Relationships Among Classical Test Theory and Item Response Theory Frameworks via Factor Analytic Models

Educational and Psychological

Measurement

2015, Vol. 75(3) 389–405

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414559071

epm.sagepub.com

Nidhi Kohli¹, Jennifer Koran², and
Lisa Henn¹

Abstract

There are well-defined theoretical differences between the classical test theory (CTT) and item response theory (IRT) frameworks. It is understood that in the CTT framework, person and item statistics are test- and sample-dependent. This is not the perception with IRT. For this reason, the IRT framework is considered to be theoretically superior to the CTT framework for the purpose of estimating person and item parameters. In previous simulation studies, IRT models were used both as generating and as fitting models. Hence, results favoring the IRT framework could be attributed to IRT being the data-generation framework. Moreover, previous studies only considered the traditional CTT framework for the comparison, yet there is considerable literature suggesting that it may be more appropriate to use CTT statistics based on an underlying normal variable (UNV) assumption. The current study relates the class of CTT-based models with the UNV assumption to that of IRT, using confirmatory factor analysis to delineate the connections. A small Monte Carlo study was carried out to assess the comparability between the item and person statistics obtained from the frameworks of IRT and CTT with UNV assumption. Results show the frameworks of IRT and CTT with UNV assumption to be quite comparable, with neither framework showing an advantage over the other.

¹University of Minnesota, Minneapolis, MN, USA

²Southern Illinois University, Carbondale, IL, USA

Corresponding Author:

Nidhi Kohli, Quantitative Methods in Education Program, Department of Educational Psychology, University of Minnesota, 161 Education Sciences Building, 56 East River Road, Minneapolis, MN 55455, USA.

Email: nkohli@umn.edu

Keywords

classical test theory, item response theory, relationship, factor analysis

Classical test theory (CTT) has been prominent in the field of educational measurement since the 1920s; however, for the last three decades, item response theory (IRT) has been the primary framework for educational measurement and psychometric issues. A commonly held belief is that the IRT framework is theoretically superior to the CTT framework for the estimation of person and item parameters because the person and item statistics based on the CTT framework are test- and sample-dependent, respectively. Specifically, the item statistics derived from the CTT-based models, item difficulty and item discrimination, are dependent on the sample of respondents selected to answer the items. If the same items are given to a different sample, and the item difficulty and item discrimination indices are computed on CTT-based models, they may vary substantially depending on the nature of the sample. Similarly, the scores earned by test takers depend on the items they have been asked to answer. If the test takers are given another set of more or less difficult items, their number-correct test scores likely are going to be lower or higher, respectively, than their number-correct scores on the original set of items.

In contrast to CTT, the person and item statistics based on IRT are considered to be stable across different samples of items and persons, respectively. As explained by Lord (1980), this perspective on the ability variable follows from viewing the item response function as a regression function of the observed test outcomes on the ability variable. The probability of observing a particular outcome likely is unaffected by how many among the test subjects has a particular level of ability. The invariance of the item parameters follows conceptually from this framing of the item response function as a regression function, with other elements in the function conceived as fixed parameters.

We have found nothing in the literature that examines this item parameter estimate stability via simulation, that is to say, that examines variation in parameter estimates at moderate sample sizes. Nonetheless, this conceptual stability property often seems attributed to parameter estimates as well.

In previous simulation studies, researchers have used IRT models both as generating and as fitting models, which confounds comparability between CTT and IRT frameworks with data-model congruity. An alternate explanation for results favoring the IRT framework is that IRT was used to generate the data, and thus, the data will more naturally fit this model. In addition, researchers have chosen traditional CTT item statistics, proportion correct and point-biserial correlations, and a traditional person-statistic, unweighted total scores, for the comparability comparison. But there is considerable literature suggesting that it may be more appropriate to use statistics based on an underlying normal variable (UNV) assumption, such as thresholds and biserial correlations.

Despite well-defined theoretical differences between the CTT and IRT frameworks, the empirical research comparing the two frameworks has failed to exhibit

differences between the two in terms of person and item parameter estimates. To explore the distinctions between the two frameworks in greater depth, we first present a review of the empirical literature comparing them. Prior simulation studies have exclusively used IRT models to generate the data, and the literature does not consider models with an UNV assumption common in factor analytic models for categorical data. However, recent literature has introduced models with an UNV assumption as an extension of the CTT framework. As both IRT models and CTT-based models with the UNV assumption have been demonstrated to be members of the class of confirmatory factor analysis models, we expect a high level of comparability between results obtained under related member models. To test this theory, we compare item and person statistics for the extended CTT framework and the IRT framework using data generated under both frameworks. That we simulate data under each framework is, to our knowledge, a unique contribution.

Literature Review

Prior studies comparing CTT and IRT frameworks have not found that differences between the two translate to advantage of one framework over another. Many works have mentioned IRT parameter invariance as part of a general introduction of the IRT model (see, e.g., Hambleton & Jones, 1993; Shirkness & DeAngelo, 2011). Rudner (1983) examined how the magnitude of an item discrimination value should change if the location of the ability variable is not the same for two groups of examinees. Cook, Eignor, and Hessay (1988) compared three administrations of a Biology achievement test in part to examine stability of IRT item parameter estimates. They found lack of stability, noting that it is affected by the advancement in skill level of the test-takers. We have found no prior literature examining IRT parameter invariance via simulation. Prior studies comparing estimates under CTT and IRT paradigms show high correlations between CTT and IRT not only for person ability but also for item difficulty (Courville, 2004; Fan, 1998; Lawson, 1991). Fan's (1998) research in particular "failed to support the IRT framework for its ostensible superiority over CTT in producing invariant item statistics" (p. 378). Item discrimination indices are less highly correlated between the two frameworks, dipping as low as 0.60, particularly when the range of difficulty parameter values exceeds 0.5 in absolute value (Fan, 1998; MacDonald & Paunonen, 2002). These correspondences are highest when the traditional CTT statistics are compared to the corresponding item statistics in one- and two-parameter logistic IRT models (Fan, 1998; MacDonald & Paunonen, 2002). All these studies use IRT as the data generation model.

Some attempts have been made to relate one framework to the other (Hambleton & Jones, 1993; Lord, 1980; Miyazaki, 2005). Lord (1980) presented approximate expressions for IRT item discrimination parameter and item difficulty parameter as functions of CTT item-biserial correlation and pass/fail threshold parameter. He called these relations "crude" and added that they "are given . . . not for practical use but rather to give an idea of the nature of the item discrimination parameter"

(pp. 33-34). Miyazaki (2005) used two-level hierarchical generalized linear models as the intermediate framework to relate these two approaches, which additionally requires a normal distributional assumption on the observed test scores and the use of the identity link. This distributional assumption is not part of the core framework of CTT, making this approach for associating the two classes of model more restrictive.

Despite the empirically demonstrated similarities between the two modeling frameworks, the results have led some researchers to nonetheless conclude that the IRT framework is superior to the CTT framework (MacDonald & Paunonen, 2002). We see two problems with the body of prior research comparing the CTT and IRT frameworks. First, prior empirical studies have exclusively used IRT to generate the data; thus, empirical results favoring IRT might be due to the design of the study. Second, prior studies comparing CTT and IRT have not considered CTT-based models or statistics with a UNV assumption. However, recent literature has presented CTT-based models with a UNV assumption as legitimate, desirable extensions of the CTT framework (Raykov & Marcoulides, 2011).

Whereas previous efforts have sought to connect IRT to CTT with approximate expressions or with hierarchical models, other research supports a relation between a class of CTT models and some IRT models using confirmatory factor analysis to delineate the associations. This approach does not require a distributional assumption on the response as with Miyazaki (2005). Where CTT observed scores are considered as a result of tests containing a single, binary item, CTT can be applied to scored responses to individual items. This is possible because CTT assumes only the existence of the mathematical expectation of the observed score not that the observed score be continuous, contrary to popular misconception (Raykov & Marcoulides, 2011). Furthermore, the item scores from several such single item tests can be assumed to fit either parallel, tau-equivalent, or congeneric¹ CTT-based models, and these CTT-based models have been demonstrated as members within the family of confirmatory factor analysis models (DeVellis, 1991; Graham, 2006; Jöreskog, 1971). An extension, CTT-based models with a UNV assumption (Ferrando, 2000; Raykov & Marcoulides, 2011), has produced models shown to be members of the family of nonlinear confirmatory factor analysis models (Raykov & Marcoulides, 2011). The two-parameter IRT model (and the one-parameter model nested within it) has likewise been shown to be mathematically equivalent to nonlinear confirmatory factor analysis model (Kamata & Bauer, 2008; McDonald, 1999; Takane & de Leeuw, 1987; Wirth & Edwards, 2007).

As both the CTT-based models with a UNV assumption and the one- and two-parameter IRT models have been demonstrated to be members of the class of nonlinear confirmatory factor analysis models, we expect a high level of similarity between results when the two frameworks are assessed under comparable conditions. We compare the two formulations using simulated data generated in each framework and compare that framework's parameter estimates to those resulting from the fit of the analogous model in the other framework. We thereby show relations between the two classes under conditions of parity.

Method

Overview

A small Monte Carlo simulation study was carried out to examine whether results reported in the literature held when the data generation model was varied (i.e., data were generated under both IRT framework and CTT with UNV assumption framework). The simulation design was composed of two conditions/factors: test length (total number of items) and number of examinees. We chose these two factors for the simulation design because we expect them to affect the magnitude of comparability between the item and person statistics arising from the CTT with UNV assumption framework and those from the IRT framework. Small sample sizes are known to affect item parameter estimates adversely, and short tests are known to affect person ability estimates adversely. The test length factor took values of 20, 40, and 60 items, whereas the number of examinees factor took values of 500 and 1,000 examinees. Thus, the combination of manipulated factors (3×2) resulted in a Monte Carlo simulation with six cells or conditions. The conditions are listed in Table 1. For each condition, 100 replications were generated.

Table 1. Simulation Conditions.

Condition	Number of items	Number of examinees
Case 1	20	500
Case 2	20	1,000
Case 3	40	500
Case 4	40	1,000
Case 5	60	500
Case 6	60	1,000

Within each manipulated condition, data sets were generated for each of four models: two models from CTT with UNV assumption, the parallel and the congeneric models, and two from IRT, the one-parameter logistic (1PL) and two-parameter logistic (2PL) models. The tau-equivalent CTT-based model with the UNV assumption was not used in the simulation study, because with the UNV assumption, the tau-equivalent model is functionally equivalent to the parallel model. This is due to the fact that constraining the loadings also results in the error variances being equal to one another. The four models used in the study are explained in detail in the following sections.

The One-Parameter Logistic Rasch Model

This model is the more restrictive of the two IRT models. All items are given the same weight in determining the level of the latent construct for an individual. This model is typically presented in its logistic form for an individual as

$$P(X_{ki} = 1 | \theta) = \frac{1}{1 + \exp[-D(\theta_i - b_k)]}, \quad (1)$$

where θ_i represents the level of the latent trait for an individual; b_k is the item difficulty parameter on a scale approximating the normal ogive scale (nonlinear factor analysis model with probit transformation), which describes how much of the latent construct an individual must possess to have a 50% probability of endorsing item k ; and $D = 1.7$ is a scaling constant that when multiplied by an item parameter approximately produces the corresponding value for the parameter on the logistic scale.

In the 1PL model, the relevant person statistic is an estimate of θ . The relevant item statistic is an estimate of the item difficulty parameter b_k .

The Two-Parameter Birnbaum Model

This model is considered to be the less restrictive model of the two IRT models. Items that are more discriminating are given greater weight in determining the level of the latent construct for an individual. This model typically is presented in its logistic form as

$$P(X_{ki} = 1 | \theta) = \frac{1}{1 + \exp[-Da_k(\theta_i - b_k)]}, \quad (2)$$

where θ_i represents the level of the latent trait for an individual; b_k is the item difficulty parameter describing how much of the latent construct an individual must possess to have a 50% probability of endorsing item k ; a_k is the item slope (discrimination) parameter on a scale approximating the normal ogive scale (nonlinear factor analysis model with probit transformation), which describes the strength of the relationship between item k and the latent trait θ ; and $D = 1.7$ is a scaling constant that when multiplied by an item parameter approximately produces the corresponding value for the parameter on the logistic scale.

In the 2PL model, the relevant person statistic is an estimate of θ . The relevant item statistics are estimates of the item difficulty parameter b_k and the item slope (discrimination) parameter a_k .

The Underlying Normal Variable Assumption

The UNV assumption is a popular approach in the latent variable modeling literature (Jöreskog, 1990; Mislevy, 1986; Muthén, 1978, 1984; Muthén & Christofferson, 1981; Takane & de Leeuw, 1987). Each observed binary item score variable X_k with two categories is assumed to be a coarse representation of an underlying unobserved continuous variable X_k^* . If X_k^* is assumed to be univariate normally distributed, then X_k^* is the UNV. A monotonic transformation matches the density of the observed binary distribution to the density of the continuous distribution:

$$X_k = \begin{cases} 0 & \text{if } X_k^* < \tau_k, \\ 1 & \text{if } X_k^* \geq \tau_k. \end{cases} \quad (3)$$

The UNV X_k^* is assumed to have a range from negative infinity to positive infinity. The notation τ_k will be used to refer to the single estimated threshold for dichotomous item k .

The Parallel Model With the UNV Assumption

Applying the UNV assumption to the parallel model concept (Ferrando, 2000), X_{ki}^* for each item k for individual i can be decomposed into true score T_i and error E_{ki} as

$$X_{ki}^* = T_i + E_{ki} \quad (4)$$

and

$$\text{var}(E_{ki}) = \text{var}(E_k),$$

where $k, k' \in \{1, \dots, p\}$ are item indices and $k \neq k'$. By definition, $E(E_{ki}) = 0$. Analogous to the parallel model without the UNV assumption, all UNV item scores are assumed to share the same true score T_i . In addition, all UNV item scores are assumed to have equal reliability (equal measurement error). There is no subscript k on the true score T_i . Because the true score T_i is assumed to be the same across items, the distinguishing subscript k is not needed.

The Congeneric Model With the UNV Assumption

Applying the UNV assumption to the congeneric model concept (Ferrando, 2000), the UNV item scores X_{ki}^* are linear functions of the same true score, and individual item error variances are not constrained to be equal. All UNV item scores X_{ki}^* have true scores that are linear functions of the same true score T_i . In equation form, the model is

$$X_{ki}^* = \lambda_k^* T_i + E_{ki}, \quad (5)$$

where λ_k^* is the unique loading for item k . By definition, $E(E_{ki}) = 0$. Note that $T_{ki} = \lambda_k^* T_i$.

In the congeneric model with the UNV assumption, the relevant person statistic is an estimate of T_i (such as a factor score). Because λ_k^* differs across items, T_i is not necessarily a linear function of the number correct score. Thus, in the congeneric model, the number correct score does not necessarily contain the same information as the true score T_i . The relevant item statistics are the item threshold τ_k and the biserial (polyserial) correlation r_{bs} . As in the parallel model with the UNV assumption, if the item score X_{ki} is the result of dichotomous scoring $\{0, 1\}$, then $1 - \text{probit}(\tau_k)$ is the proportion of individuals, where $X_{ki} = 1$ (item proportion correct π_k). Thus, in the

case of the congeneric model with the UNV assumption, the item proportion correct π_k contains the same information as the item threshold. The biserial (polyserial) correlation r_{bs} is equivalent to the standardized factor loading, which is the correlation between the UNV item score X_{ki}^* and the common factor T_i .

Data Generation

The discrimination parameter, a , for the IRT models was generated on the normal ogive scale, meaning that a scaling factor of $D=1.7$ was incorporated into the IRT model as per Equation 2. Item difficulty, b , values for the IRT models were sampled from a uniform distribution; $b \sim \text{Uni}(-2, 2)$. This is comparable to values used in the MacDonald and Paunonen (2002) study. In the 1PL model, the item discrimination, a , was fixed at 1. Item discrimination values for the 2PL were sampled from a uniform distribution; $a \sim \text{Uni}(1, 2)$. Person parameter, θ , values were drawn from the standard normal distribution. Corresponding bounds for sampling distributions for loading, λ , and threshold, τ , parameters for the CTT-based models with UNV assumption were calculated using the following equations:

$$\lambda_k^* = \frac{a_k/D}{\sqrt{1 + (a_k/D)^2}} \quad (6)$$

and

$$\tau_k = \frac{(a_k/D)b_k}{\sqrt{1 + (a_k/D)^2}} \quad (7)$$

(Wirth & Edwards, 2007).

Thus, for the parallel model with the UNV assumption, the threshold, τ , values were sampled from a uniform distribution, with $\tau \sim \text{Uni}(-1.014, 1.014)$. Factor loadings, λ , were set to 1. For the congeneric model with the UNV assumption, the threshold $\tau \sim \text{Uni}(-1, 1.52)$, while the factor loading $\lambda \sim \text{Uni}(0.5, 0.76)$. As with the IRT model, person parameter, T , values for both CTT with UNV assumption models were assumed to be standard normally distributed. Data for the models considered across all the manipulated conditions were generated by using R version 3.0.2 (R Development Core Team, 2013), with data for the CTT with UNV models generated with the R package psych, version 1.4.5 (Revelle, 2014). The IRT models were fit with the R package mirt, version 1.4 (Chalmers, 2012), using the EM algorithm, and CTT with UNV assumption models were fit with the R package lavaan, version 0.5-17.701 (Rosseel, 2012), using a weighted least square estimator with robust variance estimator, mimicking that of the WLSMV estimator in Mplus (Muthén & Muthén, 1998-2012).

Analysis

On fitting the models, the correlations of person and item estimates were calculated. Correlations were computed across pairings of two sets of statistics: the IRT statistics and the CTT with UNV assumption model statistics. The correlation of theta, θ , the person statistic obtained from the IRT models, and the factor score, T , obtained from the CTT with UNV assumption models, were calculated to assess the degree of comparability of person estimates. Correlations between the item difficulty parameter, b , obtained from the IRT models, and the item threshold, τ , obtained from the CTT with UNV assumption models, were calculated to assess the degree of comparability of item difficulty estimates. The correlations between the item discrimination parameter, a , obtained from the IRT models, and the factor loading, λ , obtained from the CTT with UNV assumption models, were obtained to assess the degree of comparability of item discrimination estimates. Finally, the median correlation and the range of correlations for each pairing across 100 replicates were computed for reporting. The correlation calculations were handled by R version 3.0.2.

Results

Results appear in Tables 2 to 9. Each table presents results for one of the model pairings. The same model used to generate the data was fit back to the data, and the resulting estimates were compared to another model. Tables 2 and 3 concern data generated by the 1PL IRT model. The 1PL IRT model fit is compared with the parallel CTT with UNV assumption model in Table 2 and to the congeneric CTT with UNV assumption model in Table 3. Tables 4 and 5 concern data generated by the 2PL IRT model, compared again to the parallel and congeneric CTT with UNV assumption models.

Tables 6 and 7 use data generated under the parallel CTT with UNV assumption model. The fit of the parallel model was compared with estimates from the 1PL IRT model in Table 6 and the 2PL IRT model in Table 7. Tables 8 and 9 make similar comparisons using data generated under the congeneric CTT with UNV assumption model.

Thus, Tables 2 and 6 make the same comparisons, estimates from the 1PL IRT model and the parallel CTT with UNV assumption model. In Table 2, the data originate from the 1PL IRT model, whereas in Table 6, the data originate from the parallel CTT with UNV assumption model. Analogous pairings occur for Tables 3 and 7, Tables 4 and 8, and Tables 5 and 9.

The results of the analyses comparing IRT item difficulty parameter estimates with those for item threshold τ from the CTT with UNV assumption appear in the center three columns in each of Tables 2 to 9. The correlations, based on the models paired in each table, indicate a high degree of comparability. Note that there is only one set of threshold estimates for both the parallel and congeneric CTT with UNV assumption models. This occurs because, when fitting the model to the data, the thresholds are estimated as a first step, and the specific model (parallel or congeneric) is fit as a

Table 2. Correlation Between IPL IRT and Parallel Fits, Data From IPL IRT Model.

IPL data vs. parallel fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	•	•	•	1.000	1.000	1.000	0.999	0.999	1.000
Case 2	•	•	•	1.000	1.000	1.000	0.999	0.999	0.999
Case 3	•	•	•	1.000	1.000	1.000	0.999	1.000	1.000
Case 4	•	•	•	1.000	1.000	1.000	0.999	1.000	1.000
Case 5	•	•	•	1.000	1.000	1.000	1.000	1.000	1.000
Case 6	•	•	•	1.000	1.000	1.000	1.000	1.000	1.000

Note. IPL IRT = one-parameter logistic item response theory. • = Discrimination parameter is not available in the IPL IRT model, and loadings are not unique in the parallel model. With these models, the correlation is not calculated.

Table 3. Correlations Between Congeneric and IPL IRT Fits, Data From IPL IRT Model.

IPL data vs. congeneric fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	•	•	•	0.993	0.998	0.999	0.997	0.998	0.999
Case 2	•	•	•	0.995	0.999	1.000	0.998	0.999	0.999
Case 3	•	•	•	0.986	0.998	0.999	0.999	0.999	0.999
Case 4	•	•	•	0.997	0.999	1.000	0.999	0.999	1.000
Case 5	•	•	•	0.995	0.998	0.999	0.999	0.999	1.000
Case 6	•	•	•	0.998	0.999	1.000	0.999	0.999	1.000

Note. IPL IRT = one-parameter logistic item response theory. • = Discrimination parameter is not available in the IPL IRT model, and loadings are not unique in the parallel model. With these models, the correlation is not calculated.

Table 4. Correlations Between Parallel and 2PL IRT Fits, Data From 2PL IRT Model.

2PL data vs. parallel fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	•	•	•	1.000	1.000	1.000	0.995	0.997	0.998
Case 2	•	•	•	1.000	1.000	1.000	0.995	0.997	0.999
Case 3	•	•	•	1.000	1.000	1.000	0.995	0.998	0.999
Case 4	•	•	•	1.000	1.000	1.000	0.996	0.998	0.998
Case 5	•	•	•	1.000	1.000	1.000	0.996	0.998	0.999
Case 6	•	•	•	1.000	1.000	1.000	0.997	0.998	0.999

Note. 2PL IRT = two-parameter logistic item response theory. • = Discrimination parameter is not available in the 1PL IRT model, and loadings are not unique in the parallel model. With these models, the correlation is not calculated.

Table 5. Correlations Between Congeneric and 2PL IRT Fits, Data From 2PL IRT Model.

2PL data vs. congeneric fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	0.656	0.909	0.983	0.991	0.997	0.999	0.998	0.999	1.000
Case 2	0.877	0.952	0.986	0.992	0.997	0.999	0.999	0.999	1.000
Case 3	0.505	0.913	0.974	0.992	0.997	0.999	0.998	0.999	1.000
Case 4	0.887	0.945	0.973	0.995	0.997	0.999	0.999	1.000	1.000
Case 5	0.696	0.914	0.963	0.994	0.997	0.999	0.998	0.999	1.000
Case 6	0.887	0.948	0.977	0.995	0.998	0.999	0.999	1.000	1.000

Note. 2PL IRT = two-parameter logistic item response theory.

Table 6. Correlations Between IPL IRT and Parallel Fits, Data From Parallel Model.

Parallel data vs. IPL fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	•	•	•	1.000	1.000	1.000	0.999	1.000	1.000
Case 2	•	•	•	1.000	1.000	1.000	0.999	1.000	1.000
Case 3	•	•	•	1.000	1.000	1.000	1.000	1.000	1.000
Case 4	•	•	•	1.000	1.000	1.000	1.000	1.000	1.000
Case 5	•	•	•	1.000	1.000	1.000	1.000	1.000	1.000
Case 6	•	•	•	1.000	1.000	1.000	1.000	1.000	1.000

Note. IPL IRT = one-parameter logistic item response theory. • = Discrimination parameter is not available in the IPL IRT model, and loadings are not unique in the parallel model. With these models, the correlation is not calculated.

Table 7. Correlations Between 2PL IRT and Parallel Fits, Data From Parallel Model.

Parallel data vs. 2PL fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	•	•	•	1.000	1.000	1.000	0.995	0.997	0.999
Case 2	•	•	•	1.000	1.000	1.000	0.997	0.998	0.999
Case 3	•	•	•	1.000	1.000	1.000	0.997	0.998	0.999
Case 4	•	•	•	1.000	1.000	1.000	0.999	0.999	0.999
Case 5	•	•	•	1.000	1.000	1.000	0.998	0.999	0.999
Case 6	•	•	•	1.000	1.000	1.000	0.999	0.999	1.000

Note. 2PL IRT = two-parameter logistic item response theory. • = Discrimination parameter is not available in the IPL IRT model, and loadings are not unique in the parallel model. With these models, the correlation is not calculated.

Table 8. Correlations Between 1PL IRT and Congeneric Fits, Data From Congeneric Model.

Congeneric data vs. 1PL fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	•	•	•	0.962	0.989	0.998	0.991	0.995	0.997
Case 2	•	•	•	0.976	0.992	0.997	0.992	0.996	0.998
Case 3	•	•	•	0.966	0.991	0.995	0.996	0.997	0.998
Case 4	•	•	•	0.982	0.991	0.995	0.996	0.997	0.998
Case 5	•	•	•	0.968	0.988	0.996	0.996	0.998	0.999
Case 6	•	•	•	0.985	0.991	0.995	0.997	0.998	0.999

Note. 1PL IRT = one-parameter logistic item response theory. • = Discrimination parameter is not available in the 1PL IRT model, and loadings are not unique in the parallel model. With these models, the correlation is not calculated.

Table 9. Correlations Between 2PL IRT and Congeneric Fits, Data From Congeneric Model.

Congeneric data vs. 2PL fit	Correlation in discrimination parameter			Correlation in difficulty parameter			Correlation in factor scores		
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Case 1	0.892	0.963	0.991	0.962	0.989	0.998	0.999	0.999	1.000
Case 2	0.899	0.97	0.991	0.976	0.992	0.997	0.999	0.999	1.000
Case 3	0.903	0.966	0.983	0.966	0.991	0.995	0.999	1.000	1.000
Case 4	0.906	0.966	0.986	0.982	0.991	0.995	0.999	1.000	1.000
Case 5	0.901	0.963	0.986	0.968	0.988	0.996	0.999	1.000	1.000
Case 6	0.918	0.967	0.984	0.985	0.991	0.995	0.999	1.000	1.000

Note. 2PL IRT = two-parameter logistic item response theory.

second step. The median correlation coefficient consistently is quite high among all cases. Furthermore, the range of correlation coefficients is very narrow.

The results of the analysis comparing IRT item discrimination parameter estimates with those for factor loading λ from the CTT with UNV assumption model appear in the first three columns of Tables 5 and 9. Item discrimination indices are not estimated in the 1PL IRT model, and loadings are not estimated in the parallel CTT with UNV assumption model. As a result, no correlations are reported for Tables 2 to 4 and Tables 6 to 8. The correlations in Tables 5 and 9, based on the models paired in those tables, were somewhat lower than the results of item difficulty estimates. Of particular interest are correlations of 2PL IRT and congeneric fits when data were generated by the 2PL IRT model (Table 5). The correlation seems more sensitive to sample size than test length, though in the main all correlations are quite good. Minima reflect the range reported by Fan (1998), especially considering that work reports averages across simulations. Table 9 reflects the same trends, though not as pronounced.

Finally, correlations of factor score estimates to those for ability parameter, the last three columns in each of Tables 2 to 9, also are quite strong across all conditions and model pairings.

There was almost no difference in correlation values for analogous pairings regardless of the model from which the data were generated. For example, the correlations calculated between the difficulty parameter estimate for 1PL IRT and threshold parameter estimate for congeneric CTT with UNV assumption when the data were generated from the 1PL IRT model, Table 2, were quite similar to correlations calculated between those parameters for those models when the data were generated from the congeneric model, Table 6. The exception is with discrimination parameter estimates; correlations were slightly lower when the data were generated with the 2PL IRT model, Table 5, than when the data were generated with the congeneric CTT with UNV assumption model, Table 9.

Discussion

The findings of this study reflect results in the literature. Item difficulty parameter estimates obtained from the IRT and the CTT with UNV assumption models were highly comparable across all conditions and model pairings. This is consistent with the findings of Fan (1998) and MacDonald and Paunonen (2002). The correlation for item discrimination parameter estimates obtained from the IRT and the CTT with UNV assumption models were lower, also reflected by Fan (1998) and MacDonald and Paunonen (2002). We find greater sensitivity to sample size than to test length, though the difference between small and large sample size is more pronounced for shorter tests.

For the most part, the correlations were high for all model pairings regardless of which element in the pair had served as the data generation model. Only in the

discrimination parameter estimate correlations is any noticeable difference seen. This finding lends support to the idea that the two modeling frameworks have equal merit.

MacDonald and Paunonen (2002) suggested high accuracy for the discrimination parameter, as measured by correlation of estimate to true value, only when the difficulty parameter is restricted to a narrow range. Such a finding can be expected in light of the characterization by McCullagh and Nelder (1989) of the relationship between logistic and probit functions as “almost linearly related over the interval $0.1 \leq \pi \leq 0.9$ ” (p. 109). As the cumulative probabilities corresponding to the bounds of the threshold parameter for our congeneric model, 0.16 to 0.88, approach these values, the deteriorated correlation in the discrimination parameter may simply reflect the deterioration of the linearization approximation between the two functional forms.

The accepted view of IRT item parameter invariance is founded on the parameters’ function in the abstract model concept. In contrast, the understood susceptibility of item statistics in the CTT framework to variations in data concerns specific sample estimates. Purported superiority results from a comparison of unlike concepts. In this article, we show comparability between the CTT with UNV assumption and IRT, both in concept and through correlation of analogous parameter estimates. The implication of this comparability is that the strength of the model concept in IRT can apply equally to a CTT-based model with UNV assumption. Conversely, cautions regarding parameter estimation in CTT-based models with UNV assumption apply to parameter estimation in IRT models as well.

The invariance of IRT item parameters at the conceptual level is not, actually, absolute. Lord (1980) notes that the location and scale of the ability variable is arbitrary. This fact means that, as noted by Rupp and Zumbo (2006), item parameters actually are invariant only up to a linear transformation unless the location and scale of the ability variable are held constant from test group to test group.

Moreover, the fact that parameter estimates often are sensitive to the sample on which the estimates are based is not inherently a defect. Such an attribute can allow the researcher to uncover variations that affect the outcome of interest and thus advance the field. Researchers need simply to keep this feature in mind when interpreting model results.

High correlations cannot distinguish the scenario of stable estimates between the two frameworks and the drift that is the same in each framework. This is a limitation of the present study. This study also does not directly examine parameter estimate stability as functions of either ability distribution of the test sample or sample size. An investigation focusing on the latter would represent an important contribution, since most performance attributes regarding parameter estimation rely on an assumption of a sufficiently large, yet unquantified, sample size. But real data are always limited in number.

The theoretical framework used here along with the correlation results show the frameworks of IRT and CTT with UNV assumption to be quite comparable, with neither framework showing an advantage over the other. This finding presents the

opportunity for CTT with UNV models to be applied in contexts where they had not been considered previously.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. Essentially parallel, essentially tau-equivalent, and essentially congeneric models are not specifically addressed. However, the results and conclusions of this study using models defined without a mean structure are equally applicable to corresponding models defined with a mean structure, as correlations are invariant to the locations of the scales.

References

- Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29. Retrieved from <http://www.jstatsoft.org/v48/i06/>
- Cook, L. L., Eignor, D. R., & Hessy, L. T. (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(1), 31-45.
- Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3141396)
- DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage.
- 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.
- Ferrando, P. J. (2000). Testing the equivalence among different item response formats in personality measurement: A structural equation modeling approach. *Structural Equation Modeling*, 7, 271-286.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, 66, 930-944.
- Hambleton, R., & Jones, R. (1993). Comparison of classical test theory and item response theory. *Educational Measurement: Issues and Practice, Fall*, 38-47.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109-133.
- Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24, 387-404.

- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modelling*, 15, 136-153.
- 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 Press.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- MacDonald, P., & Paunonen, S. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62, 921-943.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. New York, NY: Psychology Press.
- Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, 11, 3-31.
- Miyazaki, Y. (2005). Some links between classical and modern test theory via the two-level hierarchical generalized linear model. *Journal of Applied Measurement*, 6, 289-310.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.
- Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43, 551-560.
- Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, 46, 407-419.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- R Development Core Team. (2013). *R: A language and environment for statistical computing* (ISBN 3-900051-07-0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.
- Revelle, W. (2014). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, IL: Northwestern University. Retrieved from <http://CRAN.R-project.org/package=psych>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. Retrieved from <http://www.jstatsoft.org/v48/i02/>
- Rudner, L. M. (1983). A closer look at latent trait parameter invariance. *Educational and Psychological Measurement*, 43, 951-955.
- Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66, 63-84.
- Sharkness, J., & DeAngelo, L. (2011). Measuring student involvement: A comparison of classical test theory and item response theory in the construction of scales from student surveys. *Research in Higher Education*, 52, 480-507.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58-79.