DOCUMENT RESUME

ED 387 507 TM 023 684

AUTHOR Lukhele, Robert; Sireci, Stephen G.
TITLE Using IRT To Combine Multiple-Choice and Free-Response Sections of a Test onto a Common Scale Using A Priori Weights.
PUB DATE 21 Apr 95
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Assessment; *Essay Tests; *Item Response Theory; *Multiple Choice Tests; *Scaling; Scoring; *Test Construction; Test Reliability
IDENTIFIERS A Priori Tests; Calibration; *Free Response Test Items; GED Writing Skills Test; General Educational Development Tests; MULTILOG Computer Program; *Weighting (Statistical)

ABSTRACT

Free-response (FR) item formats, such as essay questions, are popular in educational assessment. The criticisms against FR items are that they are more expensive to score, take up more testing time, provide less content coverage, and are less reliable than multiple-choice (MC) items. For these reasons, FR items are often combined with MC items. Calibrating FR and MC items onto a common scale poses problems for tests constructed using item response theory (IRT), because IRT calibration may reduce the weight of the FR items below the level desired by the test developer. This paper illustrates how MC and FR items can be calibrated onto a common scale using the Multilog computer program and how an a priori weighting scheme can be incorporated within the scale. The relative worth of the MC and FR portions of the test was evaluated with respect to reliability, test information, and passing status using the Writing Skills Test of the Tests of General Educational Development (GED). Data were from a sample of 1,986 GED candidates. Results suggest that to maximize reliability using an IRT calibration model, the weights for the MC and essay sections should not be adjusted a posteriori. (Contains 5 tables, 3 figures, and 23 references.) (Author/SLD)

Using IRT To Combine Multiple-Choice And Free-Response Sections Of A Test Onto A Common Scale Using A Priori Weights

Robert Lukhele[1]

University of California at Santa Barbara

Stephen G. Sireci

GED Testing Service of the American Council on Education

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April 21, 1995

[1]This research was collaborative and the order of the authors is alphabetical. The work of the first author was done while he was an ACE-GEDTS Psychometric Fellow. We are grateful to Andrew Wiley for doing some of the factor analyses, Patricia Jones and David Messersmith for assistance in working with the data files, and Gary Skaggs for his advice on scaling procedures. We are also grateful to Davide Thissen for his guidance on calibrating and scoring the essay and multiple-choice items separately, and to Howard Wainer for suggesting to use Nunnally's formula for estimating the IRT essay weight.

ABSTRACT

Free-response (FR) item formats, such as essay questions, are popular in educational assessment. The criticisms against FR items are that they are more expensive to score, take up more testing time, provide less content coverage, and are less reliable than multiple-choice (MC) items. For these reasons, FR items are often combined with MC items. Calibrating FR and MC items onto a common scale poses problems for tests constructed using IRT, because IRT calibration may reduce the weight of the FR items below the level desired by the test developer. This paper illustrates how MC and FR items can be calibrated onto a common scale using Multilog, and how an a priori weighting scheme can be incorporated within the scale. The relative worth of the MC and FR portions of the test are evaluated with respect to reliability, test information, and passing status.

Introduction

Free-response (FR) item formats, such as essay questions, are becoming increasingly popular in educational measurement. This popularity is due in part to the current emphasis on authentic assessment and a desire to maximize construct validity. The criticisms of FR items are that they are more expensive to score, take up more testing time, provide less content coverage, and are less reliable than multiple-choice (MC) items. For these reasons, FR items are often administered in conjunction with a much larger number of MC items.

It is common knowledge that inferences derived from test scores are limited in their ability to provide information about examinees with respect to the construct measured. It is often recommended that multiple forms of assessment be used to obtain more accurate estimates of examinee proficiency (e.g., Anastasi, 1988). Unfortunately, the restrictions of large-scale testing programs often make multiple test administrations infeasible. Thus high-stakes decisions are often made on the basis of test scores derived from a single battery of tests. In considering this problem, Messick (1989) suggested using several different item formats within a test:

> measurement research in general and construct validation in particular should, as a matter of principle, entail multiple measures of each construct under scrutiny. When a single test score is employed, however, one strategy for triangulating on the referent construct is to incorporate multiple item or task formats in a total score composite. (p. 35)

Given the increased emphasis on more "authentic" assessment (e.g., Linn, 1994), it is likely that tests comprising different item types will become increasingly popular. The Writing Skills Test (WST) of the Tests of General Educational Development (GED Tests) provides an example this type of hybrid test, comprising 50 multiple-choice questions and one holistically-scored essay (scored on a 2-12 scale). Using classical test theory, the MC and essay sections of the test are weighted 64% and 36%, respectively. This weighting was chosen to allow the essay to have a significant impact on total score, while maintaining the reliability of the composite score at or above .87 (Patience & Swartz, 1987). In order to achieve a passing score on the WST, candidates must demonstrate a level of performance greater than or equal to about the 30th percentile of the GED norm group (a nationally-representative sample of U.S. graduating high school seniors). The GED Tests are reported on a 20-80 scale with a mean of 50 and standard deviation of 10. The score associated with passing the WST is typically defined as a GED standard score of 40 or 45[2].

---

[2]The passing scores on the GED Tests apply to the entire battery of tests. The passing scores are set by each state or jurisdiction participating in the GED program. The standard scores of 40 and 45 are often used to evaluate passing scores on a single subtest.

A Description of the GED Writing Skills Test

The Tests of General Educational Development (GED Tests) are a battery of five subject tests designed to measure the academic proficiency of adults seeking a high school-level diploma. The subject tests comprising the GED test battery are: Writing Skills, Social Studies, Science, Interpreting Literature and the Arts, and Mathematics. This report summarizes scaling research conducted on the GED Writing Skills Test (WST). The purpose of this paper is to present and evaluate an IRT-based method for scaling the WST.

The WST comprises 50 multiple-choice (MC) questions and a single essay. The purpose of the WST is to determine the writing proficiency of GED candidates. GED candidates have two hours to complete the WST. They are advised to use 75 minutes for the MC questions and 45 minutes for the essay. The MC portion of the WST "measures the ability to edit sentences within the context of one or more paragraphs of complete, connected discourse" (Auchter, Sireci, & Skaggs, 1993; p. 11). GED candidates are presented with paragraphs and are required to recognize errors that appear in the selection, or select an alternative way of writing one or more sentences from a list of possible "re-writes." The content areas measured by the MC section are sentence structure (35%) usage (35%), and mechanics (30%).

The essay portion of the WST measures a GED candidate's ability to compose a well-written response to an essay topic. Examinees are presented with an expository topic (e.g., state a view or present an opinion) and asked to write a response in essay form. The essays are scored on a 6-point scale by two readers; if the separate scores assigned by two readers differ by more than one point, the essay is scored by a chief reader. The total essay score is the sum of the readers' scores when graded by two readers, and twice the average of the three scores when graded by three readers. Thus, the essay scores range from two to twelve.

In suggesting weights for the MC and essay portions of the tests, the GED Writing Committee recommended that the essay be weighted around 50%. Realizing that such weighting could result in a low level of reliability of the WST composite score, Patience and Swartz (1987) calculated weights that would "weight the essay as much as possible without diminishing the estimated test-retest reliability [of the composite score] below a level professionally acceptable" (p. 5). Based on the results of their analyses, the MC and essay portions of the WST are weighted 64% and 36%, respectively (i.e., weights of .64 and .36 out of 1.0).

Statement of Problem

The procedure used by Patience and Swartz (1987) to derive weights for the MC and essay portions of the WST works well within the classical testing paradigm under which the GED Tests are currently developed. However, the GED Testing Service (GEDTS) is exploring test development procedures based on item response theory

(IRT). Assigning a priori weights to different portions of a test is problematic for tests developed using IRT, because IRT calibration optimally weights all test items to maximize test information. Such weighting may reduce the total impact of free response items, such as essays, below the level desired by the test developer. Thus, if IRT was used to scale the WST, the relative impact of the essay portion of the test could be substantially reduced. The following research questions were addressed by this study:

1. Can the entire WST be calibrated (scaled) using IRT?

2. If the WST can be calibrated via IRT, what would be the "optimal" weight assigned to the essay via an IRT model?

3. Can the current weighting scheme still be used if it is calibrated using an IRT model?

4. What is the relative worth of the essay portion of the WST in terms of reliability and test information?

5. If GEDTS switched to IRT procedures for test development, what would be the impact on passing rates?

Method

Subjects

The data used in the study come from a sample of 1,986 GED candidates who took the entire GED test battery between January and June, 1993. The candidates came from a stratified sample of GED testing centers that participate periodically in GED research projects. The sample was considered to be representative of the total GED candidate population in terms of ability, age, ethnicity/race, and sex. MC item data were also available for 116 additional candidates (no essay scores were available for these candidates). These additional data were used in calibrating the item parameters for the MC items only (total N for MC calibrations was 2,102; total N for essay calibration and for all scoring runs was 1,986).

Procedure

The Multilog 6.1 computer program (Thissen, 1991) was used to calibrate the MC and essay items of the WST. Multilog is an extremely flexible program which can fit both MC and graded-response items (such as the GED essay) in a single calibration run.

MC item calibration. The 50 multiple-choice items were calibrated using a modified version of the three-parameter IRT model (3PL). This modified 3PL included a

lower asymptote (to serve as a pseudo-chance parameter) in the model, but fixed the asymptote at .15[3]. A fixed lower asymptote was used to minimize parameter estimation (given the sample size), and because this model was found to be appropriate on other IRT applications to GED test data. The lower asymptote was fixed at .15 because previous estimates of this asymptote, using a three-parameter IRT model and larger sample sizes of GED candidates, centered around this value. This value is also consistent with the recommendations of Divgi (1984) who suggested that the lower asymptote should be fixed at $\frac{1}{k}-.05$, where k= the number of response alternatives for a MC item. Because all GED MC items have five response alternatives, fixing the lower asymptote at .15 seemed reasonable (1/5=.20, .20-.05=.15). Thus, the modified 3PL model used for the MC items was:

$$p(\theta) = c + [\frac{1-c}{1+\exp[-1.7a(\theta-b)]}]$$

where p (Θ) is the probability of choosing the correct answer as a function of Θ (writing skills proficiency), b is the difficulty level of the item, a is the slope of the item characteristic curve (ICC) at the point $\Theta=b$, and c (the lower asymptote) is .15. Normal priors were used on all MC items.

Essay item calibration. The essay item was calibrated using the graded response model (Samejima, 1969). The graded response model is designed for item formats that have ordered response categories. For a response $x=k$, where $k=1,2,...,m$, m represents the highest score category on the item, and the probability of response k is given by:

$$P(x=k) = \frac{1}{1+\exp[-a(\theta-b_{k-1})]} - \frac{1}{1+\exp[-a(\theta-b_k)]}$$

where a is the item discrimination parameter and b_k is the threshold for score category k. $P(x=k)$ represents the item characteristic curve describing the probability that the response is in category k or higher. The value b_{k-1} represents "the point on the Θ–axis at which the probability passes 50% that the response is in category k or higher" (Lukhele, Thissen & Wainer, 1993; p. 8). Samejima's graded response model assumes that an order exists for responses to an item such that response $k+1$ indicates higher proficiency than response k. Thus, Samejima's model is congruent with the 2-12 scoring scale for the WST essay.

[3]Because this model fixed the asymptote parameter, it can also be described as a modified two-parameter model (i.e., only two parameters were estimated).

In applying Samejima's model to test data, Multilog allows for up to 10 score categories per item. No candidates received a perfect score (of 12) on the essay and so the essay responses were represented by 10 response categories (2 to 11)[4].

Calibration procedure

One calibration run and three separate scoring runs were used to calibrate separate proficiency values (thetas) for the examinees. The calibration run provided item parameter estimates for the MC items and the essay on a common scale (μ=0, σ=1). Using these item parameters, a MCθ was derived using only the parameters for the MC items (as if the essay were not included on the test), an essayθ was derived using only the parameters for the essay, and a compositeθ was derived using the parameters from both the MC items and the essay.

Results

Factor Analyses

Factor analyses were performed on the MC and combined data sets. Full-information factor analyses (Bock, Gibbons, & Muraki, 1988) were conducted on the MC item data using the Testfact computer program (Wilson, Wood, & Gibbons, 1991). Two- through five-dimensional solutions were obtained. The first factor had an eigenvalue of 22.8, the second factor had an eigenvalue of 2.7, and the other three factors had eigenvalues less than 2. Though the change in the loglikelihoods exhibited statistically significant fit for all additional factors (2 through 5), the magnitude of the first factor relative to the second factor was taken to indicate essential unidimensionality. (Bock, et al. point out that the change in -2loglikelihoods is not a good indication of correct dimensionally because statistical significance is typically obtained.) A re-analysis of the MC data using the Dimtest program (Stout, 1987) corroborated the hypothesis of essential dimensionally (t=1.08, p=.14).

To evaluate the dimensionally of the complete WST an item parcel factor analysis approach was used (Dorans & Lawrence, 1987; Thissen, Wainer, & Wang, 1994). Five ten-item parcels were created out of the 50 MC items. These parcels were similar in difficulty (average item difficulty of the parcels ranged from .74 to .75) and had the same number of score categories (11) as the essay (0 to 10 versus 2-12)[5].

---

[4]Only a small proportion (.0004) of GED candidates receive a perfect score on the essay (by both readers), so it was not unusual that no perfect scores were present in this sample. Furthermore, collapsing over extreme score categories when more than 10 categories are present generally does not result in a loss of information when fitting an IRT model (Sireci, Thissen, and Wainer, 1991; Wainer, Sireci, & Thissen, 1991).

[5]To create the item parcels, the test items were ordered in ascending order of difficulty. The items were then assigned to parcels sequentially: the least difficult item was assigned to the first parcel, the fifth-least difficult item was assigned to the fifth parcel. The assignment of items to parcels was reversed after each sequence (e.g., the sixth-least difficult item was assigned to the fifth parcel, the seventh-least difficult item was assigned to the fourth parcel, etc.). This strategy (similar

A 6X6 lower-triangular correlation matrix comprising polychoric correlation's among the MC item parcels and the essay was computed using Prelis-2 (Jöreskog & Sörbom, 1993b). This correlation matrix was factor-analyzed using LISREL-8 (Jöreskog & Sörbom, 1993a) to test for significant difference between a one-factor model for all test items, and a two-factor model where the essay corresponds to a separate factor. The one-factor model fit the data very well ($G^2(9)=15$, $p=.09$) and the standardized factor loading for the essay was .76 (standardized factor loading for parcels ranged from .60 to .83). A two-factor model, where the essay was linked to a separate, uncorrelated factor did not fit the data ($G^2(10)=1,400$, $p\leq.001$). These results confirmed the hypothesis of unidimensionality of the WST. Additional hierarchical models specifying a general factor for all items and correlated unique factors for the MC and essay (e.g., Thissen et al, 1994) were not investigated since they were not central to the purpose of this study.

IRT Results

WST calibration

No problems were encountered in calibrating the MC and essay items onto a common scale using Multilog. The observed and predicted proportions of respondents in all item score categories were extremely similar (within .01), indicating good model fit. The marginal reliability for the WST, as calibrated by Multilog, was .88.

Marginal reliability and test information

Based on a separate scoring run using only the parameters for the MC items, the marginal reliability for the MC portion of the WST was .87. A scoring run using only the essay parameters yielded a marginal reliability of .34. Comparing these values to the value obtained using all item parameters (.88), we see that inclusion of the essay resulted in a very slight increase in total reliability. It is also clear that using the essay alone, without the MC items, results in very low score reliability. Using the Spearman-Brown prophecy formula, approximately 13 essay items would be needed to achieve the level of reliability obtained by the MC items. From a test information perspective, the reliability of the essay portion of the WST is equivalent to that of about four MC items.

The case for the essay is improved when looking at total test information. The Test Information Functions (TIFs) based on all items (MC plus essay), and MC items only, are presented in Figure 1. Test information is increased across the entire scale when the essay is included. Test information is maximized (for both MC-only and MC plus essay) for the Θ interval -1.5 to -1.0. The item information function for the essay is presented in Figure 2. The essay information function corroborates the finding that

---

to Thissen, Wainer, & Wang, 1994) was employed to create parcels that could be considered equivalent in terms of difficulty and content.

the essay contributes information along the entire Θ scale. However, the essay contributes relatively less information around the "middle" of the scale (-1.0 to .5) and relatively greater information towards the extremes. The item characteristic curve for the essay is plotted in Figure 3 (essay score of 2 is labeled 1, score of 3 is labeled 2, etc.). As expected, the probability of getting a higher score on the essay increases as proficiency increases. The trace lines for the two highest observed essay score categories (i.e., scores of 10 and 11) are not plotted in Figure 3 because they correspond to theta values beyond 3.0 (thresholds for these score categories are 4.3 and 5.6, respectively).

Insert Figure 1 About Here

Insert Figure 2 About Here

Insert Figure 3 About Here

Determining the essay weight

As mentioned above, IRT weights all test items according to the amount of information they contribute to the proficiency scale. Unfortunately, because IRT models assume the particular item set calibrated is a sample from the population of all test items, weights are not "assigned" to the items and are not reported. To determine the relative weight attributed to the essay, and to derive separate proficiency (Θ) scores for the candidates, the separate scoring runs were used to calculate three Θ scores for candidates: MCΘ, essayΘ, and compositeΘ. Given these separate theta scores, and the marginal reliabilities associated with them, the weight assigned to the MC and essay portions of the WST were estimated using Nunnally's (1978) formula for the reliability of a weighted linear composite:

$$r_{yy} = 1 - \frac{b_{MC}^2 + b_{Essay}^2 - (b_{MC}^2 r_{MC} + b_{Essay}^2 r_{Essay})}{b_{MC}^2 + b_{Essay}^2 + 2 b_{MC} b_{Essay} r_{MC,Essay}}$$

where: r_{yy}=the reliability of the weighted linear composite, b_{MC} and b_{Essay} are MC and essay weights, r_{MC} and r_{Essay} are the reliabilities of the MC and essay portions,

and $r_{MC,Essay}$ is the correlation between the MC and essay scores. Although Nunnally's formula is typically used to determine the reliability of a composite score, given known weights of the subcomponets, we use the formula here to work backwards and solve for the unknown weights. Substituting $1\text{-}b_{Essay}$ as the weight for the MC portion (subcomponent weights must sum to 1.0), and given the marginal reliabilities of .88 for the composite score, .87 for the MC portion, .34 for the essay, and .62 for the correlation between the MC and essay, the essay weight was estimated to be .14[6].

Weighted results

Given that separate MC and essay Θ scores were calculated for all candidates (and were on the same scale) weighting these subscores using the current .64/.36 weighting scheme was straightforward: WCΘ=.64(MCΘ) + .36(EssayΘ). The WCΘ (weighted composite theta) score was then scaled to the GED score scale using the same linear transformation procedures that are used to transform raw scores to the GED scale (lower bound of 20, upper bound of 80, mean of 50 and standard deviation of 10). This procedure allowed for placing the IRT Θ scores on the GED scale according to the a priori weighting scheme. Plugging these a priori weights into Nunnally's formula, along with the IRT-derived marginal reliability estimates for the MC (.87) and essay (.34) portions of the test, and given the correlation of .53 between MC_Θ and $essay_\Theta$ scores, the reliability of the weighted IRT-derived composite score was .82. Thus, adjusting the "optimal" IRT weights for the MC and essay portions of the test led to a reduction in total test reliability. A comparison of the reliability estimates for the composite score, MC score, and essay score are presented in Table 1. Both classical and IRT-derived (marginal) reliability estimates are presented.

Insert Table 1 About Here

Effect of weighting on passing the WST

As a first step in evaluating the impact of the essay on passing the WST, correlations were computed between the MC_Θ, $essay_\Theta$, and pass/fail on the WST. Three pass/fail variables were created: the first (Pass40) was defined as achieving a GED standard score of 40, the second (Pass45) was defined as achieving a GED standard score of 45, and the third (Pass30%) was defined as achieving a GED score that corresponded to the 30th percentile of the score distribution. This score was 43 on the current

[6]The correlation between the MC and essay Θ scores was .53 for these data; however using this value in Nunnally's formula led to weights that were indeterminate. Previous research found the correlation between the MC and essay portions of the WST to be as high as .62 (Patience & Swartz, 1987; Auchter, Sireci, & Skaggs, 1993). Using .62 as the correlation provided solutions that were not indeterminate. This procedure was tested in other examples where all the parameters in Nunnally's formula were known. The formula calculated the weights correctly in all examples.

GED score scale and 45.1 on the IRT-derived GED score scale (i.e., based on transformation from the WC_Θ scale). These correlations are presented in Table 2. The IRT-derived essay scores ($Essay_\Theta$) and the raw essay scores (GED_{Essay}) exhibited lower correlations than the MC scores and the composite scores. The raw MC scores (GED_{MC}) exhibited the highest correlations for all three passing variables, followed by the IRT-derived MC scores (MC_Θ). The correlations for the IRT-derived composite score (CS_Θ), IRT-derived weighted composite score (WC_Θ), and the current GED composite score (GED_CS) were lower than the MC correlations, due to inclusion of the (less-correlated) essay.

Insert Table 2 About Here

Table 3 presents the passing rates broken down by MC and essay information. The passing rates based on these separate portions of the test are presented for both the IRT and GED scaling procedures. The essay did not have a large impact on passing rate using IRT scaling. However, the essay resulted in a reduction in passing rate of about 6-8% using the current GED scaling procedure.

Insert Table 3 About Here

To address the question of whether the essay would affect the passing status of GED examinees differentially according to scaling procedure (IRT versus classical), "discordant" passing rates for each scaling procedure were calculated. These rates indicate the percentage of candidates whose passing status would change depending on whether the WST was all-objective, or MC plus essay. The results, presented in Table 4, indicate that passing classification differences would occur, and that IRT scaling results in fewer classification differences than the current GED scaling.

Insert Table 4 About Here

To evaluate the effect of weighting the MC and essay portions on passing status, the pass rates for Pass40, Pass45, and Pass30% were compared across: 1) the GED standard scores resulting from the optimally-weighted composite score scale, 2) the scores resulting from weighting the MC_Θ and $essay_\Theta$ scores (.64 and .36, respectively), and 3) the standard scores actually assigned to the candidates using the current GED scaling procedures (non-IRT). The results of this comparison are presented in Table 5. The passing rates across the IRT solutions were very similar.

The passing rates between the IRT and classical (GED) scaling procedures differed by 3-7 percentage points. This differences is probably due to a sensitivity to sampling fluctuation for the classically-derived GED scale.

Discussion

This research demonstrated that the WST can be calibrated using IRT. Furthermore, the results indicate that it is possible to maintain the current weighting scheme of the MC and essay portions of the WST using IRT. However, changing the optimal, IRT-derived, weight for the essay (from .14 to .36) reduced the total test reliability from .88 to .82. Using the optimal IRT "weights" did not substantially diminish the impact of the essay. The IRT-calibrated essay contributed information along the entire proficiency scale (Figure 1), and provided significant impact on passing status (Table 4). Given the decrease in total test reliability after weighting, the results suggest that the IRT-derived (optimal) weight for the essay should be used if the WST is calibrated using IRT.

Using the current GED scoring procedures, the essay is essential for determining pass/fail status. As illustrated in Table 4, about 6-10% more candidates would pass the WST if it were a MC-only examination. Thus, the essay appears to screen out candidates who are successful on the MC portion, but are relatively weaker when it comes to producing a writing sample.

The impact of the essay on passing the WST was reduced somewhat using IRT. This finding is probably due to the extreme unidimensionality of the assessment (i.e., IRT maximizes information on the unidimensional latent distribution), and the fact that the essay information decreased within the range of the Θ distribution where passing decisions are made. The small difference in passing rates based on MC_{Θ} and the IRT-derived weighted composite (Table 3) illustrates the invariance of IRT over samples of items drawn from the population of all potential items. However, the discordant passing rates (Table 4) demonstrate that even within an IRT model, about 5-6% more candidates would pass if the tests were comprised of only MC items. Therefore, including the essay within the IRT scale results in an improvement of about 5-6% in passing classification.

To increase the impact of the essay on passing the WST, the essay must discriminate better among "middle proficiency" candidates. Such improved discrimination could be accomplished by refining the scoring rubric for the middle of the essay raw score scale (e.g., improved differentiation between scores of 3 and 4 on the six-point scale). This refinement would probably increase the IRT-derived (optimal) weight of the essay above .14. An alternative to maximizing information within the Θ interval where passing decisions are made (-1.0 to -.5) would be to add a few relatively more difficult MC items. However this action would reduce the impact of the essay.

In comparing passing rates across the current GED scoring procedures and an IRT-based procedure (Table 5), it was found that IRT scoring resulted in passing rates

more closely aligned with the GED norm group. There was very little fluctuation between passing rates for Pass45 and Pass30% when using IRT, but relatively large differences under the current scoring regimen (Table 3). Because Pass30% was defined as the 30th percentile of the sample, this passing standard corresponded to a GED standard score of 43 (i.e., 70% of the sample achieved this score or better). The 30th percentile on the IRT-derived scale corresponded to a GED_Θ score of 45.1, which is identical to the scale score for the GED norm group. Thus, the sample invariance property of IRT (over samples of examinees) appears to hold across both GED candidate and high school senior populations (the GED norm group).

If GEDTS were to use IRT in test development, conversion tables linking raw MC and essay scores to the GED scale would be needed. Thissen, Pommerich, Billeaud, and Williams (1994), and Thissen, Pommerich, and Billeaud (1994) demonstrate a procedure for converting summed MC and free-response raw scores to IRT-derived scale scores. Using Thissen's et al. procedure, IRT-based methods for item tryout, norming, scaling, and equating, could be implemented without changing the look, or score reporting procedures, of the current GED test battery.

This study investigated only one particular WST form. Though it is likely that the IRT results would generalize to other test forms, the research conducted here should be applied to other test forms and essay topics. Future research should also use IRT item selection procedures to approximate target WST test information functions so that information is maximized over the Θ interval where passing score decisions are made.

The results here are consistent with the findings of Donoghue (1994) and Wainer and Thissen (1993) in terms of total IRT information achieved via MC and polytomous items. Polytomously-scored items, such as the WST essay, provide more information than MC items. However, when testing time is taken into account, MC items provide greater information within a given time period. However, the magnitude of this gain in "information-per-minute" using MC items varies according to the test studied (Donoghue, 1994). With respect to the WST, if it comprised only essays, about 13 essays (requiring 9 hours and 45 minutes of testing time) would be necessary to obtain the current level of reliability. If the WST were completely objective, about four additional well-chosen MC questions would be needed[7]. Given these trade-offs, it appears that the current composition of the WST is efficient.

This study focused on the reliability of WST composite scores. It did not compare the differential validity of the MC and essay item types. Questions such as "Do the MC and essay portions predict workplace (or collegiate) writing skills differentially?" are important and should be examined in future research. It is clear that writing skills

[7]Lukhele, Thissen & Wainer (1994) found that using all response information in MC items (i.e., giving "partial credit" for incorrect MC responses) yielded an improvement in test information beyond that of free response items. So, another option for increasing score precision would be to use a partial credit scoring scheme on an all objective test.

proficiency can be measured using both MC and essay items. Taking the concerns of test score reliability and construct validity together, it appears that hybrid tests, using MC items, essays, and other potential item types (e.g., short answer) represent a commendable compromise. In terms of scaling the WST, the results of this study suggest that to maximize reliability using an IRT calibration model, the weights for the MC and essay sections should not be adjusted a posteriori.

References

Anastasi, A. (1988). *Psychological Testing*. New York: Macmillan.

Auchter, J.C., Sireci, S.G., & Skaggs, G. (1993). The Tests of General Educational Development technical manual. Washington, D.C.: American Council on Education.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement*, 12, 261-280.

Divgi, D.R., (1984). Does small N justify use of the Rasch model? Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31*, 295-311.

Dorans, N. J., & Lawrence, I.M. (1987). The internal construct validity of the SAT (Research Report). Princeton, NJ: Educational Testing Service.

Linn, R.L. (1994). Criterion-referenced measurement: a valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice, 13*, 12-15.

Jöreskog, K.G., & Sörbom, D. (1993a) *LISREL-8 User's Reference Guide.* Mooresville, IN: Scientific Software.

Jöreskog, K.G., & Sörbom, D. (1993b) *Prelis-2 User's Reference Guide.* Mooresville, IN: Scientific Software.

Lukhele, R., Thissen, D., & Wainer, H. (1993). *On the relative value of multiple-choice, constructed response, & examinee-selected items on two achievement tests.* ETS Technical Report. Princeton, N.J.: Educational Testing Service.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement*, (#rd ed.). Washington, DC: American Council on Education.

Nunnally, J.C. (1978). *Psychometric theory* (2nd printing). New York: McGraw Hill.

Patience, W., & Swartz, R.(1987). Essay score reliability: issues in and methods of reporting the GED Writing Skills Test scores. Paper presented at the annual

meeting of the National Council on Measurement in Education, Washington, D.C.: April.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 4, Part 2, Whole #17.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Thissen, D. (1991). Multilog: Multiple categorical item analysis and test scoring using item response theory, version 6 [computer program]. Mooresville, IN: Scientific Software.

Thissen, D., Pommerich, M., & Billeaud, K. (1994). Item response theory for combining scores on tests including polychotomous items with ordered responses. Unpublished paper, University of North Carolina at Chapel Hill.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1994). *Item response theory for scores on tests including polychotomous items with ordered responses.* L.L. Thurstone Psychometric Laboratory Technical Report. Chapel Hill, N.C.: University of North Carolina.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement,* 31, 113-123.

Wainer, H., Sireci, S.G., & Thissen, D. (1991). Differential testlet functioning: definition and detection. *Journal of Educational Measurement, 28,* 197-219.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103-108.

Wilson, D., Wood, R., & Gibbons, R. (1991). Testfact: Test scoring, item statistics, and item factor analysis, 386/486 version [computer program]. Chicago, IL: Scientific Software, Inc.

Table 1

Marginal (IRT) and Classical Reliability Estimates for WST Composite and Subscores

	IRT	Classical
Composite:		
Optimally Weighted	.88	N/A
Weighted .64/.36	.82	.86
MC	.87	.87
Essay	.34	.58

Notes: The weighted composite reliabilities were estimated using Nunnally's (1978) formula. The MC classical reliability is a KR20. The classical estimate of reliability for the essay is the correlation between two essay scores obtained from a sample of graduating high school seniors tested in the Spring of 1994. N/A=not available.

Table 2

Correlations Between Subscores, Composite Scores, and Success on the WST

	Pass40	Pass45	Pass30%
GED_{MC}	.67	.78	.75
MC_{Θ}	.60	.75	.69
GED_CS	.60	.75	.70
CS_{Θ}	.52	.75	.68
WC_{Θ}	.51	.73	.66
GED_{Essay}	.52	.53	.54
$Essay_{\Theta}$	.30	.41	.37

Table 3

WST Passing Rates Using MC Only Versus MC Plus Essay

	GED-MC	GED-CS	MCΘ	WCΘ
Pass40	94.7%	88.2%	86.0%	85.6%
Pass30%	83.8%	77.0%	70.6%	70.3%
Pass45	73.8%	65.7%	72.2%	72.6%

Table 4

WST Discordant Passing Rates Due to MC Only Versus MC Plus Essay

	Current GED Scale	Weighted IRT Scale
Pass 40	6.5%	5.2%
Pass30%	8.1%	5.2%
Pass 45	10.0%	6.3%

Table 5

WST Passing Rates Across Three Scaling Solutions

	Unweighted IRT	Weighted IRT	Current GED
Pass40	85.9%	85.6%	88.2%
Pass45	72.4%	72.6%	65.7%
Pass30%	70.2%	70.3%	77.0%

Figure 1: Test Information Functions for Complete WST and MC Only

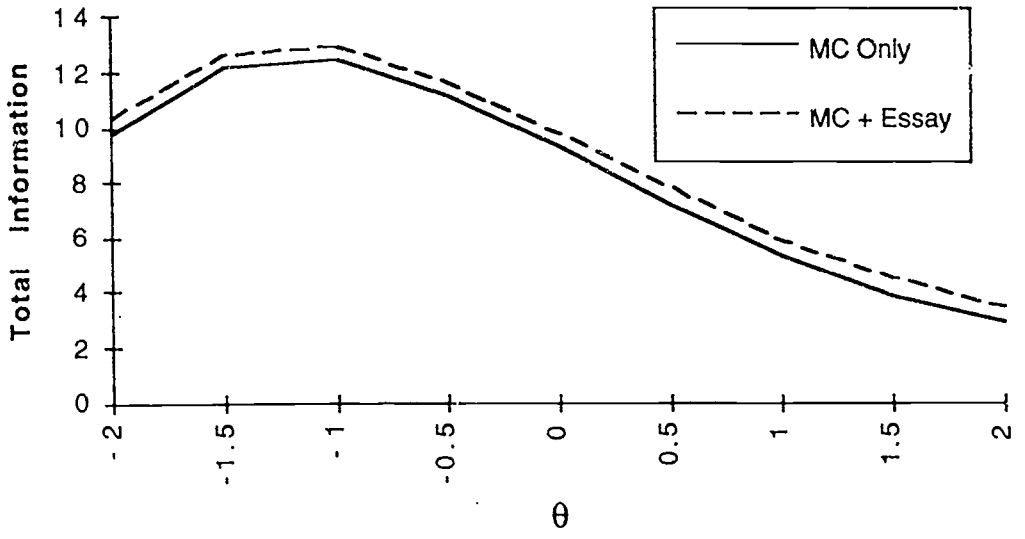

Figure 2: Information Function for the WST Essay

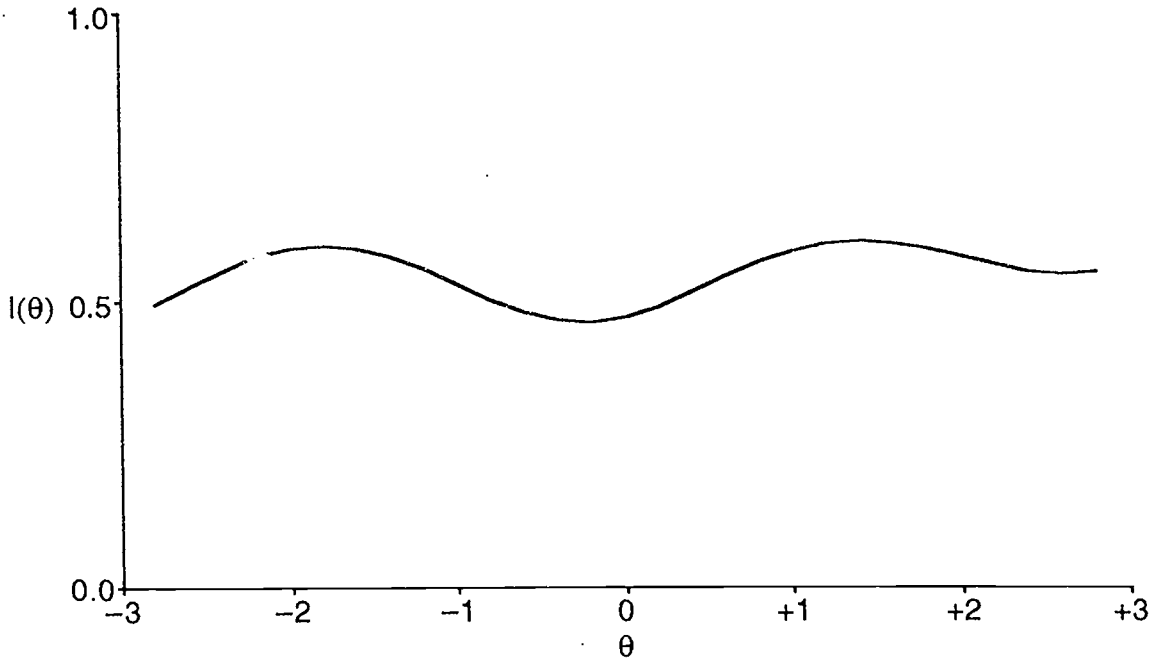

Figure 3: WST Essay Item Characteristic Curve