A MONTE CARLO COMPARISON OF ITEM AND PERSON STATISTICS BASED ON ITEM RESPONSE THEORY VERSUS CLASSICAL TEST THEORY

PAUL MACDONALD AND SAMPO V. PAUNONEN
University of Western Ontario

Despite the well-known theoretical advantages of item response theory (IRT) over classical test theory (CTT), research examining their empirical properties has failed to reveal consistent, demonstrable differences. Using Monte Carlo techniques with simulated test data, this study examined the behavior of item and person statistics obtained from these two measurement frameworks. The findings suggest IRT- and CTT-based item difficulty and person ability estimates were highly comparable, invariant, and accurate in the test conditions simulated. However, whereas item discrimination estimates based on IRT were accurate across most of the experimental conditions, CTT-based item discrimination estimates proved accurate under some conditions only. Implications of the results of this study for psychometric item analysis and item selection are discussed.

The development of achievement, ability, aptitude, interest, and personality tests is generally a multistep process that can follow one of two distinct measurement frameworks. These are usually called the classical test theory (CTT) and the item response theory (IRT) measurement strategies. Depending on which method the test constructor chooses, different steps are taken in the statistical analysis of the initial pool of test items, possibly leading to different selections of those items for the final test form. The question facing a test constructor is whether these differences will result in substantively different products. If so, is one product superior to the other in terms of the test's overall psychometric properties? The purpose of this article is to report new empirical data bearing on these questions.

Educational and Psychological Measurement, Vol. 62 No. 6, December 2002 921-943
DOI: 10.1177/0013164402238082

CTT Versus IRT

Under the CTT framework, item analysis largely consists of calculating difficulty and discrimination indices for each item. The difficulty of an item is estimated by the proportion of examinees who endorse a dichotomous item in the keyed direction (e.g., true or false) or who "pass" an item by choosing the correct response. The rate of item endorsement, or item passing, is referred to as the item mean, item difficulty, or item *p* value, whereby a value approaching 1.0 indicates an easy item and a value approaching .0 indicates a difficult item.

Item discrimination relates to the ability of an item to differentiate between examinees of varying levels of ability. The discrimination of an item is often estimated by the Pearson product–moment correlation (r_{it}) between participants' responses to the item (e.g., either 0 or 1 for items scored dichotomously) and the participants' total test scores. In some CTT applications, "corrected" item discrimination is determined by computing total scores excluding the item scores (e.g., 0 or 1) on a given item being analyzed to avoid inflating the correlation by including the influence of the item's scores in both the variables being correlated. A large item discrimination r_{it} value indicates that the item effectively differentiates between high- and low-ability examinees, whereas a near-zero or negative item discrimination r_{it} value indicates poor examinee differentiation.

Limitations of CTT indices of item difficulty and item discrimination indices have been noted by Lord (1953) and more recently by several other researchers (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Nunnally & Bernstein, 1994). The principal limitation mentioned is that the item statistics and the person statistics (i.e., observed test scores) are dependent. That is, estimates of item difficulty and discrimination are dependent on the particular group of examinees completing the test, and estimates of person ability are dependent on the particular test items administered.

To illustrate the dependency of the item and person statistics under CTT, consider a test measuring some ability of interest. In that test, examinee ability scores are dependent on the difficulty of the test items. Thus, if the test is composed of relatively easy items, the person statistics (i.e., observed test scores) will be relatively high, giving the impression that the examinees possess high levels of ability. If the test is composed of relatively difficult items, however, the person statistics will be relatively low, giving the impression that the examinees possess low levels of ability. As such, estimates of examinee ability are dependent on the difficulty of the test items. Similarly, the item difficulty estimates are dependent on the ability of the examinees. If examinees completing the test are high in ability, then item *p* values will also be high, suggesting that the items were easy. Conversely, if the examinees completing the test are low in ability, then *p* values will be similarly low, sug-

gesting that the items were difficult. Similar statistical interdependencies exist between the observed scores and the item discrimination indices of CTT.

Under the IRT framework, item analysis also consists of estimating item statistics. When dealing with items that have been scored dichotomously, three related IRT models are popular in the psychometric literature. The most complex of these model is called the three-parameter IRT model. That model takes the form

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \tag{1}$$

where c_i is an item guessing parameter, b_i is an item difficulty parameter, a_i is an item discrimination parameter, and D is a scaling constant (usually D = 1.702). Note that the probability of an examinee's responding correctly to an item, $P_i(\theta)$, is also dependent on θ, his or her level of the trait being assessed. Readers interested in a more detailed explanation of the three-parameter IRT model are directed elsewhere for a comprehensive presentation of that and other models (e.g., Baker, 1992; Crocker & Algina, 1986; McKinley & Mills, 1989.

The three-parameter IRT model can be constrained to form the simpler two-parameter IRT model by removing the item guessing parameter c_i. The reduced model, therefore, contains only estimates of item difficulty and item discrimination and has the form

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}. \tag{2}$$

A further restriction can be imposed to create the one-parameter IRT model (or Rasch model). In that model, the item discrimination parameter a_i is constrained such that all items have equal and fixed discrimination level a. Therefore, the only parameter to be estimated is the item difficulty b_i. The one-parameter IRT model takes the form

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}}. \tag{3}$$

In theory, measures based on IRT overcome the principal limitation of measures based on CTT. That is, item parameter estimates are not dependent on the particular sample of examinees who have been administered the test items, and the person ability estimates are not dependent on the particular sample of test items administered. This invariance property of IRT models has been demonstrated extensively and has been widely accepted

(Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Rudner, 1983; van der Linden & Hambleton, 1997).

Past Comparisons of CTT and IRT

Despite the theoretical advantages attributed to IRT over CTT, little has been done to demonstrate them empirically. An early attempt to contrast the two measurement frameworks was conducted by Tinsley and Dawis (1977). In their study, the authors confirmed that person ability estimates based on the one-parameter IRT model were independent of the difficulty of the test items. They also confirmed that person ability estimates based on CTT (i.e., test total score *T*) were not independent of the item difficulty. Thus, the CTT person statistic score *T* was influenced by test difficulty, but the IRT person parameter θ was not. These putative advantages of IRT models were not, however, always demonstrated in later research.

A study by Cook, Eignor, and Taft (1988) examined the stability of item statistics based on IRT and CTT. Responses of students were collected using two forms of a biology admissions test. From the two different test administrations, item statistics were estimated for the three-parameter IRT model and for CTT. Cook et al. found that item difficulty estimates were unstable for both measurement frameworks because the item estimates differed between the two test administrations in each case. Unexpectedly, the authors found that the item difficulty estimates were slightly more stable for CTT estimates than for IRT estimates.

Lawson (1991) compared item and person statistics from three data sets based on the one-parameter IRT model with statistics based on CTT. His analyses led him to report that the estimates from the two measurement frameworks were "almost identical" (p. 166). Lawson suggested further that for persons involved with the design and administration of testing instruments, IRT appeared to offer few advantages over CTT.

The above finding of no empirical advantage for IRT models when dealing with real data was supported by Ndalichako and Rogers (1997). In their study, the responses of students in a school-leaving reading comprehension exam were analyzed under IRT and CTT. The researchers found that person estimates of ability for the two strategies were correlated almost perfectly (as high as .988). With such a high degree of comparability, in combination with the ease of estimating the CTT statistics, Ndalichako and Rogers favored the continued use of CTT for test scoring and item analysis.

In a recent comparison of IRT and CTT, Fan (1998) examined item and person statistics using CTT and IRT methods. For that study, samples of response data were extracted from students (N = 193,240) who completed a math exam (60 items) and a reading exam (48 items). From each exam bank, several random samples of 1,000 participants were analyzed under the one-,

two-, and three-parameter IRT models and under CTT. Estimates of item difficulty, item discrimination, and person ability were then assessed for invariance across the random samples and for comparability across the two measurement frameworks.

Fan (1998) found that item difficulty and person ability estimates were highly comparable between the IRT and CTT measurement methods. However, the comparability of item discrimination estimates across methods ranged from high to low depending on the characteristics of the particular respondent samples being evaluated. Fan also found the invariance of item difficulty and item discrimination estimates under CTT were as good as, if not better than, estimates under IRT. He concluded that his overall findings failed to support the superiority of the IRT framework and indicated that the two measurement frameworks produce highly similar item and person statistics. In fact, he reiterated a popular quote by Robert L. Thorndike (1982) on the future of IRT 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)

Purpose of the Study

We note at this point that the findings from past empirical investigations comparing IRT- and CTT-based item and person statistics should not be generalized to all educational and psychological tests. As Fan (1998) noted in his study, research comparing IRT and CTT models have typically contrasted item and person statistics obtained from a small number of real tests. As such, the particular collection of items constituting the test may be unique in its properties (e.g., number of items, level of item difficulty, degree of item discrimination). And with those unique characteristics, the question arises as to how results based on those particular tests generalize to other tests with different characteristics. Fan suggested that future studies may overcome this limitation by using artificial tests whose characteristics can be manipulated experimentally. This study was designed to address that concern.

In this study, we sought to replicate and extend the studies by Fan (1998) and Lawson (1991), both of whom employed archival data sets in their comparisons of IRT and CTT frameworks. We chose, instead, Monte Carlo simulations to investigate the comparability, invariance, and accuracy of IRT and CTT parameter estimates under a variety of testing conditions. In the first phase of the study, simulated test items were generated based on the one- and two-parameter IRT models to create artificial tests measuring a hypothetical ability of interest. Using these tests, responses of simulated examinees were

generated to create our response data sets. In the second phase of the study, item and person parameter estimates were obtained for each simulated response data set according to IRT and CTT methods. In the final phase, the obtained parameter estimates were evaluated to assess their comparability and accuracy under CTT and IRT models.

The assessment of the IRT and CTT measurement methods in this study focused on three major questions: (a) How comparable are the item and person statistics from IRT and CTT frameworks? (b) How invariant are the item statistics of IRT and CTT across examinee samples? and (c) How accurate are the item and person statistics from IRT and CTT frameworks vis-à-vis known population parameters?

Method

Simulated Data Sets

The process of generating sets of simulated test item response involved four stages. In the first stage, a column vector of true ability scores for N = 1,000 simulated examinees was generated from a standard normal distribution. In the second stage, simulated tests were created whose characteristics varied in terms of test length (n items), item difficulty (parameter b), and item discrimination (parameter a). For each test, a column vector of n values was generated to represent the item difficulty values of the test items, and another column vector of n values was generated to represent the item discrimination values. In the third stage, the response probabilities of the N examinees to the n items of the tests were calculated based on the two-parameter IRT model. Thus, an $N \times n$ matrix of response probabilities was obtained from Equation 2.

In the fourth stage of our data generation, the $N \times n$ matrix of response probabilities was translated into an $N \times n$ matrix of discrete item responses (i.e., 1 or 0). This was done by comparing each response probability with a random number drawn from a uniform distribution of values ranging from 0 to 1.0. To illustrate, consider an example in which the calculated probability of an examinee's passing an item was .68 and a number generated randomly from a uniform distribution was .43. Because the random number is less than the response probability, an item response of 1 would be assigned for that examinee on that item. On the other hand, had the random number been greater than the response probability, an item response of 0 would have been recorded. This is a standard item response generation method as used by, for example, Harwell, Stone, Hsu, and Kirisci (1996).

For each experimental condition in our study, a simulated test was created and used to generate two different $N \times n$ response data sets. The specific characteristics of the simulated tests that we varied were (a) the number of simu-

lated items in the test: $n = 20$, 40, and 60; (b) the distributions of the true item difficulty values, all uniform from –2.0 to 2.0, –0.5 to 0.5, –2.0 to 1.0, –1.0 to 2.0, and –1.0 to 1.0; and (c) the distributions of the true item discrimination values, all uniform from 1.0 to 2.0, 0.5 to 2.5, and fixed at 1.0 for all items (the latter reduces the two-parameter IRT model to the simpler one-parameter IRT model).

Note two aspects regarding the distributional characteristics of our simulated test items. First, the scaling of the item *p* values is consistent with their presumed underlying distributions in the IRT models. Second, those specific values were chosen to reflect previous simulation studies that have used Monte Carlo techniques to investigate the performance of IRT models (e.g., Hambleton, Jones, & Rogers, 1993; Maranon, Garcia, & Costas, 1997; Park & Lautenshlager, 1990; Veerkamp & Berger, 1997).

Item and Person Statistics

IRT- and CTT-based item and person statistics were estimated for each response data set generated in this study. Results were assessed for comparability, invariance, and accuracy. This process was completed 100 times for each condition of the study. Item difficulty was measured in the CTT framework as the proportion of examinees responding successfully (e.g., 1) to each item. A high item difficulty index indicates an item for which a larger proportion of examinees responded correctly. As such, high difficulty values indicate easier items, whereas low values indicate harder items. For the IRT measurement framework, the item difficulty statistic was measured for each item as the parameter *b* (ranging from about –2 to 2) in both the one- and two-parameter IRT models. In this framework, high parameter values indicate difficult items, whereas low parameter values indicate simple items.

The second item statistic based on the CTT, item discrimination, was measured for each item as the Pearson product–moment correlation (r_{it}) between the participants' item responses (i.e., 0 or 1) and total test scores. For the IRT framework, the discrimination statistic was measured as the parameter *a* (ranging from 0 to about 3.0) in both the one- and two-parameter IRT models. For both measurement frameworks, high item discrimination values indicate items that can effectively differentiate examinees possessing varying levels of the trait.

The person statistic, trait level, derived by CTT was estimated as the sum of each examinee's responses to all of the test items (i.e., total test score *T*). For the IRT framework, that person statistic was measured as the person parameter θ (theoretically ranging from about –3 to 3), based on both the one- and two-parameter IRT models. For both measurement frameworks, an examinee who responds successfully to most of the test items will obtain a

high person statistic T or θ, indicating high ability. Conversely, a low person statistic would indicate an examinee possessing low ability.

Estimates of item and person statistics for the IRT measurement framework were obtained with the microcomputer software package PARSCALE (Muraki & Bock, 1997). PARSCALE obtains parameter estimates using the marginal maximum likelihood (MML) method and has a number of program options that can affect the parameterization of the item and person statistics. To keep the analyses simple, program defaults were used for this study, with the exception that we increased the number of expectation and maximization (EM) cycles from 5 to 40. This was done to increase the likelihood that PARSCALE would meet the convergence criterion and not stop prior to obtaining a stable solution. For the one- and two-parameter logistic models, PARSCALE sets (a) the 30 quadrature nodes from –4.0 to 4.0; (b) the convergence criterion for the item parameters at .001; and (c) the person parameter θ to be estimated from a standard normal distribution, $N(0,1)$.

Comparability of IRT and CTT Item and Person Statistics

The comparability of the item and person statistics in this study were assessed as correlations between CTT-based estimates and their corresponding IRT-based estimates obtained from the same sample of simulated examinees. For the item difficulty statistics, the CTT-obtained item difficulty p value was correlated with the IRT-based difficulty parameter b. For the item discrimination statistics, the CTT-obtained item discrimination index r_{it} was correlated with the IRT-based discrimination parameter a. The comparability of the person statistic was obtained by correlating the CTT-based person test score T and the IRT-based person parameter θ.

Invariance of IRT and CTT Item Statistics

To assess the invariance of the item difficulty statistic for CTT, item p values from two independent samples of simulated examinees responding to the same test were correlated. Similarly, for the IRT framework, the item difficulty parameter b estimates obtained from the two samples were correlated. A high invariance (i.e., correlation) coefficient would indicate that the item difficulty statistics based on two different samples of examinees yielded similar patterns of values. That is, items estimated to have high and low difficulty levels in one sample would be so estimated in the other sample.

The invariance of the item discrimination statistics was assessed following the same procedure described for measuring the invariance of item difficulty statistics using two different examinee samples. For the IRT frame-

work, the item discrimination parameter *a* estimates obtained from one sample of examinees were correlated with the *a* estimates obtained from another independent sample of examinees. Similarly, for the CTT framework, the item discrimination indices r_{it} obtained from the two samples were correlated. High correlations would suggest that item discrimination statistics assessed are sample invariant.

Accuracy of IRT and CTT Item and Person Statistics

The principal advantage of our Monte Carlo procedure, in which simulated test items and simulated examinees were generated, is that unlike in other most studies in the past (cf. Fan, 1998), item parameters and person characteristics could be explicitly controlled. Item characteristics and person abilities, therefore, were known values. In consequence, it was possible for us to calculate the accuracy of estimated item and person statistics (obtained with both IRT and CTT) with respect to the "true" item and person statistics.

To assess the accuracy of item and person statistics based on the two measurement frameworks, correlations were calculated between estimated and known values. For item difficulty, the true values were correlated with both the IRT-based item difficulty parameter *b* and the CTT-based item *p* value. For item discrimination, true values were correlated with both the IRT-based item discrimination parameter *a* and the CTT-based item index r_{it}. For the person statistic, the known values were correlated with both the IRT-based person parameter θ and the CTT-based person test score *T*. These calculations measuring the accuracy of item and person statistics were performed for each sample of simulated examinees.

Results

The results of our Monte Carlo investigation assessing the comparability, invariance, and accuracy of item and person statistics derived by IRT and CTT are presented in Tables 1 through 8. In the sections that follow, the data are summarized and interpreted in the context of their implications for test construction and item selection under the IRT and CTT frameworks.

Comparability of IRT and CTT Item and Person Statistics

The results of the computer simulations assessing comparability of person statistics obtained from IRT and CTT frameworks are summarized in Table 1. Entries in this table were derived following the procedure described by Fan (1998), involving (a) obtaining IRT- and CTT-based person ability estimates

Table 1
Comparability of Person Statistic: Average Correlations Between IRT and CTT Person Trait Estimates

		Item Discrimination		
Test Length	Item Difficulty	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.992 (.005)	.988 (.005)	.994 (.004)
	–2.0 to 1.0	.988 (.006)	.984 (.005)	.988 (.007)
	–1.0 to 2.0	.988 (.006)	.984 (.008)	.988 (.007)
	–1.0 to 1.0	.986 (.003)	.984 (.002)	.989 (.002)
	–0.5 to 0.5	.981 (.001)	.980 (.002)	.990 (.001)
40	–2.0 to 2.0	.994 (.003)	.991 (.003)	.993 (.002)
	–2.0 to 1.0	.984 (.005)	.983 (.006)	.986 (.005)
	–1.0 to 2.0	.984 (.005)	.984 (.004)	.987 (.005)
	–1.0 to 1.0	.982 (.002)	.981 (.003)	.987 (.001)
	–0.5 to 0.5	.973 (.002)	.974 (.002)	.983 (.001)
60	–2.0 to 2.0	.995 (.003)	.993 (.002)	.994 (.003)
	–2.0 to 1.0	.983 (.005)	.983 (.006)	.986 (.004)
	–1.0 to 2.0	.983 (.004)	.983 (.006)	.985 (.005)
	–1.0 to 1.0	.980 (.003)	.980 (.003)	.986 (.002)
	–0.5 to 0.5	.970 (.002)	.971 (.002)	.981 (.001)

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 200 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

from the response data set for each sample of simulated examinees, (b) correlating the IRT and CTT person ability estimates, and (c) averaging the correlations for all samples within the same experimental conditions. Each entry in the table represents the average of 200 correlation coefficients obtained from the responses of two random samples of simulated examinees ($N = 1{,}000$) to 100 simulated tests (all average correlation coefficients reported in this study were calculated using Fisher *Z* transformations).

The results in Table 1 indicate that IRT and CTT person statistics reflect very high comparability coefficients. In fact, across all experimental conditions, the obtained average correlations between IRT-based parameter θ estimates and CTT-based person test score *T* values were no less than .970 and were as high as .995, with an overall average correlation of .985. These very high correlations indicate that regardless of the measurement framework, decisions about levels of attributes in examinees grounded in either IRT- or CTT-based person statistics will not much differ.

The results in Table 2 show that the IRT-based item difficulty parameter *b* and the CTT-based item *p* value demonstrated very high comparability, with an overall average correlation of .964. Across all test sizes and distributions of true item difficulty values, the highest levels of statistical agreement

Table 2
Comparability of Item Statistics: Average Correlations Between IRT and CTT Item Difficulty Estimates

		Item Discrimination		
Test Length	Item Difficulty	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.968 (.016)	.957 (.017)	.989 (.006)
	–2.0 to 1.0	.956 (.027)	.939 (.025)	.989 (.009)
	–1.0 to 2.0	.954 (.023)	.942 (.026)	.989 (.007)
	–1.0 to 1.0	.974 (.015)	.945 (.017)	.997 (.002)
	–0.5 to 0.5	.961 (.018)	.941 (.022)	.998 (.001)
40	–2.0 to 2.0	.968 (.009)	.947 (.013)	.988 (.004)
	–2.0 to 1.0	.952 (.020)	.929 (.018)	.988 (.006)
	–1.0 to 2.0	.953 (.021)	.929 (.020)	.988 (.005)
	–1.0 to 1.0	.975 (.009)	.933 (.016)	.996 (.001)
	–0.5 to 0.5	.961 (.011)	.935 (.018)	.997 (.001)
60	–2.0 to 2.0	.969 (.006)	.946 (.010)	.989 (.003)
	–2.0 to 1.0	.953 (.017)	.926 (.016)	.988 (.004)
	–1.0 to 2.0	.952 (.016)	.925 (.016)	.988 (.005)
	–1.0 to 1.0	.974 (.006)	.931 (.013)	.996 (.006)
	–0.5 to 0.5	.963 (.008)	.930 (.016)	.997 (.001)

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 200 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

occurred when the true item discrimination values were fixed at unity (which represents the one-parameter IRT model). Under conditions of the two-parameter IRT model, the obtained correlations did vary according to the distribution of the true item discrimination values, whereby higher correlations were found in the 1.0 to 2.0 condition than in the 0.5 to 2.5 condition. However, in both conditions, the obtained correlations were still quite strong.

Table 3 presents the results of assessing the comparability of the IRT item discrimination parameter a and CTT item discrimination r_{it} index. It is in this comparison that substantial differences between the two methods occurred. Inspection of the average correlations reveals that item discrimination and difficulty distributions of the simulated items significantly impacted the agreement between the item discrimination estimated using the two measurement frameworks. Higher rates of measurement comparability were found as the simulated item discrimination values had a wider range, and higher comparability rates were found as the item difficulty values had a narrower range. In fact, IRT and CTT item discrimination estimates reached acceptable rates of agreement only when the item difficulty values had the narrowest distribution of values (–0.5 to 0.5, $M = .951$). In all other conditions, the obtained

Table 3
Comparability of Item Statistics: Average Correlations Between IRT and CTT Item Discrimination Estimates

		Item Discrimination		
Test Length	Item Difficulty	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.235 (.217)	.537 (.204)	NA
	–2.0 to 1.0	.417 (.215)	.667 (.151)	NA
	–1.0 to 2.0	.411 (.216)	.648 (.176)	NA
	–1.0 to 1.0	.714 (.122)	.880 (.059)	NA
	–0.5 to 0.5	.952 (.019)	.949 (.017)	NA
40	–2.0 to 2.0	.245 (.164)	.533 (.135)	NA
	–2.0 to 1.0	.396 (.155)	.657 (.119)	NA
	–1.0 to 2.0	.387 (.153)	.658 (.123)	NA
	–1.0 to 1.0	.731 (.081)	.877 (.039)	NA
	–0.5 to 0.5	.956 (.013)	.946 (.013)	NA
60	–2.0 to 2.0	.269 (.119)	.515 (.108)	NA
	–2.0 to 1.0	.357 (.132)	.642 (.094)	NA
	–1.0 to 2.0	.378 (.125)	.649 (.107)	NA
	–1.0 to 1.0	.730 (.065)	.873 (.033)	NA
	–0.5 to 0.5	.956 (.010)	.944 (.010)	NA

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 200 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

comparability correlations were moderate to low (overall average correlation of .676).

To summarize the results thus far, the comparability of IRT- and CTT-based item and person statistics was very high for item difficulty estimates and person ability estimates. For these two statistics, decisions about test items or test respondents based on their information would largely agree, regardless of the method of analysis. However, this high level of comparability was not the case with regard to item discrimination statistics. The simulations revealed that the comparability of IRT and CTT item discrimination estimates varied greatly depending on the underlying characteristics of the test items. With the exception of a test containing items with both a wide range of discrimination values and a narrow range of difficulty values, any expectations of high item discrimination comparability between IRT and CTT may be unfounded. Thus, the two methods might, under these circumstances, lead to mostly different items being selected in a test construction project, for example. Note, however, that a lack of comparability in the item discrimination statistic does not inform us which measurement framework, IRT or CTT, provides the more stable or accurate estimates of item characteristics. These issues are addressed in the following sections.

Invariance of IRT and CTT Item Statistics

Tables 4 and 5 present our simulation results assessing the invariance properties of IRT and CTT measurement frameworks. Entries in these two tables are average correlations (using Fisher *Z* transformations) for item difficulty estimates (Table 4) and item discrimination estimates (Table 5) derived from the *same* measurement framework on 100 simulated tests. Each test had two different samples of simulated examinees responding, for a total of 200 random samples of examinees for each condition. IRT item parameter estimates from different samples of examinees (Sample 1 vs. Sample 2) were correlated to measure invariance of the IRT-based item statistics. CTT item statistics obtained from different samples were similarly compared.

The results in Table 4 indicate that IRT and CTT item difficulty estimates from different samples of examinees were highly invariant. In particular, item difficulty estimates based on the CTT measurement framework demonstrated a remarkably high degree of invariance, averaging .994 overall. That is, the CTT-based item difficulty *p* values obtained from two different samples of examinees responding to the same test correlated almost perfectly across all conditions. Similarly, item difficulty parameter *b* estimates based on the IRT measurement framework yielded invariance correlations almost as high in value, with an overall average of .972. For both measurement frameworks, the invariance of the item difficulty statistic appears remarkably high regardless of the number of items in the test, their range of difficulty levels, or their range of discrimination values.

Table 5 presents the results of the simulations assessing the invariance of IRT and CTT item discrimination statistics. The overall pattern within the table reveals that the CTT-based item discrimination r_{it} index demonstrated higher rates of statistical invariance than the IRT-based item discrimination parameter *a* (i.e., .954 vs. .902). That is, across all levels of test length, true item difficulty levels, and true item discrimination levels, the CTT-based item discrimination estimates from two different samples of examinees obtained higher correlations than did IRT-based estimates.

It is apparent in Table 5 that two patterns can be found within the estimates of item discrimination invariance. First, it can be seen that obtained correlations were higher when the true item discrimination values were generated from the wider 0.5 to 2.5 distribution as compared to the narrower 1.0 to 2.0 distribution. This was particularly observable for the IRT-based parameter estimates but was still noticeable for the CTT-based item discrimination statistic.

The second pattern to the entries of Table 5 can be observed in the relation between the invariance correlations and the true item difficulty values. For the IRT-based estimates, the obtained correlations were highest (M = .927) when the true item difficulty values were from the narrow distribution (–0.5 to 0.5) and lowest (M = .876) when item difficulty values were from the

Table 4
Invariance of Item Statistics: Average Correlations Between IRT and CTT Item Difficulty Estimates

		IRT Item Discrimination			CTT Item Discrimination		
Test Length	Item Difficulty	1.0 to 2.0	0.5 to 2.5	1	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.983 (.023)	.979 (.024)	.991 (.006)	.999 (.001)	.999 (.001)	.998 (.001)
	–2.0 to 1.0	.970 (.049)	.972 (.039)	.987 (.018)	.998 (.001)	.998 (.001)	.997 (.001)
	–1.0 to 2.0	.967 (.052)	.969 (.040)	.984 (.012)	.998 (.001)	.998 (.001)	.997 (.002)
	–1.0 to 1.0	.979 (.026)	.972 (.027)	.989 (.005)	.997 (.002)	.996 (.002)	.995 (.002)
	–0.5 to 0.5	.945 (.043)	.952 (.031)	.971 (.011)	.987 (.006)	.984 (.007)	.978 (.010)
40	–2.0 to 2.0	.986 (.010)	.980 (.013)	.990 (.004)	.999 (.001)	.999 (.001)	.998 (.001)
	–2.0 to 1.0	.957 (.043)	.965 (.029)	.984 (.007)	.998 (.001)	.998 (.001)	.997 (.001)
	–1.0 to 2.0	.961 (.041)	.962 (.031)	.985 (.006)	.998 (.001)	.998 (.001)	.997 (.001)
	–1.0 to 1.0	.983 (.014)	.970 (.018)	.988 (.004)	.996 (.001)	.996 (.001)	.994 (.002)
	–0.5 to 0.5	.947 (.028)	.946 (.023)	.972 (.008)	.986 (.004)	.985 (.005)	.977 (.006)
60	–2.0 to 2.0	.988 (.007)	.982 (.012)	.990 (.003)	.999 (.001)	.999 (.001)	.998 (.001)
	–2.0 to 1.0	.961 (.034)	.956 (.025)	.985 (.005)	.998 (.001)	.998 (.001)	.997 (.001)
	–1.0 to 2.0	.955 (.036)	.956 (.027)	.984 (.005)	.998 (.001)	.998 (.001)	.997 (.001)
	–1.0 to 1.0	.984 (.010)	.969 (.014)	.987 (.003)	.996 (.001)	.996 (.001)	.994 (.002)
	–0.5 to 0.5	.952 (.019)	.941 (.019)	.970 (.007)	.986 (.003)	.984 (.004)	.977 (.006)

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 100 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

Table 5
Invariance of Item Statistics: Average Correlations Between IRT and CTT Item Discrimination Estimates

Test Length	Item Difficulty	IRT Item Discrimination			CTT Item Discrimination		
		1.0 to 2.0	0.5 to 2.5	1	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.806 (.098)	.919 (.095)	NA	.964 (.022)	.967 (.032)	NA
	–2.0 to 1.0	.831 (.065)	.942 (.036)	NA	.954 (.024)	.970 (.018)	NA
	–1.0 to 2.0	.830 (.079)	.938 (.034)	NA	.951 (.044)	.969 (.021)	NA
	–1.0 to 1.0	.873 (.060)	.959 (.024)	NA	.915 (.041)	.968 (.018)	NA
	–0.5 to 0.5	.883 (.050)	.964 (.018)	NA	.892 (.050)	.976 (.011)	NA
40	–2.0 to 2.0	.831 (.040)	.932 (.040)	NA	.966 (.017)	.970 (.019)	NA
	–2.0 to 1.0	.848 (.043)	.945 (.083)	NA	.956 (.023)	.972 (.014)	NA
	–1.0 to 2.0	.849 (.054)	.945 (.023)	NA	.958 (.019)	.970 (.016)	NA
	–1.0 to 1.0	.887 (.031)	.962 (.012)	NA	.918 (.028)	.973 (.009)	NA
	–0.5 to 0.5	.894 (.030)	.965 (.012)	NA	.895 (.029)	.975 (.008)	NA
60	–2.0 to 2.0	.833 (.042)	.936 (.029)	NA	.966 (.015)	.971 (.014)	NA
	–2.0 to 1.0	.851 (.041)	.941 (.109)	NA	.956 (.018)	.970 (.011)	NA
	–1.0 to 2.0	.845 (.035)	.940 (.069)	NA	.956 (.015)	.969 (.012)	NA
	–1.0 to 1.0	.884 (.027)	.961 (.010)	NA	.915 (.024)	.970 (.009)	NA
	–0.5 to 0.5	.892 (.024)	.966 (.009)	NA	.895 (.022)	.975 (.007)	NA

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 100 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

widest distribution (–2.0 to 2.0). For the CTT-based item discrimination estimates, on the other hand, this pattern was reversed. Highest correlations were obtained with the widest distribution of true difficulty values ($M = .967$), and the lowest correlations were obtained with the narrowest distribution of difficulty values ($M = .935$).

Differences in the item discrimination statistics for the two measurement models raise the question of which one is correct. To explore this question, the accuracy of item and person statistics was investigated to determine if the IRT- and CTT-based statistics also revealed different levels of agreement with the true item parameters.

Accuracy of IRT and CTT Item and Person Statistics

A principal advantage of computer investigations using simulated items and persons is the ability to manipulate systematically factors that would normally be inaccessible in real data sets. Simulated tests in this study, for example, were manipulated to vary in terms of length, item difficulty values, and item discrimination values. Because the characteristics of the simulated

items and persons were known to us, we were then able to evaluate the accuracy of item and person estimates based on the two measurement frameworks.

Tables 6, 7, and 8 present the simulation results assessing the accuracy of IRT- and CTT-based estimates of examinees' trait levels, test item difficulties, and test item discrimination values, respectively. Entries in these tables are the average correlations between statistics based on for IRT and CTT frameworks. Each correlation is based on 200 samples of simulated examinees (N = 1,000) responding to 100 simulated tests (two samples of examinees per test).

The results in Table 6 indicate that IRT- and CTT-based person statistics accurately estimate the true abilities of the simulated examinees. Across all levels of item difficulty values and item discrimination values, the IRT person parameter θ and CTT person test score T were highly correlated with true values (Ms = .965 and .952, respectively). Those results suggest that regardless of the measurement framework, test-based decisions regarding person ability estimates will be consistent and accurate.

The accuracy of IRT and CTT item difficulty statistics are presented in Table 7. Under the CTT measurement framework, very high correlations were found between item difficulty p values and true item difficulty values (M = .991). The highest correlations were obtained when true item discrimination values were fixed at 1.0 (M = .993), followed by the 1.0 to 2.0 distribution (M = .993), and the 0.5 to 2.5 distribution (M = .985). Under the IRT measurement framework, high correlations were also found between item difficulty parameter b values and true item difficulty values. Highest correlations were found when true item discrimination values were fixed at 1.0 (M = .991), followed by the 1.0 to 2.0 distribution (M = .972), and the 0.5 to 2.5 distribution (M = .958). Higher correlations were also found for the IRT framework when the true item difficulty distribution was –2.0 to 2.0 (M = .984), followed by –1.0 to 1.0 (M = .979), –2.0 to 1.0 (M = .972), –1.0 to 2.0 (M = .971), and finally –0.5 to 0.5 (M = .964). These results indicate that IRT- and CTT-based item difficulty estimates were slightly negatively affected by the range of item difficulty and item discrimination values of the test items. The overall accuracy estimates of item difficulty statistics, however, remained high to very high for both measurement frameworks.

Table 8 presents the results of the simulations assessing the accuracy of IRT and CTT item discrimination statistics. Recall that these estimates were previously found (Table 3) to have substantial differences for the two models. It is readily apparent from this table that the accuracy of item discrimination estimates is dependent on the measurement framework. Across all simulated conditions, IRT item discrimination estimates obtained higher correlations than CTT item discrimination estimates (Ms = .949 vs. .618). Furthermore, the differences in accuracy estimates ranged from slight in some conditions

Table 6
Accuracy of the Person Statistic: Average Correlations Between the True Person Parameter and Estimates Based on IRT and CTT

		IRT Item Discrimination			CTT Item Discrimination		
Test Length	Item Difficulty	1.0 to 2.0	0.5 to 2.5	1	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.954 (.004)	.952 (.007)	.932 (.005)	.947 (.007)	.941 (.009)	.930 (.005)
	–2.0 to 1.0	.952 (.006)	.950 (.007)	.933 (.007)	.940 (.011)	.935 (.011)	.926 (.009)
	–1.0 to 2.0	.952 (.006)	.950 (.009)	.932 (.006)	.940 (.011)	.935 (.015)	.926 (.009)
	–1.0 to 1.0	.954 (.004)	.953 (.005)	.939 (.003)	.941 (.005)	.938 (.006)	.933 (.004)
	–0.5 to 0.5	.948 (.003)	.946 (.003)	.939 (.003)	.930 (.004)	.928 (.004)	.930 (.004)
40	–2.0 to 2.0	.976 (.002)	.976 (.003)	.964 (.002)	.971 (.004)	.967 (.005)	.961 (.003)
	–2.0 to 1.0	.973 (.003)	.972 (.004)	.963 (.003)	.958 (.008)	.955 (.009)	.954 (.005)
	–1.0 to 2.0	.973 (.003)	.973 (.003)	.963 (.003)	.958 (.008)	.957 (.007)	.954 (.006)
	–1.0 to 1.0	.974 (.002)	.973 (.002)	.967 (.002)	.957 (.004)	.955 (.004)	.957 (.003)
	–0.5 to 0.5	.969 (.002)	.968 (.002)	.966 (.002)	.943 (.004)	.943 (.004)	.953 (.003)
60	–2.0 to 2.0	.984 (.001)	.984 (.001)	.976 (.002)	.979 (.003)	.977 (.003)	.972 (.002)
	–2.0 to 1.0	.981 (.002)	.981 (.002)	.975 (.002)	.965 (.007)	.965 (.006)	.964 (.004)
	–1.0 to 2.0	.981 (.002)	.981 (.002)	.975 (.002)	.965 (.006)	.964 (.007)	.964 (.005)
	–1.0 to 1.0	.981 (.001)	.981 (.001)	.977 (.001)	.962 (.004)	.962 (.004)	.965 (.003)
	–0.5 to 0.5	.977 (.002)	.977 (.002)	.976 (.002)	.948 (.004)	.948 (.004)	.960 (.003)

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 100 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

Table 7
Accuracy of Item Statistics: Average Correlations Between the True Difficulty Parameter and Estimates Based on IRT and CTT

		IRT Item Discrimination			CTT Item Discrimination		
Test Length	Item Difficulty	1.0 to 2.0	0.5 to 2.5	1	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.980 (.017)	.975 (.013)	.994 (.004)	.994 (.002)	.990 (.004)	.996 (.001)
	–2.0 to 1.0	.969 (.036)	.960 (.027)	.992 (.008)	.992 (.003)	.984 (.009)	.984 (.002)
	–1.0 to 2.0	.967 (.030)	.960 (.030)	.992 (.006)	.992 (.003)	.985 (.010)	.994 (.002)
	–1.0 to 1.0	.979 (.018)	.962 (.015)	.994 (.003)	.997 (.001)	.990 (.006)	.997 (.001)
	–0.5 to 0.5	.961 (.021)	.944 (.018)	.986 (.006)	.992 (.004)	.983 (.008)	.989 (.005)
40	–2.0 to 2.0	.983 (.009)	.975 (.009)	.993 (.003)	.994 (.001)	.989 (.003)	.996 (.001)
	–2.0 to 1.0	.964 (.026)	.955 (.021)	.992 (.004)	.991 (.003)	.983 (.006)	.994 (.002)
	–1.0 to 2.0	.965 (.027)	.953 (.018)	.992 (.003)	.991 (.002)	.984 (.006)	.994 (.002)
	–1.0 to 1.0	.981 (.009)	.961 (.010)	.994 (.002)	.996 (.001)	.987 (.004)	.997 (.001)
	–0.5 to 0.5	.963 (.013)	.945 (.013)	.986 (.004)	.991 (.002)	.982 (.006)	.988 (.003)
60	–2.0 to 2.0	.984 (.005)	.975 (.007)	.993 (.002)	.994 (.001)	.989 (.003)	.996 (.001)
	–2.0 to 1.0	.967 (.022)	.953 (.015)	.992 (.003)	.991 (.002)	.983 (.005)	.994 (.001)
	–1.0 to 2.0	.965 (.020)	.952 (.015)	.991 (.004)	.991 (.002)	.984 (.004)	.994 (.002)
	–1.0 to 1.0	.982 (.006)	.961 (.008)	.994 (.001)	.996 (.001)	.988 (.004)	.997 (.001)
	–0.5 to 0.5	.964 (.010)	.942 (.012)	.985 (.003)	.991 (.002)	.981 (.005)	.988 (.003)

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 100 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

Table 8
Accuracy of Item Statistics: Average Correlations Between the True Discrimination Parameter and Estimates Based on IRT and CTT

Test Length	Item Difficulty	IRT Item Discrimination			CTT Item Discrimination		
		1.0 to 2.0	0.5 to 2.5	1	1.0 to 2.0	0.5 to 2.5	1
20	–2.0 to 2.0	.901 (.052)	.961 (.061)	NA	.202 (.206)	.521 (.186)	NA
	–2.0 to 1.0	.916 (.037)	.970 (.017)	NA	.388 (.208)	.660 (.148)	NA
	–1.0 to 2.0	.911 (.044)	.969 (.019)	NA	.371 (.209)	.646 (.169)	NA
	–1.0 to 1.0	.933 (.033)	.978 (.013)	NA	.664 (.127)	.874 (.061)	NA
	–0.5 to 0.5	.941 (.026)	.982 (.010)	NA	.904 (.040)	.944 (.019)	NA
40	–2.0 to 2.0	.911 (.029)	.967 (.020)	NA	.211 (.165)	.528 (.132)	NA
	–2.0 to 1.0	.921 (.025)	.971 (.058)	NA	.363 (.153)	.660 (.106)	NA
	–1.0 to 2.0	.921 (.028)	.972 (.013)	NA	.359 (.135)	.654 (.110)	NA
	–1.0 to 1.0	.941 (.017)	.981 (.006)	NA	.685 (.089)	.871 (.041)	NA
	–0.5 to 0.5	.945 (.015)	.982 (.006)	NA	.908 (.025)	.940 (.013)	NA
60	–2.0 to 2.0	.910 (.022)	.968 (.015)	NA	.243 (.118)	.517 (.095)	NA
	–2.0 to 1.0	.922 (.023)	.971 (.081)	NA	.347 (.122)	.651 (.080)	NA
	–1.0 to 2.0	.919 (.019)	.971 (.052)	NA	.368 (.111)	.653 (.088)	NA
	–1.0 to 1.0	.941 (.014)	.980 (.006)	NA	.687 (.068)	.867 (.033)	NA
	–0.5 to 0.5	.944 (.013)	.983 (.005)	NA	.910 (.019)	.938 (.011)	NA

Note. IRT = item response theory; CTT = classical test theory. Each entry is based on the average of 100 correlations computed over 1,000 examinees. Average correlation coefficients were obtained through Fisher *Z* transformations. Standard deviations of the raw correlations appear in parentheses.

(e.g., when true item difficulty values ranged from –0.5 to 0.5, *M*s = .963 vs. .924) to very large in others (e.g., when true item difficulty values ranged from –2.0 to 2.0, *M*s = .936 vs. .370). For both measurement frameworks, high accuracy correlations were obtained when both the range of true item discrimination values was widest (i.e., 0.5 to 2.5) and when the distribution of true item difficulty values was narrowest (i.e., –0.5 to 0.5, with *M*s of .982 and .941, respectively).

To summarize, the results of the computer simulations assessing the accuracy of IRT- and CTT-based item and person statistics has provided important insight into the two measurement frameworks. First, regarding estimates of the ability levels of examinees, the IRT-based person parameter θ and CTT-based person test score *T* were both highly accurate. Second, regarding estimates of test item difficulty, the IRT-based item difficulty parameter *b* estimates and CTT-based item difficulty *p* values were also both highly accurate. But third, regarding estimates of test item discrimination, only the IRT-based item discrimination parameter *a* yielded highly accurate estimates across all conditions of this study, whereas the CTT-based item discrimination r_{it} index yielded high accuracy estimates only under certain experimental conditions. In some cases, CTT-based estimates of item discrimination were surprisingly inaccurate.

Summary and Conclusions

This Monte Carlo investigation examined the behavior of item and person statistics obtained from IRT and CTT measurement frameworks. The study focused on three main issues: (a) How comparable are the item and person statistics generated by IRT and CTT methods? (b) How invariant are the item statistics of IRT and CTT across examinee samples? and (c) How accurate are the item and person statistics from IRT and CTT frameworks? Simulated tests and simulated examinees were generated by computer programs that manipulated the length of the test, item difficulty values, and item discrimination values. For each experimental condition, 100 simulated tests were completed by two randomly generated samples of examinees of 1,000 respondents each. Prior to (a) summarizing our major findings and (b) discussing implications of our results as regards test construction, some important framing statements should first be emphasized.

Framing the Discussion

It is important at the outset to acknowledge at least three important differences between IRT and CTT. First, IRT attempts to locate each examinee on the correct point on an interval measurement scale, θ, and attempts to estimate scores on the latent ability variable. CTT, on the other hand, focuses on the observed scores, although the concept of true scores is invoked in an effort to evaluate the quality of the observed scores (i.e., to estimate score reliability). In theory, different tests may yield an invariant estimate of θ for a given examinee, whereas observed scores vary across test forms; and this may occur even if the CTT observed scores are perfectly correlated because correlations primarily evaluate only the constancy of rank orders of ability (or item discrimination) estimates and not whether they remain centered at given points.

Second, IRT does have the appeal that person abilities and item difficulties are scaled into comparable logit metrics so that items yielding the most information can be readily selected for given examinees (e.g., an examinee of ability $\theta = 1.5$ ideally would be given items with difficulties of about 1.5). Third, in practice, people invoking IRT models typically censor the data for persons who have too many responses that are unexpected under a given model (e.g., a bright person misses a number of the easiest items) and for items whose response patterns are aberrant (e.g., items that several of the most able examinees miss). Of course, in CTT, the same sort of data editing could also be done.

Yet notwithstanding these differences, it is certainly possible that both theories may still lead to identical decisions as regards item selection and the characterization of the quality of test scores. IRT parameters are not necessar-

ily magical just because they are expressed in the less familiar metric of logits or because their mathematics are more complex.

Major Findings

Three major findings of this study stand out. First, the item difficulty and person statistics from IRT and CTT frameworks were highly comparable in all conditions. However, item discrimination statistics were comparable in only some conditions. Second, the item difficulty and item discrimination statistics were highly invariant between random samples of examinees when data were evaluated from the IRT framework. For the CTT framework, item difficulty and item discrimination statistics were even more consistent across samples, yielding higher invariance estimates. Third, under both the IRT and CTT measurement frameworks, item difficulty and person statistics were highly accurate across all conditions. However, only the IRT-based item discrimination statistic accurately estimated true discrimination values across all conditions. The CTT-based item discrimination statistic was only accurate under certain test conditions.

Under the conditions investigated in this study, our findings generally support the person-invariant item statistics property of the IRT measurement framework. More important, these findings demonstrate that the IRT framework accurately estimates item and person statistics across a wide variety of simulated testing conditions. Similarly, the simulation results demonstrated that CTT-based item statistics were also person-invariant. However, only the item difficulty and person statistics of the CTT framework were shown to estimate accurately true parameter values. The CTT-based item discrimination statistic yielded accuracy estimates that were high in some testing conditions but only moderate to low in other conditions. These findings raise interesting questions regarding the differences between IRT and CTT measurement frameworks and the impact on test construction efforts.

Implications for Test Construction

Standard test construction techniques for the development of achievement, aptitude, interest, and personality measures using CTT generally involve the selection of test items according to their content and statistical characteristics. The statistics usually include indices of item difficulty p values and item discrimination r_{it} values (Hambleton & Swaminathan, 1985; Hambleton et al., 1991). Whether the test constructor selects the best set of items from a larger item pool depends, of course, on the accuracy of these two item statistics. If either, or both, of these item statistics are not accurate, the possibility exists that some good items will fail to be selected for the final test form and some poor items will fail to be rejected.

When the collection of potential test items in a pool possesses a narrow range of item difficulty values (common in personality and interest assessments), then item discrimination estimates should be largely accurate for both IRT and CTT measurement frameworks. In such a situation, item selection decisions based on either framework should result in the selection of roughly the same set of test items. On the other hand, if the range of item difficulty statistics exceeds a narrow range of item difficulty values (about –0.5 to 0.5, common in achievement and ability tests), then the accuracy of item discrimination estimates begins to decrease with CTT methods. In the worst case scenario, if the pool of potential items possesses a very wide range of item difficulty values, unacceptably low accuracy of item discrimination estimates under the CTT framework may result. In consequence, some item selection decisions may be erroneous in the sense that the final set of selected test items may not be optimal.

In contrast to the potential problems associated with item selection under the CTT framework, decisions regarding item selection under IRT models are less influenced by vagaries in the properties of the item pool. In fact, the IRT-based item statistics maintained high levels of accuracy across all experimental conditions in this study. This finding suggests that a test constructor's item selection decisions based on item difficulty and discrimination estimates are more likely to result in the best possible subset of test items with IRT methods. And, as alluded to above, this consideration might be most relevant to the domain of aptitude and abilities measurement, where a wide range of item difficulties is typically regarded as desirable.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

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 Statistics*, *25*, 31-45.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*, 357-381.

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*, 38-47.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, *30*, 143-155.

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.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101-125.

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. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*, 517-548.

Maranon, P. P., Garcia, M. I. B., & Costas, C. S. L. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement*, *57*, 559-568.

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.

Muraki, E., & Bock, R. D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data [Computer program]. Chicago: Scientific Software International.

Ndalichako, J. L., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement*, *57*, 580-589.

Nunnally, J. C., & Berstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Park, D. G., & Lautenshlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement*, *14*, 163-173.

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

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)

Tinsley, H. E., & Dawis, R. V. (1977). Test-free person measurement with the Rasch simple logistic model. *Applied Psychological Measurement*, *1*, 483-487.

van de Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van de Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*, 203-226.