The Reliability and Precision of Total Scores and IRT Estimates as a Function of Polytomous IRT Parameters and Latent Trait Distribution

Steven Andrew Culpepper

Abstract

A classic topic in the fields of psychometrics and measurement has been the impact of the number of scale categories on test score reliability. This study builds on previous research by further articulating the relationship between item response theory (IRT) and classical test theory (CTT). Equations are presented for comparing the reliability and precision of scores within the CTT and IRT frameworks. This study presented new results pertaining to the relative precision (i.e., the test score conditional standard error of measurement for a given trait value) of CTT and IRT, and the new results shed light on the conditions where total scores and IRT estimates are more or less precisely measured. The relative reliability of CTT and IRT scores is examined as a function of item characteristics (e.g., locations, category thresholds, and discriminations) and subject characteristics (e.g., the skewness and kurtosis of the latent distribution). CTT total scores were more reliable when the latent distribution was mismatched with category thresholds, but the discrepancy between CTT and IRT declined as the number of scale categories increased. This article also considered the appropriateness of linear approximations of polytomous items and presented circumstances where linear approximations are viable. A linear approximation may be appropriate for items with two response options depending on the item discrimination and the match between the item location and latent distribution. However, linear approximations are biased whenever items are located in the tails of the latent distribution and the bias is larger for more discriminating items.

Keywords

reliability, scale construction, classical test theory, item response theory, polytomous items, information function

The impact of the number of scale categories on reliability is a classic topic in psychology (Symonds, 1924), and an extensive body of research has examined the effect of the number of

1University of Illinois at Urbana–Champaign, USA

Corresponding Author:
Steven Andrew Culpepper, University of Illinois at Urbana–Champaign, 116D Illini Hall, MC-374, 725 South Wright Street, Champaign, IL 61820, USA.
Email: sculpepp@illinois.edu
scale categories on the corresponding reliability of total scores ($\tau$). In fact, researchers examined the impact of the number of scale categories on total score reliability using empirical data (Adelson & McCoach, 2010; Bendig, 1954; Chafouleas, Christ, & Riley-Tillman, 2009; L. Chang, 1994; Komorita & Graham, 1965; Matell & Jacoby, 1971; Weng, 2004), Monte Carlo simulations (Aguinis, Pierce, & Culpepper, 2009; Bandalos & Enders, 1996; Cicchetti, Shoinralter, & Tyrer, 1985; Enders & Bandalos, 1999; Greer, Dunlap, Hunter, & Berman, 2006; Jenkins & Taber, 1977; Lissitz & Green, 1975), and analytic derivations (Krieg, 1999).

Methodological developments have bridged the concept of reliability between the item response theory (IRT) and classical test theory (CTT) frameworks and discussed concepts that have traditionally been reserved for IRT (e.g., item information functions [IIFs]) within the context of CTT. The purpose of this article is to understand the circumstances where researchers should prefer estimating true scores (i.e., $\theta$) with test scores derived from IRT (i.e., $\hat{\theta}$) versus CTT (i.e., $\tau$). This article compares the precision and reliability of $\theta$ and $\tau$, and equations are presented for investigating how item and subject characteristics affect the reliability of $\theta$ and $\tau$. Mellenbergh (1996) noted that reliability is a population-dependent quantity that is affected by characteristics of latent distributions, whereas the conditional standard error of measurement (CSEM) quantifies error variance (i.e., precision or the inverse of information) for a specific $\theta$ value. New equations are presented that compare the relative precision of $\theta$ and $\tau$. The results show that IRT and CTT have CSEMs that are roughly the mirror image across values of $\theta$. That is, $\theta$ is measured more precisely in portions of the $\theta$ continuum that include relatively more category thresholds, whereas $\tau$ tends to be measured more precisely for $\theta$ values that are further from category thresholds.

It is important to articulate the contributions of this article to existing research. First, methodological advances concerning dichotomous items established a link between the IRT and CTT frameworks by articulating the reliability of total scores and percentile ranks with corresponding IRT item parameters, such as item difficulty, discrimination, and guessing parameters (Bechger, Maris, Verstralen, & Beguin, 2003; Dimitrov, 2003; Kolen, Zeng, & Hanson, 1996; May & Nicewander, 1994). Additional research has studied the impact of IRT item parameters on the reliability of gain scores (May & Jackson, 2005), the lower and upper bounds of the IRT reliability coefficient (Kim & Feldt, 2010), the reliability of scale scores and performance assessments using polytomous IRT (Wang, Kolen, & Harris, 2000), and the reliability of subscores using unidimensional (Haberman, 2008) and multidimensional IRT (Haberman & Sinharary, 2010). Researchers have also found that IRT scores provide more accurate estimates of interaction effects within the contexts of analysis of variance (Embretson, 1996) and multiple regression (Kang & Waller, 2005; Morse, Johanson, & Griffeth, 2012). This article offers new information about the connection between the concepts of reliability and precision in CTT and IRT. That is, no study has provided theoretical results about the relative precision of $\theta$ and $\tau$ for a given $\theta$ value. The new derivations provide theoretical rationale for circumstances when CTT scores include relatively more or less measurement error than IRT scores. Moreover, no research study has analytically studied the interactive effect of item characteristics and latent distribution shape on the relative reliability of $\theta$ and $\tau$. This article uses Fleishman’s (1978) power transformation (PT) method probability density function to study the impact of nonnormal latent distributions on the reliability of total scores.

Second, additional research has extended the concept of item and test information to CTT under the assumptions that observed measurements are continuous, rather than polytomous, and $\theta$ is linearly related to $\tau$ (Ferrando, 2002, 2009; McDonald, 1982; Mellenbergh, 1996). Ferrando notes that a linear model provides a good approximation when item discrimination indices are relatively small in value and coarsely measured items have five or more response categories. Moreover, Ferrando and Mellenbergh showed that, for the linear model, the CTT
IIF is horizontal and unrelated to $\theta$. However, whenever the relationship between latent and observed total scores is nonlinear (which occurs when items are polytomous), the CTT IIFs are no longer unrelated to $\theta$. In fact, this article shows that the conditional standard error of $x$ given $\theta$ is a downward-facing function where scores near category thresholds have the least amount of precision. The accuracy of a linear relationship is also explored, and the results in this article examine the effect of test characteristics (e.g., item locations and discrimination) and subject characteristics (e.g., latent distribution shape) on the appropriateness of linear approximations of polytomous items.

Third, previous Monte Carlo studies that have studied the relative performance of CTT and IRT estimates are limited by the combination of parameter values and type of reliability studied. For instance, Greer et al. (2006) studied the impact of skewness on coefficient alpha with the constraint that item variances were equal, which may not occur frequently in practice. The results in this article can be used to study any combination of IRT parameter values and latent distribution shape and provide more general results than previous Monte Carlo simulations. In addition, Wang et al. (2000) presented equations for computing the reliability of scale scores from performance assessments using the generalized partial credit model. However, unlike Wang et al., this study presents equations for evaluating how the number of items, number of scale categories, and the shape of the latent distribution affects the reliability of $\hat{\theta}$ and $x$. Furthermore, R code (R Development Core Team, 2010) is available at http://publish.illinois.edu/sculpepper/, and researchers can use the R code to compute the reliability of $\hat{\theta}$ and $x$ for different item characteristics and latent distributions. Consequently, this study presents new results and provides applied researchers with guidance for reliably scoring tests in different situations.

This article includes five sections. The first section presents equations for the reliability and precision of scores within the CTT and IRT paradigms and includes new results about the CSEM for CTT. The second section compares IRT and CTT in terms of CSEMs to provide a general understanding of the circumstances researchers should prefer CTT versus IRT scores. The third section compares the relative reliability of $\hat{\theta}$ and $x$ for different item (e.g., item locations and number of response categories) and subject distribution characteristics (e.g., the skewness and kurtosis of $\theta$), and the fourth section examines how item and subject characteristics affect the appropriateness of a linear approximation of polytomous items. The last section discusses the results and provides recommendations and concluding remarks.

**Equations for the Reliability of CTT and IRT Estimates of $\theta$**

Let $\theta$ represent a latent variable and $u_i$ be an observed polytomous response, where $i$ indexes items ($i = 1, \ldots, I$). The observed polytomous response for item $i$ can be expressed as a function of a true score ($E\{u_i|\theta\}$) and random error ($e_i$), such that $u_i = E\{u_i|\theta\} + e_i$. That is, the observed $u_i$ equals an item true score $E\{u_i|\theta\}$ (May & Nicewander, 1994), which is a nonlinear function of $\theta$, plus an error, $e_i$. Let $j$ index category thresholds ($j = 1, \ldots, J$) and $J+1$ is the corresponding number of categories for item $i$. That is, $j$ is used to index categories as well (e.g., $J+1 = 4$ implies that $u_i$ has four categories). This article assumes that researchers code $u_i$ using integers from 1 to $J+1$. Several polytomous models exist to describe the relationship between $\theta$ and the chance that $u_i$ equals one of $J+1$ categories, which include the graded response (Molenaar, Dolan, & De Boeck, 2012; Muraki, 1990; Samejima, 1969), partial credit (Masters, 1982; Muraki, 1992, 1993), and rating scale (Andrich, 1978a, 1978b) models. The derivations in this article use Muraki’s (1990) modified graded response model,
where $b_i$ and $\alpha_i$ are the item difficulty and discrimination parameters, respectively, and $c_j$ is the $j$th threshold, which is equal for all $I$ items. Equation 1 represents the chance that $u_i > j$ given $\theta$ and the probability that item $i$ equals the $j$th category is

$$P_{ij}(u_i > j|\theta) = \begin{cases} 1, & j = 0 \\ \frac{1}{1 + \exp[-\alpha_i(\theta - (b_i - c_j))]}, & 0 < j \leq J \\ 0, & j = J + 1 \end{cases}$$

(1)

Note that Muraki’s model was chosen because the form of Equation 1 is easier to manipulate analytically and the derivations of expressions for derivatives are less cumbersome. Muraki’s model assumes that the $c_j$ are constant across items and are equally spaced. The goal of this article is not to estimate abilities or item parameters with Muraki’s model, and these assumptions can be relaxed to evaluate the impact of unequal item thresholds on the reliability and precision of $x$ and $\hat{\theta}$. In fact, the new expressions are applicable for any polytomous model, and Muraki’s model is only used for computational examples. Furthermore, the aforementioned models tend to yield scores that are highly correlated (Embretson & Reise, 2000), so we should not anticipate the results would change significantly if the partial credit or rating scale models were used.

The derivations below require the specification of a distribution for $\theta$. In this article, $\theta$ is assumed to follow a Fleishman PT distribution, $\phi(\theta|\Omega)$, where $\Omega = (\mu, \sigma^2, \kappa_3, \kappa_4)$ to indicate that $\theta$ has a mean $\mu$, variance $\sigma^2$, and skewness and kurtosis of $\kappa_3$ and $\kappa_4$, respectively ($\phi(\theta|\Omega)$ is discussed in greater detail in the Appendix). Note that any univariate distribution could be chosen for $\theta$. One advantage of using Fleishman’s PT distribution is that it is flexible enough to explore how changes in $\mu$, $\sigma^2$, $\kappa_3$, and $\kappa_4$ affect the relative reliability of IRT and CTT $\theta$ estimates. However, as noted by an anonymous reviewer, researchers often set $\sigma^2 = 1$ to estimate item discriminations (i.e., the scale indeterminacy problem). Accordingly, the results in this article also use $\sigma^2 = 1$ to understand how manipulating item discriminations affects reliability. Fleishman’s PT distribution does not encompass the universe of all univariate distributions, so future researchers can modify the associated R code to examine the reliability of CTT and IRT estimates when $\theta$ follows other distributions. Furthermore, the discussion below provides an argument as to how reliability within the CTT and IRT frameworks is dependent on the match between the shape of $\phi(\theta|\Omega)$ and the topography of the conditional variance of $x$ and $\hat{\theta}$.

Reliability and Precision Within IRT Framework

One of the strengths of IRT over CTT relates to the well-known measure of precision for estimated trait scores, $\hat{\theta}$. Reliability is specific to a group and is a function of the unconditional standard error of measurement (SEM). CTT has traditionally calculated the SEM for a group of scores, whereas the CSEM is a measure of precision that corresponds to a specific trait level within the IRT framework. The CSEM of $\hat{\theta}$ is related to the test information function (TIF), which is derived using the concept of Fisher’s information to measure the amount of information that a single observation provides about $\theta$. In fact, the inverse of a TIF indicates the variance of $\theta$ for a given $\theta$. Previous research (Muraki, 1993; Samejima, 1994) discussed TIFs for polytomous IRT models and noted that the IIF for item $i$ is
\[ I_i(\theta) = -\sum_{j=1}^{J+1} P_{ij}(u_i = j|\theta) \frac{\partial^2}{\partial \theta^2} \ln(P_{ij}(u_i = j|\theta)) \]
\[ = \sum_{j=1}^{J+1} \left[ \frac{\partial P_{ij}(u_i = j|\theta)}{\partial \theta} \right]^2 - \frac{\partial^2 P_{ij}(u_i = j|\theta)}{\partial \theta^2} \right]. \tag{3} \]

For Muraki’s (1990) modified graded response model, \( P_{ij}(u_i = j|\theta) \) is a function of \( P_{ij}^*(u_i = j|\theta) \), which is a logistic function. The first two derivatives of \( P_{ij}^*(u_i = j|\theta) \) are as follows:

\[
\begin{align*}
\frac{\partial P_{ij}^*(u_i = j|\theta)}{\partial \theta} &= \alpha_i P_{ij}^*(u_i = j|\theta) \left[ 1 - P_{ij}^*(u_i = j|\theta) \right], \\
\frac{\partial^2 P_{ij}^*(u_i = j|\theta)}{\partial \theta^2} &= \alpha_i^2 P_{ij}^*(u_i = j|\theta) \left[ 1 - 2P_{ij}^*(u_i = j|\theta) \right]. \end{align*} \tag{4} \]

Accordingly, \( I_i(\theta) \) can be computed using the first and second derivatives of the item category probabilities, \( P_{ij}(u_i = j|\theta) \), which are as follows:

\[
\begin{align*}
\frac{\partial P_{ij}(u_i = j|\theta)}{\partial \theta} &= \frac{\partial P_{ij}^*(u_i = j|\theta)}{\partial \theta} \frac{\partial \alpha_i}{\partial \theta}, \\
\frac{\partial^2 P_{ij}(u_i = j|\theta)}{\partial \theta^2} &= \frac{\partial^2 P_{ij}^*(u_i = j|\theta)}{\partial \theta^2} \frac{\partial \alpha_i}{\partial \theta}. \end{align*} \tag{5} \]

The TIF for a test of polytomous items is the sum of the respective IIFs:

\[ \text{TIF}(\theta) = \sum_{i=1}^{I} I_i(\theta). \tag{6} \]

The relationship between true and observed IRT estimates can be written as \( \hat{\theta} = \theta + \varepsilon \). If \( \theta \) and \( \varepsilon \) are independent, the variance of \( \hat{\theta} \) is the sum of the true and error variances, \( \sigma^2 \{ \hat{\theta} \} = \sigma^2 \{ \theta \} + \sigma^2 \{ \varepsilon \} \). The conditional variance of \( \hat{\theta} \) given \( \theta \) is defined as \( \sigma^2 \{ \hat{\theta} | \theta \} = \sigma^2 \{ \varepsilon | \theta \} = (\text{TIF}(\theta))^{-1} \). The expected conditional variance of \( \hat{\theta} \) for a specific distribution of \( \theta \), \( \varphi(\theta|\Omega) \), is as follows:

\[ E \{ \sigma^2 \{ \hat{\theta} | \theta \} \} = \int_{-\infty}^{\infty} \sigma^2 \{ \hat{\theta} | \theta \} \varphi(\theta|\Omega) d\theta. \tag{7} \]

The expected reliability of \( \hat{\theta} \) is the well-known ratio of true to observed variance,

\[ \rho_{\theta \hat{\theta}} = \frac{\sigma^2}{\sigma^2 + E \{ \sigma^2 \{ \hat{\theta} | \theta \} \}}, \tag{8} \]

where \( \sigma^2 \) is the variance of \( \theta \) specified in \( \varphi(\theta|\Omega) \). Clearly, \( \rho_{\theta \hat{\theta}} \) is dependent on the characteristics of the test (i.e., \( \sigma^2 \{ \hat{\theta} | \theta \} \)) and the distribution of latent scores (i.e., \( \varphi(\theta|\Omega) \)).
Reliability and Precision Within CTT Framework

Let \( x \) be the total score, or sum, of the \( I \) \( u_i \), such that \( u_i = E\{u_i|\theta\} + e_i \). For a subject with a given \( \theta \), \( u_i \) equals one of \( J + 1 \) categories each with probability \( P_j(u_i=j|\theta) \), so \( E\{u_i|\theta\} \) is

\[
E\{u_i|\theta\} = \sum_{j=1}^{J+1} j P_j(u_i=j|\theta). \tag{9}
\]

That is, \( E\{u_i|\theta\} \) is a weighted average of the category scores (i.e., \( j = 1 \) to \( J + 1 \)) and the chance that subjects with a specific \( \theta \) have an observed score \( u_i \). The previous section discussed an expression for the expected conditional variance of \( \theta \) as a function of the inverse TIF and distribution of \( \theta \). This section analogously derives an expression for \( E\{\sigma^2\{x|\theta\}\} \), which is the expected variance of \( x \) for a given value of \( \theta \). An equation for the reliability of \( x \) is presented as a function of \( E\{\sigma^2\{x|\theta\}\} \) and the variance of the expected true scores, \( \sigma^2\{E\{u_i|\theta\}\} \). This subsection proceeds by first deriving an expression for the conditional error variance of \( x \), \( E\{\sigma^2\{x|\theta\}\} \), and then identifies an equation for the true score variance, \( \sigma^2\{E\{u_i|\theta\}\} \).

Consider item \( i \) where the error is \( e_i = u_i - E\{u_i|\theta\} \). Note that \( u_i \) is a coarse measure of \( E\{u_i^2|\theta\} \) in that \( u_i \) is ordinal and equals one of the \( J + 1 \) values, whereas \( \theta \) is measured on an interval scale. One immediate observation is that \( E\{e_i|\theta\} = 0 \). Recall that \( u_i \) is a polytomous item, so \( e_i = j - E\{u_i|\theta\} \) for \( j = 1 \) to \( J + 1 \). The conditional expectation of the error within item \( i \) is

\[
E\{e_i|\theta\} = \sum_{j=1}^{J+1} P_j(u_i=j|\theta) (j - E\{u_i|\theta\}) = \sum_{j=1}^{J+1} j P_j(u_i=j|\theta) - E\{u_i|\theta\} = 0. \tag{10}
\]

The variability of observed \( u_i \) around expected values is determined by the error variance or precision; that is, \( \sigma^2\{u_i|\theta\} = E\{e_i^2|\theta\} \) and is defined as

\[
\sigma^2\{u_i|\theta\} = E\{(u_i - E\{u_i|\theta\})^2|\theta\} = \sum_{j=1}^{J+1} P_j(u_i=j|\theta) (j - E\{u_i|\theta\})^2. \tag{11}
\]

Equation 11 is new to the literature and the following sections compare the properties of \( \sigma^2\{u_i|\theta\} \) with \( \sigma^2\{\theta|\theta\} \).

\( \sigma^2\{u_i|\theta\} \) is the conditional variance for a single item and an expression is needed for the conditional variance of \( x \). An important observation is that errors within two polytomous items, say, \( u_i \) and \( u_h \) (let categories for \( u_h \) be indexed by \( k \)), are independent whenever the items are locally independent, which assumes that \( P(u_i=j, u_h=k|\theta) = P_j(u_i=j|\theta)P_{hk}(u_h=k|\theta) \). Specifically, the covariation between \( e_i \) and \( e_h \) conditioned on \( \theta \) is

\[
\sigma\{e_i, e_h|\theta\} = E\{(u_i - E\{u_i|\theta\})(u_h - E\{u_h|\theta\})|\theta\} = \sum_{j=1}^{J+1} \sum_{k=1}^{J+1} (j - E\{u_i|\theta\})(k - E\{u_h|\theta\}) P(u_i=j, u_h=k|\theta) = \sum_{j=1}^{J+1} (j - E\{u_i|\theta\}) P_j(u_i=j|\theta) \sum_{k=1}^{J+1} (k - E\{u_h|\theta\}) P_{hk}(u_h=k|\theta) = 0. \tag{12}
\]

The finding that errors within items are independent, if local independence is assumed, is particularly important, because the conditional error variance of the total score, \( x \), given \( \theta \) is simply the sum of the conditional item variances, \( \sigma^2\{u_i|\theta\} \).
Recall that $x = \sum_{i=1}^{I} u_{i}$. Equation 9 implies that the expected value of $x$ given $\theta$ is

$$E\{x|\theta\} = \sum_{i=1}^{I} E\{u_{i}|\theta\} = \sum_{i=1}^{I} \sum_{j=1}^{J} P_{ij}(u_{i}=j|\theta).$$  \(13\)

The conditional variance of $x$ given $\theta$ is the sum of conditional variances for the $I u_{i}$, under the assumption of local independence (see Equation 12), which implies that $\sigma^{2}\{x|\theta\} = \sum_{i=1}^{I} \sigma^{2}\{u_{i}|\theta\}$. The expected conditional variance for a specific distribution of $\theta$ is

$$E\{\sigma^{2}\{x|\theta\}\} = \int \sigma^{2}\{x|\theta\} \varphi(\theta|\Omega) d\theta.$$  \(14\)

Recall that in IRT the expected variance of maximum likelihood estimates is the expected value of $\sigma^{2}\{\hat{u}_{i}|\theta\}$ across the distribution of $\theta$. Similarity, the expected conditional variance of $x$ is found by replacing $\sigma^{2}\{\hat{u}_{i}|\theta\}$ with $\sigma^{2}\{x|\theta\}$. Consequently, researchers can compare the CSEM (i.e., $\sigma\{x|\theta\}$ and $\sigma\{\hat{u}_{i}|\theta\}$) to understand which values of $\theta$ are associated with relatively more precision within the CTT and IRT frameworks.

Recall that $E\{x|\theta\}$ is the expected total score for subjects with a given $\theta$ and the variance of $E\{x|\theta\}$ across subjects provides a measure of the amount of true score variance. First, note that the unconditional mean of $x$ is

$$E\{x\} = E\{E\{x|\theta\}\} = \int_{-\infty}^{\infty} E\{x|\theta\} \varphi(\theta|\Omega) d\theta.$$  \(15\)

The variance of $E\{x|\theta\}$ across $\theta$ is

$$\sigma^{2}\{E\{x|\theta\}\} = E\{(E\{x|\theta\} - E\{x\})^{2}\} = \int_{-\infty}^{\infty} (E\{x|\theta\} - E\{x\})^{2} \varphi(\theta|\Omega) d\theta.$$  \(16\)

The reliability of $x$ is the ratio of true to observed variance:

$$\rho_{xx} = \frac{\sigma^{2}\{E\{x|\theta\}\}}{\sigma^{2}\{E\{x|\theta\}\} + E\{\sigma^{2}\{x|\theta\}\}}.$$  \(17\)

**Factors That Affect the Precision of CTT and IRT Scores**

The previous section derived expressions for the CSEMs for $x$ and $\hat{\theta}$ (i.e., $\sigma\{x|\theta\}$ and $\sigma\{\hat{\theta}|\theta\}$). The characteristics of polytomous IRT CSEMs are well understood. For example, $\sigma\{\hat{\theta}|\theta\}$ tends to be smaller at points along the latent continuum where category thresholds are located and $\sigma\{\theta|\theta\}$ declines in regions where more discriminating items are located. In contrast, the expression for $\sigma\{x|\theta\}$ is new and the purpose of this section is to compare $\sigma\{x|\theta\}$ with $\sigma\{\hat{\theta}|\theta\}$ to provide researchers with a conceptual understanding of the testing situations where $x$ may be preferred to $\hat{\theta}$ in terms of measurement precision.

As an example, consider a test consisting of three items that each have four response categories. Moreover, let $\alpha$ and $\beta$ be three-dimensional vectors of item discriminations and locations, such that $\alpha = (0.5, 1.5, 2.5)$ and $\beta = (-1, 0, 1)$. Moreover, to simplify this example, let
the category thresholds be equally spaced with values of \((-1.6, 0, 1.6)\) units below and above \(\theta - b_i\) (i.e., the thresholds for the first item are located at \(-2.6, -1.6, \) and 0.6 on the \(\theta\) scale).

Figure 1 presents \(\sigma^2\{u_i|\theta\}\) (see Equation 11) and \([I_i(\theta)]^{-1}\) (see Equation 3) for the three items. The IRT conditional variances exhibit expected behavior. That is, \([I_i(\theta)]^{-1}\) is lowest at the category thresholds and \([I_j(\theta)]^{-1}\) is generally smaller than the other two items because the third item is more discriminating. One additional nuance is that \([I_j(\theta)]^{-1}\) has local minima at the category thresholds, whereas Items 1 and 2 have inverse information functions that appear smoother. Stated differently, in IRT, items with larger discriminations tend to have IIFs with more topography in regions near category thresholds.

As expected, \(\sigma^2\{u_i|\theta\}\) is smaller for more discriminating items; however, the general behavior of \(\sigma^2\{u_i|\theta\}\) is different from \([I_j(\theta)]^{-1}\). Namely, \(\sigma^2\{u_i|\theta\}\) tends to be a downward-facing function where measurement error is largest at category thresholds. For example, \(\sigma^2\{u_3|\theta\}\) tends to be the smallest of the three items, but \(\sigma^2\{u_3|\theta\}\) has maxima near the category thresholds, which differs from IRT where measurements are more precise at category thresholds. Under a
CTT framework, $\sigma^2\{u_i|\theta\}$ is smallest for either more extreme $\theta$ or for $\theta$ values that lie between category thresholds. In short, $\sigma^2\{u_i|\theta\}$ is roughly the mirror image of $[I_i(\theta)]^{-1}$, because $[I_i(\theta)]^{-1}$ tends to be lower in segments of the $\theta$ continuum where $\sigma^2\{u_i|\theta\}$ is larger and vice versa.

Figure 2 includes the same items discussed in Figure 1, but with the exception that the thresholds are no longer equally spaced. Figure 2 shows that $\sigma^2\{u_i|\theta\}$ and $[I_i(\theta)]^{-1}$ respond inversely to unequal item thresholds. For instance, $\sigma^2\{u_i|\theta\}$ tends to increase in portions of the latent continuum that include more item thresholds, whereas $[I_i(\theta)]^{-1}$ is smaller in segments of the latent continuum where there are more item thresholds and increases wherever there are fewer thresholds.

Figure 3 demonstrates CTT and IRT test CSEMs along with error bars around conditional expected values, $E\{x|\theta\}$ and $E\{\theta|\theta\}$. The top row of Figure 3 illustrates the CTT and IRT CSEMs for $x$ and $\theta$, respectively, for the hypothetical three-item test. Note that the vertical lines in the top row of panels in Figure 3 indicate category thresholds for Items 1, 2, and 3. Figure 3 shows that $\sigma^2\{x|\theta\}$ is larger for $\theta$ values that are near category thresholds, whereas $\sigma^2\{\theta|\theta\}$ is smaller at points on the latent continuum that have more thresholds and more discriminating
items. For the three-item test, $\sigma^2 \{ \hat{\theta} | \theta \}$ appears more responsive to item discriminations than $\sigma^2 \{ x | \theta \}$ as indicated by the fact that the slope of $\sigma^2 \{ \hat{\theta} | \theta \}$ is steeper than $\sigma^2 \{ x | \theta \}$ in the $\theta$ range measured by Item 1.

The second row of panels in Figure 3 plots $E \{ x | \theta \}$ and $E \{ \hat{\theta} | \theta \}$ as well as $\pm 2$ times the CSEMs. As discussed previously, $E \{ x | \theta \}$ is a more accurate indicator of $\theta$ for more extreme values on the $\theta$ continuum, whereas $\hat{\theta}$ is a better indicator in the range where items and categories are located. The overall reliability of $x$ and $\hat{\theta}$ is dependent on the shape and location of the latent distribution. For example, if $\theta \sim N(0, 1)$ most subjects lie in the middle range of the latent continuum where $x$ is less precisely measured relative to $\hat{\theta}$. In fact, the differences in $\sigma^2 \{ x | \theta \}$ and $\sigma^2 \{ \hat{\theta} | \theta \}$ contribute to $\hat{\theta}$ being significantly more reliable than $x$ (i.e., 0.65 vs. 0.49). Certainly, $\rho_{xx}$ and $\rho_{\hat{\theta}\hat{\theta}}$ will change depending on the shape and location of the $\theta$ distribution.

Figure 3. Test score CTT and IRT conditional standard error of measurement and expected value plots with $\pm 2$ error curves for a hypothetical three-item test with four response categories. Note: CTT = classical test theory; IRT = item response theory. Thresholds are indicated by dashed-vertical lines for Item 1 ($b_1 = -1$, $a_1 = 0.5$), Item 2 ($b_2 = 0$, $a_2 = 1.5$), and Item 3 ($b_3 = 1$, $a_3 = 2.5$) and $c_j = (-1.6, 0, 1.6)$. $\rho_{xx}$ and $\rho_{\hat{\theta}\hat{\theta}}$ were calculated under the assumption that $\theta \sim N(0, 1)$.
Figure 4. Reliability of CTT and IRT scores of a 10-item test for different item locations, number of response categories, and subject latent distribution shapes. Note: CTT = classical test theory; IRT = item response theory. $\omega_i = 1.0$ and category thresholds were equally spaced along the latent continuum.
The Reliability of $x$ and $\hat{\theta}$ for Item and Subject Characteristics

Figure 3 included results for a simple example to demonstrate the theoretical differences between $\sigma^2\{x|\theta\}$ and $\sigma^2\{\theta|\theta\}$, which are useful pieces of information for developing tests and instruments. That is, $\sigma^2\{x|\theta\}$ and $\sigma^2\{\theta|\theta\}$ provide applied researchers with an understanding of ranges of $\theta$ values that are best measured with $x$ or $\hat{\theta}$. Moreover, $\sigma^2\{x|\theta\}$ and $\sigma^2\{\theta|\theta\}$ offer researchers information about which measurement framework is most beneficial for various item characteristics and subject populations.

Equations 8 and 17 were used to compare the reliability of $x$ and $\hat{\theta}$ as a function of the number of scale categories, purpose of measurement (i.e., item locations dispersed along the continuum or clustered in a given region), and $\theta$ distribution shape. More specifically, Figure 4 includes $\rho_{00}$ and $\rho_{xx}$ across scale categories (i.e., 2–10 response options) for three types of item locations and four types of distributions for $\theta$. Item discriminations and test length were not manipulated and were fixed at 1.25 and 10, respectively. It is well known that increasing either item discrimination or test length increases reliability, and these parameters were held constant to focus on the other parameters.

To simplify the discussion, the three scenarios assume that items have equally spaced category thresholds. Specifically, the item category thresholds (i.e., the $J c_j$) were equally spaced between $-2.0$ and $2.0$ on the $\theta - b_i$ continuum. Let $c$ be the vector of category thresholds that are defined as $c = 2(2J(J + 1)^{-1} - 1)$ where $J$ is a vector with elements equal to the integers from 1 to $J$. For example, the threshold is zero for items with two scale categories (i.e., $J+1 = 2$) and items with four scale categories have three thresholds at $-1$, $0$, and $1$. Whereas the following discussion assumes the thresholds are equally spaced, researchers can input any set of category thresholds into the R code, which is available at the author’s website.

Let $b$ be a vector of item locations and $I$ be a vector that includes integers from 1 to $I$. The three item location scenarios represent situations where researchers would be interested in measuring $\theta$ values in a narrow range in the lower or upper tails or measuring $\theta$ values across the latent continuum. The item locations for the three scenarios represent the following uniform distributions: $b_1 \sim U(-2.5, -1.5)$ (i.e., $b_1 = 0.5(2I/J+1)^{-1} - 1 - 2$), $b_2 \sim U(-2.0, 2.0)$ (i.e., $b_2 = 2(2I/J+1)^{-1} - 1$), and $b_3 \sim U(1.5, 2.5)$ (i.e., $b_3 = 0.5(2I/J+1)^{-1} - 1 + 2$).

As noted, four subject distributions were examined to evaluate how the density of the population at various points on the latent continuum affected $\rho_{00}$ and $\rho_{xx}$. Specifically, the distributions were negatively skewed ($\gamma_3 = -1.5$, $\gamma_4 = 4.0$), normal ($\gamma_3 = 0$, $\gamma_4 = 0$), symmetric and peaked ($\gamma_3 = -0.5$, $\gamma_4 = 4.0$), and positively skewed ($\gamma_3 = 1.5$, $\gamma_4 = 4.0$).

Figure 4 includes 12 panels corresponding to the three item locations and four subject distribution types. The middle row of panels in Figure 4 demonstrates that CTT and IRT yield similar reliabilities in situations where tests consist of items that are evenly placed along the latent continuum. Given that $\sigma^2\{x|\theta\}$ is smaller in areas with fewer category thresholds, one explanation for the slight advantage of IRT is that CTT CSEMs decline outside of the $(-2, 2)$ range of the items where there are fewer subjects in the skewed and symmetric distributions. More precisely, Figure 3 showed that $\sigma^2\{\theta|\theta\}$ is significantly larger relative to $\sigma^2\{x|\theta\}$ in portions of the latent continuum that has fewer category thresholds. In contrast, $\sigma^2\{x|\theta\}$ is relatively smoother across $\theta$ values and tends to decline in segments where there are fewer category thresholds. Furthermore, in the case where items are located across the latent continuum, $\theta$ is understandably more reliable than $x$ when the latent distribution is more peaked (i.e., positive kurtosis). The middle row of Figure 4 also shows the expected positive relationship that increasing scale categories has on reliability for both $\theta$ and $x$. However, as found in previous research (e.g., see a review of relevant literature in Chafouleas et al., 2009, and Weng, 2004),
the value of an additional response category increases at a decreasing rate and reliability does not increase significantly in value beyond four or five response categories.

In theory, \( x \) is expected to be more reliable than \( ^u \) when the latent distribution is mismatched with item locations. The top and bottom panels of Figure 4 demonstrate the theoretical results discussed in this article concerning the effect that subject and item characteristics have on CTT and IRT reliability. For instance, CTT is superior to IRT whenever the items are located at the extremes of the latent distribution. Furthermore, the difference between \( \rho_{xx} \) and \( \rho_{uu} \) is largest for two category items and approaches zero as the number of categories increase. As discussed earlier, IRT is a tool for measuring specific \( u \) values more precisely. In fact, the difference between \( r_{xx} \) and \( r_{uu} \) is smallest when the majority of the latent distribution overlaps with item locations, which occurs when items are located in the lower portion of the latent continuum and the latent distribution is positively skewed (e.g., \( b_i \sim U(−2.5, −1.5) \), \( \gamma_3 = 1.5 \), and \( \gamma_4 = 4.0 \)) or when items are in the upper tail and the distribution is negatively skewed (i.e., \( b_i \sim U(1.5, 2.5) \), \( \gamma_3 = −1.5 \), and \( \gamma_4 = 4.0 \)).

In short, the results in this section provide new information about the influence of item and subject characteristics on the relative reliability of \( x \) and \( ^u \). In fact, the findings reflect the expected results given the nature of \( s^2_f \) and \( s^2_f^u \). That is, total scores were more reliable whenever items were mismatched with the location of the latent \( u \) distribution. The following section addresses another important concern related to the accuracy of linear approximations of polytomous items.

### Appropriateness of a Linear Approximation of Polytomous Items

One approach for modeling polytomous items is to use a linear approximation of the relationship between \( u \) and \( u_i \) (e.g., a common factor model, Culpepper, 2012b). As noted in Equation 13, the relationship between \( u \) and \( u_i \) is nonlinear. Ferrando (2009) and Mellenbergh (1996) noted that a linear relationship tends to provide a reasonable approximation when item discrimination indices are relatively low and there are five or more response categories. The purpose of this section is to explore the appropriateness of linear approximations of polytomous items in greater detail. This section includes two subsections. The first subsection describes a measure for quantifying the appropriateness of a linear approximation, whereas the second subsection presents findings about the accuracy of a linear approximation for different item characteristics (e.g., items locations, number of response categories, and item discriminations) and subject characteristics (e.g., the skewness and kurtosis of the latent distribution).

#### A Measure for the Appropriateness of a Linear Approximation

Let \( \lambda_i(\theta) \) be the loading relating latent \( \theta \) to observed \( u_i \) for a given \( \theta \). The relationship between \( \theta \) and \( u_i \) for a specific value of \( \theta \) is the first derivative of Equation 9 with respect to \( \theta \):

\[
\lambda_i(\theta) = \frac{\partial E\{u_i|\theta\}}{\partial \theta} = \sum_{j=1}^{j+1} \frac{\partial P_{u}(u_i=j|\theta)}{\partial \theta}.
\] (18)

The relationship between \( u_i \) and \( \theta \) will be larger for some values of \( \theta \) and near zero at the extremes of the \( \theta \) continuum (i.e., portions of the latent continuum where \( E\{u_i|\theta\} \) plateaus). A linear approximation is dependent on the characteristics of the examinee population. The expected loading for a given population of subjects can be found by averaging \( \lambda_i(\theta) \) over the distribution of \( \theta \):
\[ E\{\chi_i(\theta)\} = E\left\{ \frac{\partial E\{u_i|\theta\}}{\partial \theta} \right\} = \sum_{j=1}^{J+1} \int_{-\infty}^{\infty} \frac{\partial P_j(u_i=j|\theta)}{\partial \theta} \phi(\theta|\Omega) d\theta. \] (19)

The function that constrains the relationship between \( \theta \) and \( u_i \) to be linear for a given population of examinees is

\[ L\{u_i|\theta\} = E\{u_i\} + E\{\chi_i(\theta)\}(\theta - \mu), \] (20)

where \( E\{u_i\} = \int E\{u_i|\theta\} \phi(\theta|\Omega) d\theta, \mu = E\{\theta\} \), and \( L\{u_i|\theta\} \) denote the linear approximation of the relationship between \( \theta \) and \( u_i \). Also, note that \( L\{u_i|\theta\} \) has the same form as the least squares projection where \( L\{u_i|\theta\} \) passes through the centroid for \( \theta \) and \( u_i (\mu, E\{u_i\}) \) and the relationship is the expected slope for a given population of examinees, \( E\{\chi_i(\theta)\} \).

Let \( d_i = u_i - L\{u_i|\theta\} \) be the error when predicting the observed item \( u_i \) with a linear approximation. Unlike, \( E\{e_i|\theta\}, E\{d_i|\theta\} \neq 0 \),

\[ E\{d_i|\theta\} = E\{u_i - L\{u_i|\theta\}|\theta\} = E\{u_i|\theta\} - L\{u_i|\theta\}. \] (21)

That is, \( d_i \) will be biased for certain values of \( \theta \) but \( E\{E\{d_i|\theta\}\} = 0 \) across the \( \theta \) range for a given population. The bias in \( d_i \) does not translate into larger conditional variances for \( L\{u_i|\theta\} \). In fact, \( \sigma^2\{d_i|\theta\} = \sigma^2\{u_i|\theta\} \),

\[ \sigma^2\{d_i|\theta\} = E\left\{ (d_i - E\{d_i|\theta\})^2|\theta \right\} = E\left\{ (u_i - E\{u_i|\theta\})^2|\theta \right\} = \sigma^2\{u_i|\theta\}. \] (22)

Furthermore, the bias in \( d_i \) does not introduce dependence among residuals. Let \( d_h = u_h - L\{u_h|\theta\} \) be a residual when linearly approximating \( u_h \). The covariance between errors is

\[ \sigma\{d_i, d_h|\theta\} = E\left\{ (d_i - E\{d_i|\theta\})(d_h - E\{d_h|\theta\})|\theta \right\} = E\left\{ (u_i - E\{u_i|\theta\})(u_h - E\{u_h|\theta\})|\theta \right\} = 0, \] (23)

where the last inequality was obtained by recalling that the assumption of local independence implies \( E\{e_i e_j|\theta\} = 0 \) as shown in Equation 12.

Ferrando (2009) described a measure of the appropriateness of a linear model, which is based on restricting observed scores to fall within feasible ranges (e.g., 1 and \( J + 1 \) for a polytomous item with \( J + 1 \) categories). More precisely, \( 1 < L\{u_i|\theta\} < J + 1 \) implies that a linear approximation is appropriate for \( \theta \) scores within the following range:

\[ \frac{1 - E\{u_i\}}{E\{\chi_i(\theta)\}} + \mu < \theta < \frac{J + 1 - E\{u_i\}}{E\{\chi_i(\theta)\}} + \mu. \] (24)

Following Ferrando, the floor and ceiling indices for the appropriateness of a linear approximation of \( u_i \) are \( \delta_0 = (1 - E\{u_i\})/E\{\chi_i(\theta)\} + \mu \) and \( \delta_1 = (J + 1 - E\{x\})/E\{\chi_i(\theta)\} + \mu \), respectively. Ferrando’s measure of appropriateness is the proportion of examinees with \( \theta \) scores in the viable range:

\[ P(\delta_0 < \theta < \delta_1) = \int_{\delta_0}^{\delta_1} \phi(\theta|\Omega) d\theta. \] (25)

Consider the examples presented in Figure 5, which compares \( E\{u_i|\theta\} \) with \( L\{u_i|\theta\} \) for three items with \( h_i = (-2, 0, 2) \) and the following parameter values held constant, \( \alpha_i = 2 \): for all \( i \), four response categories (\( J + 1 = 4 \)), category thresholds of \( c = -1.6, 0, 1.6, 0 \), and \( \Omega = (0, 1, 0, 0) \). The
middle panel in Figure 5 includes the case where \( b_1 = 0 \). Figure 5 shows that \( E\{u_i|\theta\} \) and \( L\{u_i|\theta\} \) deviate at the ends of the latent continuum. However, given that the distribution of \( \theta \) is standard normal, nearly all of the subjects (i.e., 98.9%) have approximated scores between \( \delta_1 \) and \( \delta_2 \). The top and bottom rows of Figure 5 include examples of items that are located at \( -2 \) and 2, respectively. In fact, when items are located at the extremes of the \( \theta \) continuum \( P(\delta_0 < \theta < \delta_1) = 0.876\% \), which implies that 12.4% of subjects have approximated scores outside the viable bounds. The examples in Figure 5 demonstrate that item location and the shape of the latent distribution affect the appropriateness of linear approximations. For instance, Figure 5 provides insight that \( P(\delta_0 < \theta < \delta_1) \) is larger in value if items are located near the mean of a standard normal distribution.

**Impact of Item and Subject Characteristics on the Appropriateness of a Linear Approximation**

The previous subsection discussed \( P(\delta_0 < \theta < \delta_1) \) as a measure of the appropriateness of a linear model. The purpose of this subsection is to present more general findings about the role of item
and subject characteristics on the accuracy of a linear approximation. Specifically, this subsection discusses the appropriateness of a linear approximation of $u_i$ for 12 scenarios as defined by three response category scenarios ($J + 1 = 2-10$) and four $\theta$ distribution scenarios (e.g., the same distributions examined in Figure 4) with the addition that the item discrimination indices are compared for $\alpha_i = 0.5, 1.0, 1.5$, and 2 and category thresholds are defined as discussed for the scenarios in Figure 4.

Figures 6 and 7 plot $P(\delta_0 < \theta < \delta_1)$ against item locations (i.e., $b_i$) and provide evidence about circumstances where a linear approximation is appropriate. Note that the rows of Figures 6 and 7 correspond to the number of response categories (i.e., $J + 1$), whereas columns relate to distribution shape. Comparing rows demonstrates that $P(\delta_0 < \theta < \delta_1)$ increases as the number of response options increases; however, the accuracy of a linear approximation does not improve significantly beyond four response categories. For instance, a linear approximation is appropriate and $P(\delta_0 < \theta < \delta_1) > 0.90$ for items with five or more response categories that are located within one standard deviation of the mean. Furthermore, larger item discriminations have a negative effect on $P(\delta_0 < \theta < \delta_1)$ and the number of response categories and item discrimination has an interactive effect where $P(\delta_0 < \theta < \delta_1)$ declines more as $\alpha_i$ increases and there are fewer response categories. Moreover, Figure 6 illustrates that a linear approximation is appropriate for as few as two response options when $\alpha_i = 0.5$ or whenever $\alpha_i = 1.0$ and items are located in the middle of the distribution.

Comparing columns provides an indication of the effect of latent distribution shape and item location on $P(\delta_0 < \theta < \delta_1)$. Specifically, the relationship between $b_i$ and $P(\delta_0 < \theta < \delta_1)$ reflects the shape of the latent distribution, and $P(\delta_0 < \theta < \delta_1)$ is smallest when $b_i$ is located in the tails of the latent distribution. For instance, the $P(\delta_0 < \theta < \delta_1)$ curves for the normal distribution appear more bell shaped when the distribution is normal as opposed to peaked (i.e., $\kappa_3 = 0, \kappa_4 = 4$). In fact, $P(\delta_0 < \theta < \delta_1)$ is smaller when $\theta \sim N(0, 1)$, ceteris paribus. The relationship between item location and $P(\delta_0 < \theta < \delta_1)$ is cubic when the latent distribution is skewed, and $P(\delta_0 < \theta < \delta_1)$ is smaller in skewed distributions in segments where there is less density in the latent distribution (e.g., the right tail if $\kappa_3 = -1.5$ and the left tail if $\kappa_3 = 1.5$).

Figure 6 also demonstrates the effect of item discrimination on $P(\delta_0 < \theta < \delta_1)$. Ferrando (2009) noted that a linear approximation is most appropriate when $\alpha_i$ is smaller. In fact, Ferrando’s recommendations are supported given that a linear approximation is expected to be appropriate for almost all subjects and item locations when $\alpha_i = 0.50$. Figures 6 and 7 also show conditions where linear approximations are appropriate even when items are more discriminating (e.g., $\alpha_i = 2$). For example, $P(\delta_0 < \theta < \delta_1)$ is larger when $J + 1 \geq 3$ and items are located near the middle of the latent distribution. $P(\delta_0 < \theta < \delta_1)$ declines as $\alpha_i$ increases for all of the scenarios included in Figure 6 albeit at different rates. In short, the results in Figures 6 and 7 imply that linear approximations are best for items that measure $\theta$ near the central portion of the latent distribution. Furthermore, $L[u_i|\theta]$ is least accurate for highly discriminating items, and a linear approximation is inaccurate even if there are more than four response categories when items are located in the tails of the latent distribution.

**Discussion**

The findings in this study offer guidance to applied researchers interested in the construction and analysis of polytomous items. In short, this article presented new theoretical results concerning item and subject characteristics that affect the relative precision and reliability of $x$ and $\theta$. This section summarizes the findings for psychometricians and applied researchers and offers concluding remarks.
Figure 6. The proportion of viable scores from a linear approximation of polytomous items with two to four response categories by item location and discrimination, and subject latent distribution shapes.

Note: Category thresholds were equally spaced along the latent continuum.
Figure 7. The proportion of viable scores from a linear approximation of polytomous items with five to seven response categories by item location and discrimination, and subject latent distribution shapes.

Note: Category thresholds were equally spaced along the latent continuum.
This article studied the reliability of two alternative test scoring approaches: a CTT total score versus an IRT θ estimate. The majority of applied researchers in education and psychology are familiar with x, whereas fewer have knowledge about θ and IRT. Consequently, it is important to offer applied researchers information about the relative merits of x and θ. This article offered additional insights into fundamental differences between x and θ. One salient factor studied in this article was the interactive effect that item locations and θ distribution shape had on ρ_x and ρ_θθ. Suppose the purpose of testing is to accept high-scoring examinees into an institution as is the case with examinations for actuarial sciences where the latent distribution is either normal or positively skewed. For instance, the first actuarial exam includes items clustered in the upper portion of the θ continuum to identify those students who are most competent in the foundations of calculus and probability theory. The scores for examinees who are near the passing cutoff are measured more precisely than those who are in the middle or lower portions of the θ distribution as indicated by smaller values for σ^2{x|θ}. In contrast, the results in this article show that σ^2{x|θ} is larger near the cutoff score and would be relatively smaller for θ values in the lower and middle portions of the examinee distribution. Certainly, decision makers would prefer IRT scoring versus a total score in this instance, because the purpose of measurement is to evaluate whether test takers exceed a minimum proficiency level. However, secondary users of the actuarial test score data should prefer total scores, because, as indicated in Figure 4, x tends to be more reliable for a population of subjects than θ if items are difficult and located in the tails of the test-taker population distribution. For instance, some secondary users may want to gather validity evidence and correlate examinee scores with other indicators, such as undergraduate or graduate grade point average (Culpepper, 2010; Culpepper & Davenport, 2009), other aptitude tests, or job performance (Aguinis, Culpepper, & Pierce, 2010). In these instances, the validity coefficient associated with θ would be smaller than the coefficient for x because ρ_x > ρ_θθ.

In addition, the results provide evidence that CTT and IRT scoring methods perform similarly when the goal of testing is to measure a construct across values of a latent continuum. For example, federal and state testing programs measure what students know in relation to given standards. The results in Figure 4 imply that ρ_x and ρ_θθ are similar and testing programs could report total scores, which may be easier to explain to certain stakeholders (e.g., teachers, parents, and students).

The results in this article also offer recommendations for the construction of scales that include polytomous items, such as educational or employment performance assessments, behavioral ratings, or affective measurements. In either case, researchers may prefer x to θ if the purpose of instrument development is to conduct correlational research rather than to measure specific trait levels. However, x should probably only be preferred to θ when researchers fit simple models rather than more complicated interactive models (Embretson, 1996; Kang & Waller, 2005; Morse et al., 2012).

In addition, the findings in this article provide information about the optimal number of item response options. The results in Figure 4 examined the reliability of total scores and IRT estimates across a range of parameter values for the number of scale categories, item locations, and the shape of the latent distribution. The findings in Figure 4 imply that using more than five or six scale categories does not significantly improve the reliability of x or θ regardless of the shape of the latent distribution or location of items. However, it is important to note that adding an additional scale value had a larger effect on ρ_θθ than on ρ_x.

Another relevant finding for applied researchers (and methodologists) relates to the appropriateness of a linear approximation of polytomous items. The results in this article confirm arguments in previous research (Ferrando, 2009) that linear approximations are more accurate for less discriminating items and items with more response categories. Additional findings suggest...
that linear approximations of polytomous items seem appropriate for items that measure trait levels in denser segments of the latent distribution, and linear approximations were least appropriate for items located in the tails of the $\theta$ distribution. In contrast to previous research, the results provided new evidence that researchers need to consider item locations, in addition to item discriminations and the number of response categories, when employing linear approximations.

Last, another contribution of this study is the availability of the associated R code. More specifically, researchers can use the R code when designing instruments in an effort to evaluate conditions where total scores and IRT scores are more or less reliable and precise. Furthermore, the R code has pedagogical value, as well, for computational applications of the theoretical results.

There are several directions for future research to build on this study. First, this article offers recommendations for researchers who are interested in using $x$ as a measure of $\theta$ in applied research. Specifically, designing a reliable $x$ requires the inclusion of items located at the boundaries of the $\theta$ range of interest. For instance, if the measurement goal is to distinguish high versus low scorers on some trait, the results pertaining to CTT CSEMs dictate that the items should be located in the middle of the distribution, because low and high scorers will be measured more precisely. Likewise, items should be located around a cutscore (and not at the cut-score) if the purpose of measurement is to make inferences about whether examinees exceed or fall below some minimum proficiency level. The behavior of CSEMs under CTT is counterintuitive, because, in contrast, item design under the IRT framework dictates that developers should write items that are specific to certain $\theta$ levels and measurement purposes. Additional research is needed to understand differences in optimal test assembly (H. Chang & Ying, 2009) within the IRT and CTT frameworks.

Second, there could be benefits in revisiting some topics in modern IRT within the context of CTT. For instance, there may be new insights available by reexamining topics in CTT, such as computer adaptive testing (H. Chang & Ying, 1996) or equating techniques (Kolen & Brennan, 2004), which could lead to new methodologies, refinements of existing approaches, or other unanticipated discoveries.

Third, researchers could extend the results in this article to understand the impact of using $x$ and $\theta$. Specifically, researchers use total scores as dependent variables and predictors in every subdiscipline of psychology and education. Despite the widespread use of total scores, few methodological studies have examined the impact of using total scores on the power and Type I error rates of tests that researchers employ (Embretson, 1996; Kang & Waller, 2005; Morse et al., 2012). Furthermore, with the exception of Embretson (1996), previous studies utilized Monte Carlo techniques that are limited by the parameter values studied. Consequently, future analytic explorations could provide additional insights into the effect of using total scores, and future research should accordingly examine the effect that the number of scale categories, the shape of the latent distribution, and IRT parameters have on the performance of commonly used statistical tests (Culpepper, 2012a; Culpepper & Aguinis, 2011).

Fourth, this study examined the theoretical reliability of $x$ and $\theta$ using a polytomous IRT model. As one anonymous reviewer noted, this article addressed reliability from a mathematical perspective and does not consider factors related to subjects’ cognitive decision making. For example, this article did not address issues related to category labels; however, previous research identified a causal effect of scale labels, category position, and rating scale intensity and length on certain observed item characteristics (Dunham & Davison, 1991; Lam & Stevens, 1994; Murphy & Constans, 1987). For instance, existing evidence suggests that scale labels can affect the observed item means, but there is less evidence that manipulating category labels alters observed item variances (L. Chang, 1997; Dunham & Davison, 1991). Moreover,
positive-packed scales tend to affect item means (Dunham & Davison, 1991; Lam & Kolic, 2008), and semantic compatibility of category labels improves reliability (Lam & Kolic, 2008). Most of the previous literature on category labels used CTT or generalizability theory, and additional empirical research is needed to understand how category labeling decisions affect polytomous IRT item parameters (e.g., item locations, discriminations, and category thresholds) and consequently alter the reliability of $x$ and $\theta$. Future research may identify relationships between category labels, cognitive decision making, and IRT parameters, and the results in this article provide mathematical arguments for describing how subsequent changes in IRT parameters affect test score reliability.

In conclusion, this study presented new results concerning the relative reliability and precision of total scores and IRT scores. The derivations in this article offer the most extensive analysis of the reliability of total scores by linking parameters of polytomous IRT models with CTT. In addition, new equations were discussed that described the CTT CSEM to provide new conceptual understanding of differences in the precision of scores estimated within the CTT and IRT frameworks.

## Appendix

### Fleishman’s Power Transformation (PT) Method Probability Density Function (PDF)

Fleishman (1978) developed a PT method for generating nonnormal univariate random variables. The PT method uses the following function to transform a standard normal random variable, $y$, into a metric with a given $\kappa_3$ and $\kappa_4$,

$$\theta = f(y|\omega) = \sum_{r=1}^{4} \omega_r y^{r-1},$$

(A1)

where the vector of Fleishman coefficients, $\omega = (\omega_1, \omega_2, \omega_3, \omega_4)$, is identified so that $\theta$ has a predetermined mean, variance, skewness, and kurtosis. Additional research extended Fleishman’s method to higher order PTs (Headrick, 2002) and multivariate circumstances (Headrick & Sawilosky, 1999; Lyhagen, 2008; Vale & Maurelli, 1983).

One critique of Fleishman’s method was that there was no known probability distribution for variables generated using Fleishman’s technique (Tadikamalla, 1980). Recent research derived the PDF for Fleishman’s PT method (Headrick & Kowalchuk, 2007) and showed that the PDF for Fleishman’s PT method is

$$\phi(\theta|\Omega) = \frac{\phi(f^{-1}(\theta|\omega))}{\sigma f'(f^{-1}(\theta|\omega)|\omega)},$$

(A2)

where $\phi(\theta)$ is the standard normal distribution, $f^{-1}(\theta|\omega)$ is the inverse of Equation A1 that specifies values of $y$ as a function of $\theta$, and $f'(y|\omega)$ is the derivative of Equation A1 with respect to $y$. An expression for $f^{-1}(\theta|\omega)$ can be found using Cardano’s formula for the inverse of a cubic polynomial. The equation for the only real root of $f^{-1}(\theta|\omega)$ is defined as

$$y = f^{-1}(\theta|\omega) = \left[q + \sqrt{q^2 + (r - p^2)^3}\right]^{1/3} - \left[q - \sqrt{q^2 + (r - p^2)^3}\right]^{1/3} + p,$$

(A3)

where the second cube root is the absolute value of $q - \sqrt{q^2 + (r - p^2)^3}$. Furthermore, $p$, $q$, and $r$ are a function of $\omega$ and $\theta$ as shown in the following:
This article uses $\varphi(\theta|\Omega)$ to understand the impact of nonnormal latent distributions on the reliability of $x$ and $\theta$. Specifically, $\omega$ is first identified for a given $\kappa_3$ and $\kappa_4$, and the $p$, $q$, and $r$ in Equation A4 are computed to find $f^{-1}(\theta|\omega)$ to use in Equation A3.

It is important to note that Fleishman’s PT method yields a valid PDF only when $\theta$ is a monotonically increasing function of $y$ (i.e., $f'(y|\omega)$ must be positive for all values of $y$). Headrick and Kowalchuk (2007) proved that the PT method produces a valid PDF if the Fleishman coefficients satisfy the following constraints:

\[
\omega_4 > \frac{1}{5} \sqrt{\frac{5 + 7\omega_2^2}{3}} - \frac{2}{5} \omega_2, \quad 0 < \omega_2 < 1.
\]

Consequently, Fleishman’s PT method can only be used to study the impact of nonnormal latent distributions when $|\kappa_3| < 4.5$.

Declaration of Conflicting Interests
The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

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

References


