.20), but the observed score for student 1 was different than the expected score, thus the "Bias Measure" indicated a value of – 1.68. The other judges for student 1 also indicated judge bias; consequently the "Fair average" reflects an adjustment to the students score in Table 4.

Table 6. Judge Bias/Interaction

```
---------------------------------------------------------------------------------------
 Obsvd     Exp.  Obsvd  Obs-Exp| Bias    Model         |Infit Outfit|
 Score    Score  Count  Average| Measure S.E.  Z-Score| MnSq   MnSq | st measr ju measr
---------------------------------------------------------------------------------------
    3       2.0    1      0.99|  -1.68  1.87   -0.9 |  0.8    0.8 |  1   1.23 1   1.20
    1       1.1    1     -0.09|   0.00  3.47    0.0 |  0.1    0.1 |  2  -2.66 1   1.20
    2       2.4    1     -0.45|   1.47  1.82    0.8 |  0.0    0.0 |  3   2.68 1   1.20
    2       2.0    1     -0.01|   0.03  1.82    0.0 |  0.0    0.0 |  4   1.23 1   1.20
    1       1.1    1     -0.09|   0.00  3.47    0.0 |  0.1    0.1 |  5  -2.66 1   1.20
    1       1.1    1     -0.09|   0.00  3.47    0.0 |  0.1    0.1 |  6  -2.66 1   1.20
    1       1.6    1     -0.64|   0.64  1.90    0.3 |  0.7    0.7 |  7   0.01 1   1.20
    1       1.3    1     -0.32|   0.00  2.08    0.0 |  0.4    0.4 |  8  -1.13 1   1.20
    2       1.3    1      0.68|  -2.32  1.82   -1.3 |  0.0    0.0 |  9  -1.13 1   1.20
    2       2.8    1     -0.83|   3.12  1.82    1.7 |  0.0    0.0 |  1   1.23 2  -1.91
    2       1.8    1      0.23|  -0.74  1.82   -0.4 |  0.0    0.0 |  2  -2.66 2  -1.91
    3       3.0    1      0.05|   0.00  4.81    0.0 |  0.0    0.0 |  3   2.68 2  -1.91
    2       2.8    1     -0.83|   3.12  1.82    1.7 |  0.0    0.0 |  4   1.23 2  -1.91
    2       1.8    1      0.23|  -0.74  1.82   -0.4 |  0.0    0.0 |  5  -2.66 2  -1.91
    2       1.8    1      0.23|  -0.74  1.82   -0.4 |  0.0    0.0 |  6  -2.66 2  -1.91
    3       2.6    1      0.43|   0.00  1.93    0.0 |  0.7    0.7 |  7   0.01 2  -1.91
    3       2.2    1      0.76|  -1.30  1.97   -0.7 |  0.6    0.6 |  8  -1.13 2  -1.91
    2       2.2    1     -0.24|   0.77  1.82    0.4 |  0.0    0.0 |  9  -1.13 2  -1.91
    2       2.2    1     -0.16|   0.52  1.82    0.3 |  0.0    0.0 |  1   1.23 3   0.71
    1       1.1    1     -0.14|   0.00  2.85    0.0 |  0.2    0.2 |  2  -2.66 3   0.71
    3       2.6    1      0.41|   0.00  1.94    0.0 |  0.6    0.6 |  3   2.68 3   0.71
    3       2.2    1      0.84|  -1.43  1.93   -0.7 |  0.7    0.7 |  4   1.23 3   0.71
    1       1.1    1     -0.14|   0.00  2.85    0.0 |  0.2    0.2 |  5  -2.66 3   0.71
    1       1.1    1     -0.14|   0.00  2.85    0.0 |  0.2    0.2 |  6  -2.66 3   0.71
    2       1.8    1      0.21|  -0.69  1.82   -0.4 |  0.0    0.0 |  7   0.01 3   0.71
    1       1.4    1     -0.45|   0.00  1.90    0.0 |  0.7    0.7 |  8  -1.13 3   0.71
    1       1.4    1     -0.45|   0.00  1.90    0.0 |  0.7    0.7 |  9  -1.13 3   0.71
---------------------------------------------------------------------------------------
Mean 1.9    1.9    1.0  -0.00|   0.00  2.27    0.0 |  0.3    0.3 |
S.D. 0.8    0.6    0.0   0.47|   1.17  0.75    0.6 |  0.3    0.3 |
---------------------------------------------------------------------------------------
```

CONCLUSION

The three judges did not use the Official GED scoring rubric correctly. Two judges used 1.5 and 2.5 rather then the scoring rubric 1 to 4. Consequently, these two judges will need to be trained on how to use the scoring rubric. The ratings for these two judges were rounded up and then analyzed, however, future ratings will have to be checked to make sure they adhere to the Official GED scoring guide.

The classical theory using G-theory analysis of variance components reported intraclass reliability coefficients that did not indicate consistent GED essay ratings of the nine students by the three judges. The intraclass reliability coefficients ranged from .090 (absolute decision) to .125 (relative decision) for a single judge. The intraclass reliability coefficients ranged from .238 (absolute decision) to .298 (relative decision) for the average of the three judges. The intraclass reliability coefficient was based on a mixed model with judges (fixed) and students (random). In addition, an absolute decision was desired because students had to achieve an average rating of 2.0 or higher to pass the GED essay portion of the GED test.

The latent trait theory used Rasch measurement to compute students' essay writing ability, judges' rating ability, and determine the difference between the observed ratings and expected probability ratings, i.e., judge bias. Raw score ratings ranged from 4 to 8 for the three judges with only 4 students receiving an observed average greater than 2.0, i.e., students 1, 3, 4, and 7 (Table 4). The judges' ratings were consistent with a reported reliability coefficient of .74 (Table 5). Two judges, 1 and 3, had total ratings across the nine students of 14 and 15, respectively. The second judge had a higher sum of student ratings indicating that judge gave higher GED essay rating values, i.e., more lenient rater. Judge bias was indicated in Table 6 with students 1, 3, and 4 receiving biased ratings. Consequently, some judge rater bias was present with a few students.

A comparison of classical and latent trait theory is possible given the two separate analyzes. It is obvious that G-theory that reports an intraclass reliability coefficient using the mean squares from a mixed design analysis of variance doesn't capture the differences in student essay writing ability, judges' rating ability, nor the presence of judge bias in rating particular students. Basically, judge inter-rater reliability is present, but not captured by the classical theory intraclass reliability coefficient. More clarity of interpretation is present with Rasch measurement using the Facets program that it is recommended for this type of analysis.

REFERENCES

ACE. (2003a). *History of the GED Tests.* [Online] Available: http://www.acenet.edu/calec/ged/history-A.cfm

ACE. (2003b). *Introduction.* [Online] Available: http://www.acenet.edu/calec/ged/intro-A.cfm

ACE. (2003c). *What's New? Language Arts, Writing.* [Oneline] Available: http://www.acenet.edu/calec/ged/whatsNew_K/aw_/a.cfm

Allen, J.M. & Schumacker, R.E. (1998). Team assessment utilizing a many-facet Rasch Model. *Journal of Outcome Measurement*, *2*(2), 142-158.

Crocker, L., & Algina, J. (1986). *Introduction of Classical & Modern Test Theory.* Belmont, CA: Wadsworth Publishing.

Gallagher, C. A. (1999, March). Juvenile offenders in residential placement, 1997. *Fact Sheet.* Washington, DC: U.S. Dept. of Justice, Office of Juvenile Delinquency Prevention.

Linacre, J.M. (1994). *A User's Guide to Facets*. Chicago, IL: MESA Press.

Lunz, M.E. & Schumacker, R.E. (1997). Scoring and analysis of performance examinations: A comparison of methods and interpretations. *Journal of Outcome Measurement*, *1*(3), 219-238.

McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.

Schumacker, R.E. & Lunz (1997). Interpreting the chi-square statistics reported in the many-faceted Rasch model. *Journal of Outcome Measurement*, *1*(3), 239-257.

Schumacker, R.E. (1999). Many-faceted Rasch Analysis with Crossed, Nested, and Mixed Designs. *Journal of Outcome Measurement*, *3*(4), 323-338.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability Theory: A Primer.* Newbury Park: CA, Sage Publications.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.

Sickmund, M. (2002, March). Juvenile offenders in residential placement 1997-1999. *OJJDP Fact Sheet.* Washington, DC: U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention.

Susswein, G. (2000)*Austin American Statesman Home Page. [Online].* Available: http://www.austin360.com/statesman/edition/wednesday/metro-state-.2html

U.S Department of Justice (2003). *Bureau of Justice Statistics Special Report: Education and Correctional Populations.*[Online] Available: http:www.ojp.usdoj.gov/bjs/

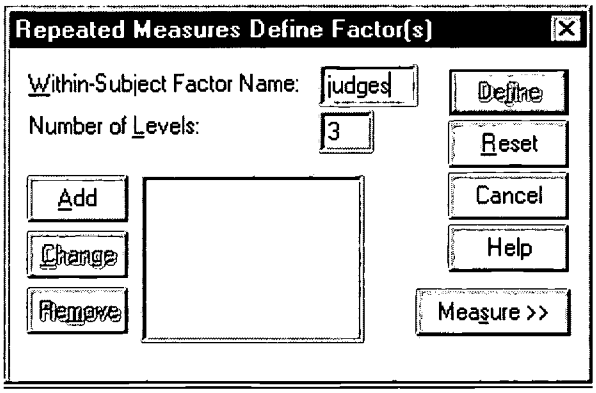
Winer, B.J. (1971). *Statistical Principles in Experimental Desig*n. New York: NY, McGraw-Hill.

Wolford, B. I. (2002). Youth education in the Juvenile Justice System. *Corrections Today,62, 126-130.*

Appendix

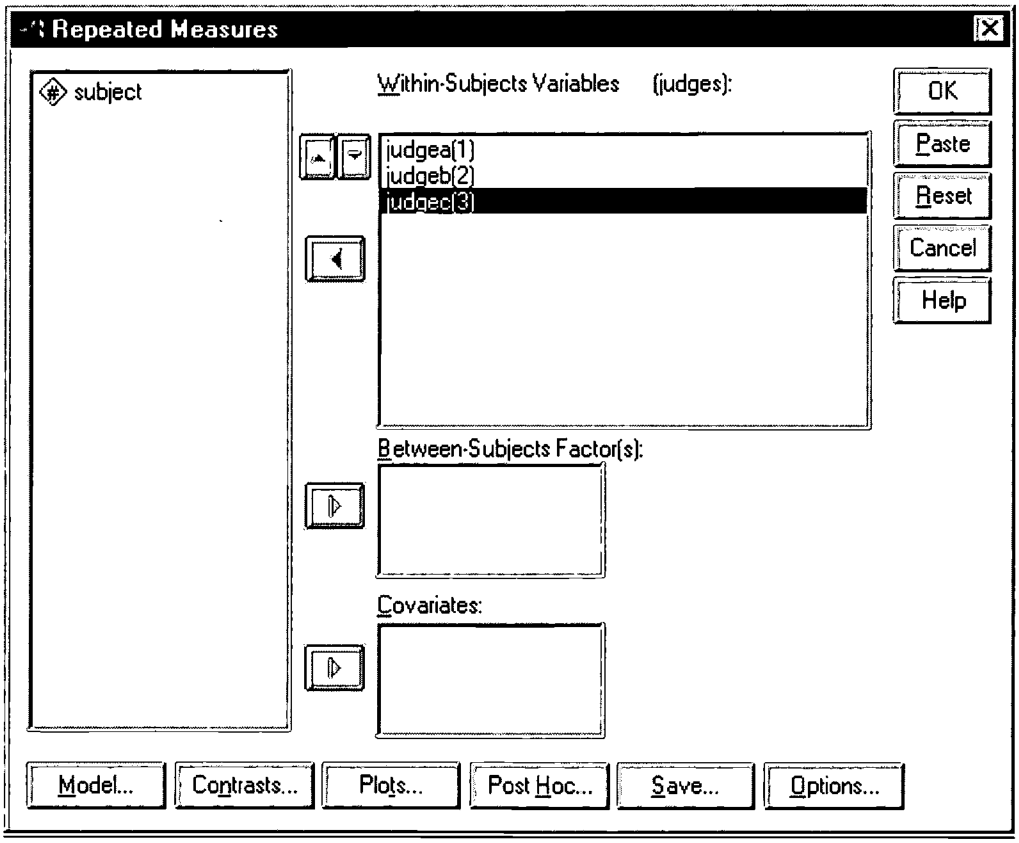
GED Official Essay Scoring Guide—Chart Format

	1	2	3	4
	Inadequate	**Marginal**	**Adequate**	**Effective**
	Reader has difficulty identifying or following the writer's ideas.	**Reader occasionally has difficulty understanding or following the writer's ideas.**	**Reader understands writer's ideas.**	**Reader understands and easily follows the writer's expression of ideas.**
Response to the Prompt	Attempts to address prompt but with little or no success in establishing a focus.	Addresses the prompt, though the focus may shift.	Uses the writing prompt to establish a main idea.	Presents a clearly focused main idea that addresses the prompt.
Organization	Fails to organize ideas.	Shows some evidence of an organizational plan.	Uses an identifiable organizational plan.	Establishes a clear and logical organization.
Development and Details	Demonstrates little or no development; usually lacks details or examples or presents irrelevant information.	Has some development but lacks specific details; may be limited to a listing, repetitions or generalizations.	Has focused but occasionally uneven development; incorporates some specific detail.	Achieves coherent development with specific and relevant details and examples.
Conventions of EAE	Exhibits minimal or no control of sentence structure and the conventions of Edited American English (EAE).	Demonstrates inconsistent control of sentence structure and the conventions of Edited American English (EAE).	Generally controls sentence structure and the conventions of Edited American English (EAE).	Consistently controls sentence structure and the conventions of Edited American English (EAE).
Word Choice	Exhibits weak and/or inappropriate words.	Exhibits a narrow range of word choice, often including inappropriate selections.	Exhibits appropriate word choice.	Exhibits varied and precise word choice.



Within-Subject Factor Name: judges

Number of Levels: 3



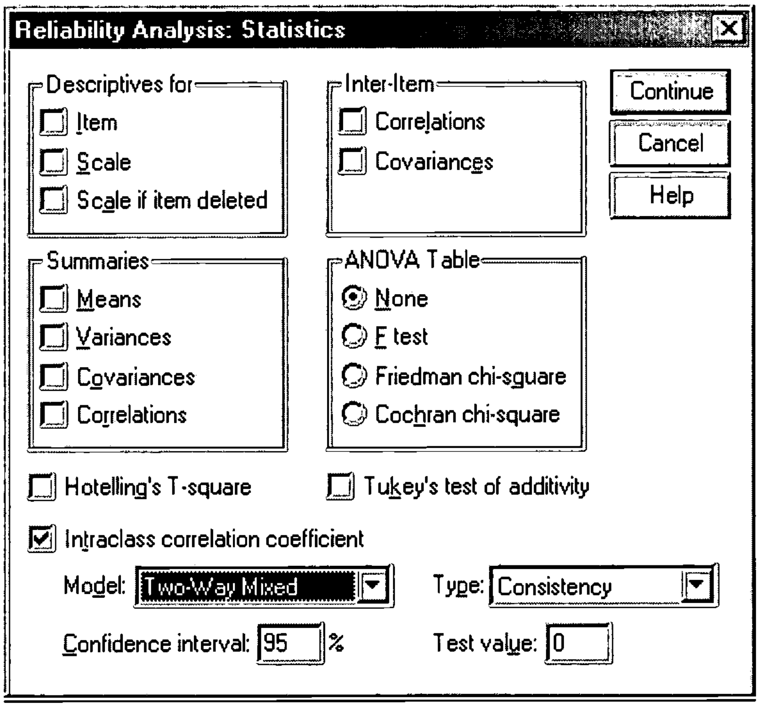
subject

Within-Subjects Variables (judges):

judgea(1)
judgeb(2)
judgec(3)

Between-Subjects Factor(s):

Covariates:



Reliability Analysis: Statistics

Descriptives for
- □ Item
- □ Scale
- □ Scale if item deleted

Inter-Item
- □ Correlations
- □ Covariances

Summaries
- □ Means
- □ Variances
- □ Covariances
- □ Correlations

ANOVA Table
- ● None
- ○ F test
- ○ Friedman chi-square
- ○ Cochran chi-square

□ Hotelling's T-square

□ Tukey's test of additivity

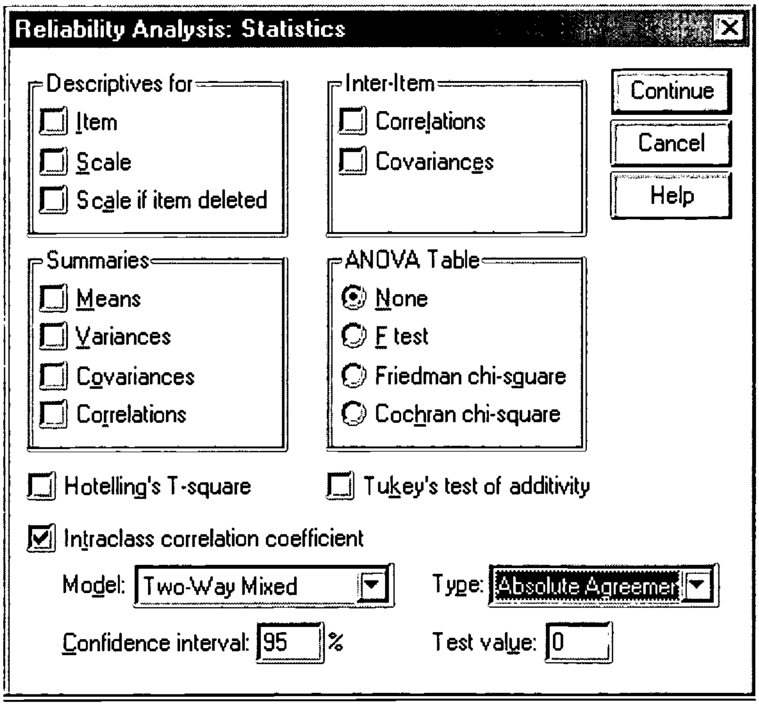
☑ Intraclass correlation coefficient

Model: Two-Way Mixed

Type: Consistency

Confidence interval: 95 %

Test value: 0



Reliability Analysis: Statistics

Descriptives for
- □ Item
- □ Scale
- □ Scale if item deleted

Inter-Item
- □ Correlations
- □ Covariances

Summaries
- □ Means
- □ Variances
- □ Covariances
- □ Correlations

ANOVA Table
- ● None
- ○ F test
- ○ Friedman chi-square
- ○ Cochran chi-square

□ Hotelling's T-square

□ Tukey's test of additivity

☑ Intraclass correlation coefficient

Model: Two-Way Mixed

Type: Absolute Agreemer

Confidence interval: 95 %

Test value: 0

RASCH MEASUREMENT PROGRAMS

Rating Data File (school.asc)
(Student - column 1; Judge - column 3; Rating - columns 5-7)

```
1 1 3.0
1 2 2.0
1 3 2.0
2 1 1.0
2 2 2.0
2 3 1.0
3 1 2.0
3 2 3.0
3 3 3.0
4 1 2.0
4 2 2.0
4 3 3.0
5 1 1.0
5 2 2.0
5 3 1.0
6 1 1.0
6 2 2.0
6 3 1.0
7 1 1.0
7 2 3.0
7 3 2.0
8 1 1.0
8 2 3.0
8 3 1.0
9 1 2.0
9 2 2.0
9 3 1.0
```

FACFORM program to create comma separated data set for Facets
(school.key)

```
; This program reads rating data from file school.asc and outputs
; school.fac and school.spe files
$Input  = school.asc    ; flat file - ascii raw data
$Output = school.fac    ; facform comma separated file
$Spoutput = school.spe  ; specifications file
$Facets=2                ; student and judge
$Flabel=1,"student"
$Flabel=2,"judge"
; Get ratings on the first line
$DO=1
        $Label = 1,$S1W1    ;student id in column 1
        $Label = 2,$S3W1    ;first judge in column 3
        $Rating = $S5W3     ;first rating in column 5-7
; Get ratings from the second line
$Nextline
        $Label = 2,$S3W1    ;second judge in column 3
        $Rating = $S5W3     ;second rating in column 5-7
; Get ratings from the third line
$Nextline
        $Label = 2,$S3W1    ;third judge in column 3
        $Rating = $S5W3     ;third rating in column 5-7
; Repeat for all subjects
$AGAIN
```

Comma Separated Data Set (school.fac)

```
1,1-3,3.0,2.0,2.0
2,1-3,1.0,2.0,1.0
3,1-3,2.0,3.0,3.0
4,1-3,2.0,2.0,3.0
5,1-3,1.0,2.0,1.0
6,1-3,1.0,2.0,1.0
7,1-3,1.0,3.0,2.0
8,1-3,1.0,3.0,1.0
9,1-3,2.0,2.0,1.0
```

Facets program file (school.spe)

```
Title = Facets Analysis of 9 student GED Essays by 3 Judges
Facets = 2
Data file = school.fac ; reads in comma separated rating data
Scorefile=school       ; output file with judge bias values
Output=school.out      ; output file with student and judge calibrations
Models = ?,?,R4
*
; Positive = 1
; Noncenter = 1
Labels =
1,student
1-9
*
2,judge
1-3
*
```

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034937

I. DOCUMENT IDENTIFICATION:

Title: AN ANALYSIS OF CORRECTIONAL EDUCATION GED ESSAYS

Author(s): DR(S) Michael P. Monahan & Randall E. Schumacker

Corporate Source: University of NORTH TEXAS

Publication Date: 2/15/03

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

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.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

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.

Sign here, → please

Signature: Schumacker Michael Monahan

Printed Name/Position/Title: Randall Schumacker Professor

Organization/Address: University of NORTH TEXAS

Telephone: 940 565 3962

FAX: 940 565 2185

E-Mail Address: rschumacker@unt.edu

Date: 2/15/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org