

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

by Xitao Fan

Despite theoretical differences between item response theory (IRT) and classical test theory (CTT), there is a lack of empirical knowledge about how, and to what extent, the IRT- and CTT-based item and person statistics behave differently. This study empirically examined the behaviors of the item and person statistics derived from these two measurement frameworks. The study focused on two issues: (a) What are the empirical relationships between IRT- and CTT-based item and person statistics? and (b) To what extent are the item statistics from IRT and those from CTT invariant across different participant samples? A large-scale statewide assessment database was used in the study. The findings indicate that the person and item statistics derived from the two measurement frameworks are quite comparable. The degree of invariance of item statistics across samples, usually considered as the theoretical superiority IRT models, also appeared to be similar for the two measurement frameworks.

© COPYRIGHT 1998 Sage Publications, Inc.

Classical test theory (CTT) and item response theory (IRT) are widely perceived as representing two very different measurement frameworks. However, few studies have empirically examined the similarities and differences in the parameters estimated using the two frameworks. Prior to exploring this issue in some detail, some readers may appreciate a brief review of related theories. Additional detail is provided elsewhere (cf. Crocker & Algina, 1986; McKinley & Mills, 1989).

Brief Review of CTT and IRT

Although CTT has served the measurement community for most of this century, IRT has witnessed an exponential growth in recent decades. The major advantage of CTT are its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). Relatively weak theoretical assumptions not only characterize CTT but also its extensions (e.g., generalizability theory). Although CTT's major focus is on test-level information, item statistics (i.e., item difficulty and item discrimination) are also an important part of the CTT model.

At the item level, the CTT model is relatively simple. CTT does not invoke a complex theoretical model to relate an examinee's ability to success on a particular item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored). This success rate of a particular pool of examinees on an item, well known as the p value of the item, is used as the index for the item difficulty (actually, it is an inverse indicator of item difficulty, with higher value indicating an easier item). The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination, which is often expressed statistically as the Pearson product-moment correlation coefficient between the scores on the item (e.g., 0 and 1 on an item scored right-wrong) and the scores on the total test. When an item is dichotomously scored, this estimate is often computed as a point-biserial correlation coefficient.

The major limitation of CTT can be summarized as circular dependency: (a) The person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent. This circular dependency poses some theoretical difficulties in CTT's application in some measurement situations (e.g., test equating, computerized adaptive testing).

Despite the theoretical weakness of CTT in terms of its circular dependency of item and person statistics, measurement experts have worked out practical solutions within the framework of CTT for some otherwise difficult measurement problems. For example, test equating can be accomplished empirically within the CTT framework (e.g., equipercentile equating). Similarly, empirical approaches have been proposed to accomplish item-invariant measurement (e.g., Thurstone absolute scaling) (Englehard, 1990). It is fair to say that, to a great extent, although there are some issues that may not have been addressed theoretically within the CTT framework, many have been addressed through ad hoc empirical procedures.

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

IRT, on the other hand, is more theory grounded and models the probabilistic distribution of examinees' success at the item level. As its name indicates, IRT primarily focuses on the item-level information in contrast to the CTT's primary focus on test-level information. The IRT framework encompasses a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. For test items that are dichotomously scored, there are three IRT models, known as three-, two-, and one-parameter IRT models.

Although the one-parameter model is the simplest of the three models, it may be better to start from the most complex, the three-parameter IRT model; the reason for this sequence of discussion will soon become obvious. The IRT three-parameter model takes the following form:

(1) [MATHEMATICAL EXPRESSION NOT REPRODUCIBLE IN ASCII]

where $c_{.sub.i}$ is the guessing factor, $a_{.sub.i}$ is the item discrimination parameter commonly known as item slope, $b_{.sub.i}$ is the item difficulty parameter commonly known as the item location parameter, D is an arbitrary constant (normally, $D = 1.7$), and $[\Theta]$ is the ability level of a particular examinee. The item location parameter is on the same scale of ability, $[\Theta]$, and takes the value of $[\Theta]$ at the point at which an examinee with the ability-level $[\Theta]$ has a 50/50 probability of answering the item correctly. The item discrimination parameter is the slope of the tangent line of the item characteristic curve at the point of the location parameter.

When the guessing factor is assumed or constrained to be zero ($c_{.sub.i} = 0$), the three-parameter model is reduced to the two-parameter model for which only item location and item slope parameters need to be estimated:

(2) [MATHEMATICAL EXPRESSION NOT REPRODUCIBLE IN ASCII]

If another restriction is imposed that stipulates that all items have equal and fixed discrimination, then $a_{.sub.i}$ becomes a constant rather than a variable, and as such, this parameter does not require estimation, and the IRT model is further reduced to

(3) [MATHEMATICAL EXPRESSION NOT REPRODUCIBLE IN ASCII]

So, for the one-parameter IRT model, constraints have been imposed on two of the three possible item parameters, and item difficulty remains the only item parameter that needs to be estimated. This one-parameter model is often known as the Rasch model, named after the researcher who did pioneer work in the area. It is clear from the discussion that the three-parameter model is the most general model, and the other two IRT models (two- and one-parameter models) can be considered as models nested or subsumed under the three-parameter model.

Purpose of the Study

Theoretically, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person statistics. As a result, in theory, IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This invariance property of item and person statistics of IRT has been illustrated theoretically (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) and has been widely accepted within the measurement community.

The invariance property of IRT model parameters makes it theoretically possible to solve some important measurement problems that have been difficult to handle within the CTT framework, such as those encountered in test equating and computerized adaptive testing (Hambleton et al., 1991). However, as the cornerstone of IRT, the importance of the invariance property of IRT model parameters cannot be overstated, because, without this crucial property, the complexity of IRT models can hardly be justified on either theoretical or practical grounds.

Because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between the IRT- and CTT-based item and person statistics. Theoretically, such relationships are not entirely clear, except that the two types of statistics should be

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

monotonically related under certain conditions (Crocker & Algina, 1986; Lord, 1980). But such relationships have rarely been empirically investigated, and, as a result, they are largely unknown.

The empirical studies available in this area have primarily focused on the application of the two methods in test equating (e.g., Becker & Forsyth, 1992; Harris, 1991). With regard to test equating, Hambleton et al. (1991) suggested that, theoretically, the invariance property of the IRT item statistics obviated the need of equating tests; instead, it is (linear) scaling, rather than equating, that is necessary within the framework of IRT. The discussion implies that IRT models handle equating tasks better than the CTT equating approaches. The empirical studies in this area, however, provide a mixed picture, with some indicating the superiority of IRT approaches (e.g., Peterson, Cook, & Stocking, 1983), some suggesting better results from CTT ad hoc approaches (e.g., Clemans, 1993; Kolen, 1981; Skaggs & Lissitz, 1986a), and still some finding that both CTT and IRT equating methods produce very comparable results (Skaggs & Lissitz, 1988). The mixed picture has prompted some researchers to suggest that it might be unrealistic to expect one method to provide the best equating results for all types of tests (e.g., Skaggs & Lissitz, 1986b).

A literature search revealed only one study that empirically examined the comparability of IRT-based and CTT-based item and person statistics. Lawson (1991) compared IRT-based (one-parameter Rasch model) and CTT-based item and person statistics for three different data sets, and showed exceptionally strong relationships between the IRT- and CTF-based item and person statistics. The results of the study, although the study was based on somewhat small data sets and only examined the most restrictive one-parameter IRT model, suggest that information from the two approaches about items and examinees might be very much the same.

Similarly, the invariance property of IRT item/person parameters has been little explored empirically, although invariance has been illustrated theoretically (e.g., Hambleton & Swaminathan, 1985; Rudner, 1983). However, Miller and Linn (1988), using an extant large data set, did report the results of a study examining the variations of item characteristic functions in the context of instructional coverage variations. They reported relatively large differences in item curve responses, suggesting lack of invariance of IRT item parameters. Lack of invariance was also reported by Cook, Eignor, and Taft (1988) for both CTT- and IRT-based item difficulty estimates.

Given the limited number of empirical studies directly or indirectly addressing the invariance issue, there is an obvious lack of systematic investigation about the absolute invariance of the item and person statistics obtained from either CTF or IRT frameworks and a lack of studies that empirically compare the relative invariance of item and person statistics obtained from CTT versus those from IRT. The major criticism for CTT is its inability to produce item/person statistics that would be invariant across examinee/item samples. This criticism has been the major impetus for the development of IRT models and for the exponential growth of IRT research and applications in the recent decades.

It is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scarce. It appears that the superiority of IRT over CTT in this regard has been taken for granted by the measurement community, and no empirical scrutiny has been deemed necessary. The empirical silence on this issue seems to be an anomaly. This lack of empirical investigation has prompted some researchers to state that item response modeling has been too focused on mathematical elaboration at the expense of empirical exploration (Goldstein & Wood, 1989).

The present study was designed to "constructively replicate" (Gall, Borg, & Gall, 1996; Lykken, 1968) the study by Lawson (1991). The present study focused on two major issues: (a) How comparable are the item and person statistics from the CTT framework with those from the IRT framework? and (b) How invariant are the item statistics of CTT and IRT across examinee samples? More specifically, the study addressed the following five research questions:

1. How comparable are the CTT-based and IRT-based examinee ability estimates?
2. How comparable are the CTT-based and IRT-based item difficulty estimates?
3. How comparable are the CTT-based and IRT-based item discrimination estimates?
4. How invariant are the CTT-based and IRT-based item difficulty estimates across different participant samples?

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

5. How invariant are the CTT-based and IRT-based item discrimination estimates across different participant samples?

Methods

Data Source

The data used in this study are from the Texas Assessment of Academic Skills (TAAS) tests administered in October 1992 to 11th-grade students. Designed for assessing the mastery of school instructional objectives, TAAS is a state-mandated, criterion-referenced test battery consisting of reading, math and writing tests. The writing test contained both multiple-choice and essay items. The reading (48 items) and the math (60 items) tests consisted of multiple-choice items scored dichotomously as either correct or incorrect. Unattempted items were scored as incorrect responses. Data from the reading and math tests were used in the present study. The participant pool for the database had more than 193,000 participants. Table 1 presents the demographic information of the participant pool of this database.

Table 1 Ethnicity and Gender Composition of the Participant Pool (n = 193,240)

Group	Frequency	Percentage
Ethnicity		
American Indian	526	0.3
Asian American	5,815	3.0
African American	24,714	12.8
Hispanic	59,918	31.0
White	98,166	50.8
Unknown/not indicated	4,101	2.1
Gender		
Female	98,240	50.8
Male	94,610	49.0
Unknown/not indicated	390	0.2

TAAS was designed to be a test battery for assessing the minimum competency of students in Texas public schools in several academic areas. As is generally the case for mastery tests, TAAS test items were primarily curriculum content based, and test score distributions were not normally distributed; rather, the score distributions exhibit obvious ceiling effects, as indicated by the frequency distributions in Figure 1.

[Figure 1 ILLUSTRATION OMITTED]

Participant Sampling

To examine the issues related to IRT and CTT statistics, three sampling plans were implemented for math and reading test data so that the behaviors of IRT and CTT statistics could be examined under different examinee sample conditions. The three sampling plans generated samples that were progressively more dissimilar. This sampling strategy allowed the examination of the behaviors of IRT and CTT statistics across progressively less comparable participant samples. All samples in the present study had a sample size of 1,000, which is considered sufficiently large even for the estimation of IRT parameters in the three-parameter IRT model.

Random samples. Random samples of examinees, each consisting of 1,000 examinees, were drawn from the entire participant pool. Twenty such samples were drawn for math test data and 20 for reading, making the total number of random participant samples 40. Because these were random samples from the same population, the random samples should be comparable with each other within the limits of statistical sampling error.

Gender group samples. Samples of female participants and those of male participants were randomly drawn separately for math and reading test data. Twenty female samples and 20 male samples were drawn for each test, making the total number of gender samples 80. Because the female and male samples were drawn from different populations, as defined by the demographic variable gender, theoretically there should be more dissimilarity between a female sample and a male

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

sample than between the two random samples described in the previous section. Table 2 presents the performance statistics for the female and male groups. It is seen that the female and male groups had comparable performance for reading, although there is a slight difference at the test level for math.

Table 2
Performance Characteristics of Female and Male Groups

Test	Gender	M	SD	Q1	Median	Q3
Reading	Female	37.17	7.43	33	39	43
	Male	37.48	7.48	33	39	43
Math	Female	41.81	11.02	34	43	51
	Male	43.26	11.26	36	45	53

Note. Q1 = first quartile (25th percentile); median = second quartile (50th percentile); Q3 = third quartile (75th percentile).

Truncated high-ability and low-ability group samples. This sampling plan generated samples that were different in terms of performance on the tests. The high-ability group was defined as those whose scores fell within the 15th to 100th percentile range on the math or the reading test. The low-ability group was defined as those whose scores fell within the 0 to 85th percentile range on the math or the reading test. Twenty samples were randomly drawn from each of the two groups, separately for each test, making the total number of high-ability and low-ability samples 80. Because these two groups were defined in terms of test performance, not in terms of a demographic variable as in the gender group sampling, there should be more dissimilarity between a high-ability sample and a low-ability sample than between a female and a male sample pair.

Comparability of IRT and CTT Person Statistics

The comparability of IRT- and CTT-based person statistics (ability [θ] in IRT vs. obtained score T in CM) was assessed by correlating the [θ] and T estimates obtained from the same sample of participants. [θ] values were obtained through the IRT program BILOG (PC Version 3.07, for one-, two-, and three-parameter IRT models, respectively), and the obtained score T in CTT was simply the raw score. CTT-obtained score T was correlated with IRT ability [θ] estimated through one-, two-, and three-parameter IRT models. All IRT estimations were carried out using the marginal maximum likelihood (MML) method, which is the default for the BILOG program. Analyses were replicated for different participant samples (random, gender, and truncated ability group samples) and for items of both math and reading tests.

Comparability of IRT and CTT Item Statistics

The comparability of IRT- and CTT-based item statistics was examined by correlating IRT and CTT item statistics obtained from the same sample of participants. Two types of item statistics were compared: (a) item difficulty parameter b (item location parameter) from IRT models with CTT item difficulty p value and (b) IRT item discrimination parameter a (item slope parameter from two- and three-parameter IRT models) with CTT item discrimination index [$r_{\text{sub.pb}}$] (item-test, point-biserial correlation). The [$r_{\text{sub.pb}}$] for CTT was bias corrected (i.e., the contribution of an item score to the total score was removed before calculating the [$r_{\text{sub.pb}}$] for the item).

Degree of Invariance of IRT and CTT Item Statistics

As discussed in the literature review sections, the invariance property of item statistics is crucial. The degree of invariance of item statistics was assessed by correlating item parameter estimates of two different samples within each measurement framework. The three sampling plans discussed previously allowed the assessment of item statistics invariance across progressively dissimilar samples: (a) between two random samples of the same population, (b) between female and male samples, and (c) between high- and low-ability samples. This progression of dissimilarity between samples facilitated the assessment of the degree of invariance of item statistics for the two measurement frameworks.

Transformations for CTT p Value and Item-Test Correlation

- Reprinted with permission. Additional copying is prohibited. -

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

In CTT, the item difficulty index p (p value), the proportion of examinees passing an item, expresses item difficulty on an ordinal scale not on an interval scale. This p value, however, can easily be transformed to an interval scale so that it is more appropriate for statistical analyses. The transformation simply requires the assumption that the underlying trait being measured by an item is normally distributed. The transformation is achieved by finding the z score corresponding to the $(1 - p)$ th percentile from the z distribution. For example, if the p value of an item is .84 (84% of the examinees passed the item), the z value for such a p value will be -1 , as indicated in Figure 2. This normalization removes the curvilinearity in the relationship between two sets of item p values (Anastasi, 1988).

[Figure 2 ILLUSTRATION OMITTED]

This transformation of the CTT item difficulty index has been widely used in different measurement situations, such as in Thurstone absolute scaling (Donlon, 1984; Thurstone, 1947) and in research related to item bias detection (Angoff, 1982; Cole & Moss, 1993). In the present study, correlation analyses were carried out both between original p values obtained from sample pairs and between normalized p values to assess the invariance characteristic of the CTT item difficulty index.

In CTT item discrimination is expressed as the item-test, Pearson product-moment correlation (point-biserial correlation). Because the correlation coefficient is not linearly scaled (Hinkle, Wiersma, & Jurs, 1988), the Fisher z transformation is usually recommended before statistical analyses are applied to correlation coefficients. For this reason, in the assessment of the invariance characteristic of the CTT item discrimination index, correlation analyses were applied to both original item point-biserial coefficients and to Fisher z -transformed point-biserials between two samples of examinees.

Results and Discussion

The results of the study are discussed as responses to the five research questions presented previously. Whenever appropriate, relevant interpretation and discussion about the meaning and implications of the results are presented together with the results. But before the results related to the research questions are presented, the question of IRT model fit is addressed.

IRT Model Fit Assessment

In any application of the IRT model, it is important to assess to what extent the IRT model assumptions are valid for the given data and how well the testing data fit the IRT model selected for use in that particular situation. The violation of IRT model assumptions, or misfit between the IRT model used and the testing data, may lead to erroneous or unstable IRT model parameter estimates. In the present study, the assessment of IRT model fit was conducted on a simple random sample of 6,000 examinees for TAAS math and reading tests (equal sample size for the two tests, but different samples). The large sample size used in assessing IRT model assumption and model fit should have provided stable and trustworthy results about the model assumption and model fit.

Unidimensionality is the most important assumption common for all IRT models. This assumption is sometimes empirically assessed by investigating whether a dominant factor exists among all the items of the test (Hambleton et al., 1991). The first three eigenvalues for the 60 test items on the TAAS math test were 11.4, 1.5, and 1.3. The first three eigenvalues for the 48 test items on the TAAS reading test were 8.4, 1.5, and 1.3. Based on these results, it appeared reasonable to conclude that the unidimensionality assumption for the IRT models held for the data used in the study.

Model-data fit was assessed by checking if the individual test items misfit the given IRT model. In BILOG (V 3.07), a likelihood-ratio [chi square] test (for a test with more than 20 items), which assesses the discrepancy between the expected response pattern and the actual response pattern of the subjects on a particular item in relation to their performance on the test as a whole, is conducted for each item (Mislevy & Bock, 1990). Table 3 summarizes the number of items identified as misfitting the given IRT model at the $[\text{Alpha}] = .01$ level.

Table 3
Number of Misfitting Items Identified for the Two Tests
([Alpha] = .01)

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

Test	Number of Item	IRT Models		
		1P	2P	3P
Math	60	20	2	1
Reading	48	18	3	1

Note. IRT = item response theory; 1 P = one parameter, 2P = two parameter; 3P = three parameter.

It is worth pointing out that, given the subject sample size of 6,000 used in the analyses for assessing the IRT model fit, the statistical test for identifying misfitting items has a lot of statistical power. Even with the powerful statistical test, only one or two items are identified as misfitting the two- and three-parameter IRT model. The results indicate that the data fit the two- and three-parameter IRT models exceptionally well. The fit of the data for the one-parameter model, however, is obviously very questionable, with about 30% of the items identified as misfitting the IRT model for either test. Because there is the obvious misfit between the data and the one-parameter IRT model, and because the consequences of such misfit are not entirely clear (Hambleton et al., 1991), the results related to the one-parameter IRT model presented in later sections should be viewed with extreme caution.

Research Question 1

For the first research question ("How comparable are the CTT-based and IRT-based examinee ability estimates?"), Table 4 presents the results for both math and reading tests and for samples under different sampling conditions. Three steps were involved in arriving at each entry in Table 4: (a) from each sample of examinees, both CTT- and IRT-based (one-, two-, and three-parameter IRT models, respectively) ability estimates were obtained; (b) the CTT- and IRT-based ability estimates from the same sample were correlated; and (c) the correlations between CTT- and IRT-based ability estimates from individual samples were averaged across samples for the same test and under the same sampling condition. Each table entry is the average of 20 correlation coefficients obtained from 20 random samples ($n = 1,000$ each). In Table 4, and all the following tables, an average correlation coefficient was obtained through (a) transforming individual correlation coefficients to Fisher Zs, (b) averaging the Fisher Zs, and (c) transforming the average Fisher Z back to the Pearson correlation coefficient.

Table 4 Comparability of Person Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Person Ability Estimates

Sampling Plan	Tests	IRT Models(a)		
		1P	2P	3P
Random samples	Math	.983(.001)	.975(.001)	.983(.001)
	Reading	.978(.002)	.969(.002)	.984(.002)
Gender group sampling	Math	.985(.001)	.976(.002)	.984(.001)
	Reading	.979(.002)	.970(.002)	.974(.001)
Female samples	Math	.985(.001)	.976(.002)	.984(.001)
	Reading	.979(.002)	.970(.002)	.974(.001)
Male samples	Math	.980(.001)	.973(.002)	.983(.002)
	Reading	.976(.001)	.966(.001)	.971(.002)
Truncated ability group sampling	Math	.985(.001)	.977(.001)	.982(.001)
	Reading	.988(.001)	.980(.001)	.966(.002)
High-ability samples	Math	.997(.000)	.987(.000)	.990(.000)
	Reading	.994(.000)	.974(.001)	.977(.001)
Low-ability samples	Math	.997(.000)	.987(.000)	.990(.000)
	Reading	.994(.000)	.974(.001)	.977(.001)

Note. CTT = classical test theory; IRT = item response theory. An average correlation coefficient was obtained through (a) transforming individual correlation coefficients to Fisher Zs, (b) averaging the Fisher Zs, and (c) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations are presented in parentheses.

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

(a.) These are average correlations between CTT ability estimates and those derived from one-, two-, and three-parameter IRT models, respectively.

Table 4 shows that the CTT- and IRT-based examinee ability estimates correlate extremely highly with each other for both math and reading tests, for different samples, and for all three (one-, two-, and three-parameter) IRT models, with average correlations between CTT- and IRT-based ability estimates greater than .96 for all conditions. These very high correlations indicate that CTT- and IRT-based person ability estimates are very comparable with each other. In other words, regardless of which measurement framework we rely on, the same or very similar conclusions will be drawn regarding the ability levels of individual examinees.

Research Question 2

Table 5 presents the results associated with the second research question ("How comparable are the CTT-based and IRT-based item difficulty estimates?"). Again, from the same sample, CTT-based item difficulty estimates were correlated with IRT item difficulty estimates derived from IRT models (one-, two-, and three-parameter IRT models), and each table entry is the average of within-sample correlations between CTT- and IRT-based item difficulty estimates from 20 samples. The IRT-based item difficulty estimates were correlated both with original CTT p values and with normalized CTT p values. The CTT p values were reversed in direction so that the higher the value, the more difficult the item. This linear reversal of p value direction had no statistical effect other than to make the correlations in Table 5 positive in sign.

Table 5 Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Item Difficulty Indexes

Sampling Plan	Subsample	Test	IRT Models(a)
			CTT p Values
			1P
Random samples		Math	.990(.001)
		Reading	.980(.001)
Gender group sampling	Females	Math	.989(.001)
		Reading	.980(.002)
	Males	Math	.989(.001)
		Reading	.980(.002)
Truncated ability group sampling	High ability	Math	.981(.002)
		Reading	.953(.003)
	Low ability	Math	.991(.001)
		Reading	.984(.001)

Sampling Plan	IRT Models(a)	
	2P	3P
Random samples	.910(.012)	.934(.008)
	.920(.021)	.931(.015)
Gender group sampling	.901(.021)	.926(.011)
	.923(.019)	.932(.013)
	.906(.016)	.934(.011)
	.915(.023)	.925(.009)
Truncated ability group sampling	.803(.029)	.884(.017)
	.841(.029)	.877(.020)
	.918(.016)	.936(.011)
	.916(.016)	.940(.009)

Sampling Plan	IRT Models(a)		
	Transformed p Values		
	1P	2P	3P

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

Random samples	.999(.000)	.925(.012)	.947(.007)
	.999(.000)	.927(.015)	.952(.010)
	.999(.000)	.916(.020)	.942(.009)
Gender group sampling	.999(.000)	.935(.013)	.956(.008)
	.999(.000)	.925(.013)	.945(.008)
	.999(.000)	.919(.022)	.946(.009)
Truncated ability group sampling	.999(.000)	.830(.029)	.918(.014)
	.998(.000)	.862(.033)	.932(.018)
	.999(.000)	.928(.014)	.944(.010)
	.999(.000)	.905(.011)	.936(.006)

Note. CTT = classical test theory; IRT = item response theory. An average correlation coefficient was obtained through (a) transforming individual correlation coefficients to Fisher Zs, (b) averaging the Fisher Zs, and (c) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations are presented in parentheses.

(a.) Correlations between CTT item difficulty with IRT item difficulty estimates derived from one-, two-, and three-parameter IRT models, respectively.

(b.) Correlations between IRT item difficulty estimates and CTT item p values.

(c.) Correlations between IRT item difficulty estimates and CTT normalized item p values.

As the tabled results indicate, for the IRT Rasch model (i.e., the one-parameter model), the relationship between CTT- and IRT-based item difficulty estimates is almost perfect. For the two- and three-parameter IRT models, the relationships between CTT- and IRT-based item difficulty estimates appear somewhat weaker, although still quite strong, because the majority of the coefficients were above .90 under most conditions.

It is interesting to note that the relationship between CTT- and IRT-based item difficulty estimates seems slightly, but invariably, weaker for the two-parameter IRT model than for the three-parameter IRT model. It is not clear why this should be the case or whether this may be due to the idiosyncracies of the item samples used in the study. Overall, with a few cases in which the average correlations between CTT- and IRT-based item difficulty estimates were in the .80s, the CTT- and IRT-based item difficulty estimates were quite comparable with each other for the two different tests and under the three different sampling conditions.

Because the IRT Rasch model (one-parameter IRT model) assumes fixed item discrimination and no guessing for all items, the model only provides estimates for item parameter of difficulty. Because item difficulty parameter estimates of the Rasch model were almost perfectly related to CTT-based item difficulty indexes (both original and normalized), it appears that the one-parameter model provides almost the same information as CTT with regard to item difficulty but at the cost of considerable model complexity. Unless Rasch model estimates could show superior performance in terms of invariance across different samples over that of CTT item difficulty indexes, the results here would suggest that the Rasch model might not offer any empirical advantage over the much simpler CTT framework. The degree of invariance of item statistics of the two measurement frameworks will be discussed shortly under Research Questions 4 and 5.

Research Question 3

Table 6 presents the results associated with the third research question ("How comparable are the CTT-based and IRT-based item discrimination estimates?"). Each table entry is the average of within-sample correlations between CTT item point-biserial correlations (original and Fisher Z transformed) and IRT discrimination estimates (IRT item slopes). Because the one-parameter model assumes fixed item discrimination for all items, no correlation between CTT point-biserials and the single fixed Rasch model item discrimination--a constant--could be computed; hence, N/A (not applicable) is entered under the column for the one-parameter IRT model in the table.

Table 6 Comparability of Item Statistics From the Two Measurement Frameworks: Average Correlations Between CTT- and IRT-Based Item Discrimination Indexes

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

			IRT Models(a)	
			CTT p Values(b)	
Sampling Plan	Subsample,	Test	1P	2P
Random samples		Math	N/A	.900(.014)
		Reading	N/A	.788(.036)
Gender group sampling	Females	Math	N/A	.893(.006)
		Reading	N/A	.750(.041)
	Males	Math	N/A	.894(.012)
		Reading	N/A	.776(.035)
Truncated ability group sampling	High ability	Math	N/A	.831(.021)
		Reading	N/A	.507(.046)
	Low ability	Math	N/A	.947(.007)
		Reading	N/A	.907(.011)

			IRT Models(a)	
			CTT p Values(b)	Transformed p Values(c)
Sampling Plan	3P	1P	2P	
Random samples	.688(.063)	N/A	.900(.015)	
	.260(.096)	N/A	.793(.037)	
Gender group sampling	.705(.049)	N/A	.895(.007)	
	.358(.100)	N/A	.756(.042)	
	.728(.069)	N/A	.898(.013)	
	.192(.135)	N/A	.781(.035)	
Truncated ability group sampling	.829(.043)	N/A	.832(.021)	
	.635(.052)	N/A	.507(.045)	
	.771(.050)	N/A	.950(.007)	
	.883(.016)	N/A	.912(.012)	

		IRT Models(a)
		Transformed p Values(c)
Sampling Plan	3P	
Random samples	.693(.062)	
	.264(.093)	
Gender group sampling	.712(.048)	
	.358(.098)	
	.732(.068)	
	.199(.133)	
Truncated ability group sampling	.835(.041)	
	.638(.053)	
	.776(.050)	
	.893(.014)	

Note. CTT = classical test theory; IRT = item response theory. An average correlation coefficient was obtained through (a) transforming individual correlation coefficients to Fisher Zs, (b) averaging the Fisher Zs, and (c) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations are presented in parentheses.

(a.) Correlations between CTT item discrimination indexes with IRT item discrimination estimates derived from one-, two-, and three-parameter IRT models, respectively.

(b.) Correlations between IRT item discrimination estimates and CTT item point-biserials.

(c.) Correlations between IRT item discrimination estimates and Fisher Z transformed CTT point-biserials.

- Reprinted with permission. Additional copying is prohibited. -

GALE GROUP

Information Integrity

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

In contrast to Tables 4 and 5, in which the overwhelmingly strong relationships between CTT- and IRT-based estimates of examinee ability and item difficulty indicate extraordinary comparability between the two measurement frameworks, the relationship between CTT and IRT item discrimination indexes is weaker, as indicated by the majority of the averaged within-sample correlations being within the range of .60 to .90. Furthermore, this relationship shows considerable variation across tests (math vs. reading), across sampling conditions, and across IRT models (two- vs. three-parameter IRT models).

Although the relationship between the CTT-based and IRT-based item discrimination indexes in Table 6 could be considered strong or somewhat strong under some conditions (high .80s to low .90s), the relationship is precariously low (.20s to .30s) in a few cases. All the extremely low correlations occurred for the reading test items under IRT three-parameter model. In general, the item discrimination indexes from the IRT three-parameter model correlated somewhat less with CTT point-biserials than did those from IRT two-parameter model.

The results in Table 6 show that the item discrimination indexes from the CTT and IRT frameworks tend to be less comparable than the person ability estimates and the item difficulty estimates presented previously. The lower comparability between the discrimination indexes derived from CTT and IRT implies that, in some cases, CTT and IRT may yield noticeable discrepancies with regard to which items have more discrimination power, which, in turn, may lead to the selection of different items for a test, depending on which framework is used in the estimation of item discrimination.

Up to this time, we have solely focused on the question of comparability between estimates derived from the two measurement frameworks. Low comparability between item discrimination indexes of CTT and IRT in some cases does not inform us about which measurement framework provides more stable, or more invariant, item parameter estimates across different samples. To understand the invariance characteristics of the item statistics of the two measurement frameworks, we turn now to Research Questions 4 and 5.

Research Question 4

The fourth research question ("How invariant are the CTT-based and IRT-based item difficulty estimates across different participant samples?") addresses one crucial question about CTT and IRT. As discussed previously, the assumption of item parameter invariance across different participant samples has played the most important role in the development and application of IRT models.

Table 7 presents the results for this research question. Notice that the averaged correlations in this table (and, similarly, in Table 8) are correlations between item difficulty estimates from two different samples derived from the same measurement framework. For example, the entry under CTT--between female-male samples for math test--is the average of the correlations between item CTT p values obtained from a female sample and those CTT p values obtained from a male sample. One hundred such female-male sample pairs were formed and p values correlated within each pair. The average of these 100 correlation coefficients was .945 (SD = .007). Other entries in Table 7 were obtained in the same fashion. It is important to note that the invariance property of item parameters only can be investigated by administering the same items to different samples and then comparing item parameter estimates obtained across samples.

Table 7 Invariance of Item Statistics From the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Difficulty Indexes

Invariance Across	Test
Random samples	Math
Female-male samples	Reading
	Math
High-low ability samples	Reading
	Math
	Reading
	CTT

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

Invariance Across	p Values	Transformed p values(a)
Random samples Math	.988(.002)	.988(.002)
	.993(.001)	.991(.002)
Female-male samples	.945(.007)	.947(.007)
	.978(.003)	.974(.974)
High-low ability samples	.974(.005)	.978(.005)
	.988(.002)	.982(.003)

IRT

Invariance Across	1P	2P	3P
Random samples Math	.988(.002)	.968(.010)	.965(.009)
	.991(.002)	.966(.012)	.969(.009)
Female-male samples	.947(.007)	.929(.014)	.926(.014)
	.973(.004)	.955(.010)	.955(.009)
High-low ability samples	.978(.005)	.907(.029)	.925(.014)
	.979(.003)	.862(.029)	.877(.025)

Note. CTT = classical test theory; IRT = item response theory; 1P = one parameter, 2P = two parameter; 3P = three parameter. An average correlation coefficient was obtained through (a) transforming individual correlation coefficients to Fisher Zs, (b) averaging the Fisher Zs, and (c) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations are presented in parentheses.

(a.) These are average between-sample correlations between normalized p values.

Table 8 Invariance of Item Statistics From the Two Measurement Frameworks: Average Between-Sample Correlations of CTT and IRT Item Discrimination Indexes

Invariance Across	Test	
Random samples	Math	
	Reading	
Female-male samples	Math	
	Reading	
High-low ability samples	Math	
	Reading	

CTT		
Fisher Z Transformed		
Invariance Across	Point-Biserial	Point-Biserial(a)
Random samples	.924(.016)	.924(.016)
	.855(.029)	.855(.028)
Female-male samples	.898(.018)	.898(.018)
	.804(.035)	.805(.034)
High-low ability samples	.605(.033)	.604(.034)
	.102(.063)	.106(.063)

IRT			
Invariance Across	1P	2P	3P
Random samples	N/A	.906(.019)	.857(.037)
	N/A	.891(.024)	.920(.025)
Female-male samples	N/A	.877(.023)	.837(.029)
	N/A	.864(.029)	.880(.044)
High-low ability samples	N/A	.748(.034)	.631(.055)
	N/A	.636(.078)	.020(.089)

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

Note. CTT = classical test theory; IRT item response theory; 1P = one parameter, 2P = two parameter; 3P three parameter. An average correlation coefficient was obtained through (a) transforming individual correlation coefficients to Fisher Zs, (b) averaging the Fisher Zs, and (c) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations are presented in parentheses.

(a.) These are average between-sample correlations between Fisher Z transformed item discrimination indexes.

The comparison of the tabled CTT with IRT entries indicates that CTT item difficulty estimates are closely comparable with IRT item difficulty estimates in terms of their invariance properties in the sense that the average between-sample correlation coefficients of item difficulty estimates are very high and are comparable between CTT and IRT. If there is any trend at all, it appears that CTT item difficulty estimates are slightly more invariant than IRT item difficulty estimates in almost all conditions, because the average between-sample correlations of p values appear to be slightly higher than the average between-sample correlations of IRT location parameters for all three of the IRT models (one-, two-, and three-parameter models). This empirical observation about the invariance property of the item difficulty indexes of the two measurement frameworks is quite interesting in light of the strong arguments in favor of the IRT framework due to its ostensible advantage over CTT with regard to invariance.

The average between-sample correlations for the original CTT p values and those for the normalized p values differ very little. Also, the IRT one-parameter model difficulty estimates appear to be slightly more invariant across samples than the two- and three-parameter model item difficulty estimates. Considering that "invariance only holds when the fit of the model to the data is exact in the population" (Hambleton et al., 1991, p. 23), does this result imply that the one-parameter model fits the data slightly better than the two- and three-parameter models? The results in Table 3 evaluating model fit indicate that the reverse is probably true. Also, from the point of view of statistical modeling, this seems to be unlikely, because the one-parameter model can be considered simply as a submodel nested under the two- or three-parameter model. Theoretically, a model higher in a model hierarchy tends to provide better fit than a model nested under it, because the lower model has more constraints. A constrained parameter will tend to increase the misfit of the model, and the question is usually, How much? If the misfit caused by the constrained parameter is minimal relative to the gain in model parsimony, the simpler and more restrictive model will generally be preferred.

Research Question 5

The last research question asked was, "How invariant are the CTT-based and IRT-based item discrimination indexes across different participant samples?" Table 8 presents the results of correlation analyses for CTT with IRT item discrimination indexes. As explained before, because the IRT one-parameter (Rasch) model does not provide item discrimination estimates for individual items, and instead assumes fixed item discrimination for all items, no correlations could be computed for the one-parameter model; hence, N/A is listed under the one-parameter IRT column in the table. It is also worth pointing out again that each table entry is either the average of correlations of point-biserials of CTT between two samples or the average of correlations of item slopes of IRT between two samples. Each entry is the average of 100 correlation coefficients obtained from 100 sample pairs.

The item discrimination indexes of both CTT and IRT were less invariant across participant samples than were the item difficulty indexes presented in Table 7. This result parallels what was observed about comparability between CTT and IRT item statistics in Tables 5 and 6. Also, with higher correlations of CTT point-biserials in some cases and higher correlations of IRT item slopes in others, no systematic advantage of one framework over another is obvious. In most cases, the average between-sample correlations of item discrimination indexes of CTT and those of IRT were quite comparable and moderately high (high .80s to low .90s), indicating reasonable invariance across samples.

But the invariance of item discrimination indexes from both CTT and IRT decreased with the increase of dissimilarity between samples. In other words, the item discrimination indexes of both CTT and IRT were most invariant across random samples, they were less invariant across female-male samples (i.e., the female-male sample pair was more dissimilar than the random sample pair), and they were least invariant across high-low ability samples (i.e., the high-low ability sample pair was the most dissimilar among the three sampling conditions).

For the last condition (reading test, between high-low ability samples), the CTT point-biserials totally collapsed in terms of

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

their cross-sample invariance ($r = .102$ and $r = .106$). IRT two-parameter model item slopes maintained moderate invariance ($r = .636$), but the three-parameter model item slopes again collapsed ($r = .020$). This empirical observation is somewhat puzzling.

As discussed previously, theoretically, if parameters could be adequately estimated for the given sample size, a higher order (less restrictive) model should tend to provide better fit than a lower order (more restrictive) model, although such better fit comes at the expense of model parsimony. If a better fit is obtained, greater invariance of item parameters would be expected (Hambleton et al., 1991). The fact that, in this situation, the two-parameter IRT model had moderately invariant item discrimination indexes and the three-parameter IRT model item discrimination indexes showed no invariance for the same data is contrary to both intuition and theoretical expectation.

Summary and Conclusion

The present study empirically examined the behavior of item and person statistics obtained from the CTT and IRT measurement frameworks. The study focused on two main issues: (a) How comparable are the item and person statistics from CTT with those from IRT? and (b) How invariant are the CTT item statistics and the IRT item statistics, respectively? A large-scale test database from a statewide assessment program was used as the empirical basis of the investigation. The test item pool was composed of two tests (math and reading) with 60 and 48 dichotomously scored items in each, and the participant pool had more than 193,000 examinees who took both tests. Random samples ($n = 1,000$) were drawn from the participant pool under three sampling plans, producing progressively more dissimilar sample pairs. The increasing dissimilarity between samples facilitated the assessment of the degree of invariance of the CTT and IRT item statistics.

The major findings were as follows:

1. The person statistics (examinee ability estimates) from CTT were highly comparable with those from IRT for all three IRT models.
2. The item difficulty indexes from CTT were very comparable with those from all IRT models and especially from the Rasch model.
3. Compared with item difficulty indexes, the item discrimination indexes from CTT were somewhat less comparable with those from IRT. Although under the majority of the conditions, the comparability was moderately high to high, there were a few cases where the comparability was very low.
4. Both CTT and IRT item difficulty indexes exhibited very high invariance across samples, even across samples that were quite different (samples from high- and low-ability groups). The degree of invariance of the CTT item difficulty index was highly comparable with, if not better than, that of IRT item difficulty parameter estimates.
5. Both the CTT and IRT item discrimination estimates were somewhat less invariant than their item difficulty estimates. For both CTT and IRT item discrimination estimates, the degree of invariance decreased steadily as samples became more dissimilar, implying that item discrimination parameters from neither CTT nor IRT could maintain a high degree of parameter invariance across populations that are different. The degree of invariance of CTT item discrimination estimates was highly comparable with that of IRT item discrimination estimates.

Overall, the findings from this empirical investigation failed to discredit the CTT framework with regard to its alleged inability to produce person-invariant item statistics; conversely, the findings failed to support the IRT framework for its ostensible superiority over CTT in producing person-invariant item statistics. The findings here simply show that the two measurement frameworks produced very similar item and person statistics both in terms of the comparability of item and person statistics between the two frameworks and in terms of the degree of invariance of item statistics from the two competing measurement frameworks. These findings pose some interesting questions about how to view the differences between IRT and CTT models both in theory and in testing practice.

As discussed at the beginning of this article, the invariance property of item and person statistics is the most important

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

model feature that needs to be evaluated, because the argument that IRT item and person statistics possess invariance whereas CTT item and person statistics do not has been the impetus for the development and the use of IRT measurement models. This argument has been widely accepted within the measurement community.

Unfortunately, the view that the argument is moot seems to have occurred largely in the vacuum of empirical evidence, because the literature fails to show that this important premise has been subjected to systematic and rigorous empirical investigation. It is my view that in psychological measurement, as in any other areas of science, theoretical models are important in guiding our research and practice. But the merits of a theoretical model should ultimately be validated through rigorous empirical scrutiny.

Of course, the present empirical study, like many other research studies, had its share of limitations that may potentially undermine the validity of its findings. First of all, the characteristics of the test items used in the study may be somewhat unique. As discussed at the beginning of the Methods section and as indicated by Figure 1, the test score distributions show strong ceiling effects, as is generally the case for minimum-competency tests or other criterion-referenced mastery tests. The strong ceiling effects suggest that many items tended to be very easy. Although it is unclear what systematic impact this characteristic of the data may have had on the results, it would be desirable in future studies to replicate the present study using data from norm-referenced testing, which usually involves items varying more in item difficulty and in item discrimination.

The second shortcoming of the investigation is the somewhat limited item pool used in the study. Although the examinee pool is quite adequate in the sense that a variety of different samples can be drawn from it, the same cannot be said about the item pool. Ideally, the test item pool should be larger and more diverse in terms of item characteristics so that items can be sampled from the pool to study the behaviors of CTT and IRT item statistics under different conditions of item characteristics. Future studies may benefit from using several different testing databases, and Monte Carlo studies that artificially specify different item characteristics will greatly help to address the issues.

Early in the last decade, Robert L. Thorndike (1982) made the following comment with regard to IRT measurement models:

For the large bulk of testing, both with locally developed and with standardized tests, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties. (p. 12)

The findings of the present study suggest that Thorndike's skepticism regarding IRT models possessing inherent unique advantages over CTT estimates appears to have been warranted.

References

- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.
- Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement*, 29, 341-354.
- Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment*, 1, 329-347.

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

- Cole, N. S., & Moss, P. A. (1993). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 201-219). Phoenix, AZ: Oryx Press.
- Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, 25, 31-45.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Donlon, G. (1984). *The college board technical handbook for the Scholastic Aptitude Test and achievement tests*. New York: College Entrance Examination Board.
- Englehard, G., Jr. (1990, April). Thorndike, Thurstone and Rasch: A comparison of their approaches to item-invariant measurement. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC Document Reproduction Services No. ED 320 921)
- Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. White Plains, NY Longman.
- Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology*, 42, 139-167.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 3847.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Harris, D. J. (1991). A comparison of Angoff's Design I and Design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement*, 28, 221-235.
- Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1988). *Applied statistics for the behavioral sciences* (2nd ed.). Boston: Houghton Mifflin.
- Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, 18, 1-11.
- Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lykken, D. T. (1968). Statistical significance of psychological research. *Psychological Bulletin*, 70, 155-159.
- McKinley, R., & Mills, C. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 71-135). Greenwich, CT: JAI.
- Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, 25, 205-219.
- Mislevy, R.J., & Bock, R. D. (1990). *BILOG3: Item analysis and test scoring with binary logistic models*. Chicago: Scientific Software International.

Item response theory and classical test theory: an empirical comparison of their item/person statistics.

Peterson, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, 8, 137-156.

Rudner, L. M. (1983). A closer look at latent trait parameter invariance. *Educational and Psychological Measurement*, 43, 951-955.

Skaggs, G., & Lissitz, R. W. (1986a). An exploration of the robustness of four test equating models. *Applied Psychological Measurement*, 10, 303-317.

Skaggs, G., & Lissitz, R. W. (1986b). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, 56, 495-529.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement*, 12, 69-82.

Thorndike, R. L. (1982). Educational measurement: Theory and practice. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology: Contributions of latent trait theory* (pp. 3-13). Princeton, NJ: ERIC Clearinghouse of Tests, Measurements, and Evaluations. (ERIC Document Reproduction Service No. ED 222 545)

Thurstone, L. L. (1947). The calibration of test items. *American Psychologist*, 2, 103-104.

The author thanks the Texas Education Agency for the data used in the study. Please address correspondence about this article to the author at the Department of Psychology, Utah State University, Logan, UT 84322-28 10; e-mail fafan@cc.usu.edu.