This article provides analytic evaluations of population true-score measures for binary items given their item response theory (IRT) calibration. Under the assumption of normal trait distribution, the expected values of marginalized true scores, error variance, true score variance, and reliability for norm-referenced and criterion-referenced interpretations are presented as a function of the item parameters. The proposed formulas have methodological and computational value in bridging concepts of IRT and true score theory. They provide information about the individual contribution of IRT calibrated items to marginal true-score measures and may have valuable applications in test development and analysis. For example, given a bank of IRT calibrated items, one can select binary items to develop a test with known true-score characteristics prior to administering the test (without information about raw scores or trait scores.) Calculations with the proposed formulas are easy to perform using basic statistical programs, spreadsheet programs, or even hand-held calculators. Index terms: true score theory, item response theory, reliability.

True-score measures and reliability are used in substantive and measurement studies even when item response theory (IRT) information about items and persons is available (e.g., with standardized tests). Traditionally, such measures represent a common focal point between test developers and practitioners as they place the scores and their accuracy in the original scale of measurement (e.g., number-right [NR] score). True (or domain) scores are readily interpretable; for example, when pass-fail decisions are made, a cutting score is typically set on the domain-score scale (e.g., Hambleton, Swaminathan, & Rogers, 1991, p. 85). Therefore, it seems totally appropriate to argue that IRT estimates and classical estimates of scores and their reliability are not mutually exclusive and may coexist in making adequate interpretations and decisions based on test data. Combining IRT information about trait scores with readily interpretable true-score information will positively affect the quality of test development and analysis. This, however, requires better understanding of the relationships between IRT and classical concepts from methodological and technical perspectives. As a step in this direction, this article investigates relationships between expected values of marginal true-score measures and IRT parameters of binary items. Analytic expressions of such relationships can be useful in test development and analysis from both methodological and technical perspectives.
Before presenting the theoretical framework for bridging true-score measures to IRT item parameters, an important clarification should be made. As is known, the accuracy of measurement in IRT varies across the levels of a latent trait, $\theta$, that underlies the persons’ responses on test items. One should, however, distinguish the conditional error variance at $\theta$ associated with the estimate of the trait score, $\hat{\theta}$, from the conditional error variance at $\theta$ associated with the number-correct score, $X$. Thus, the IRT conditional error variance at $\theta$, $\sigma^2_{\theta|\theta}$, inversely related to the information provided by the test at $\theta$ (Birnbaum, 1968), is not to be confused with the conditional raw-score variance at $\theta$, $\sigma^2_{\theta|X}$. The expected value of the latter (when $\theta$ varies from -4 to 4) is the marginal error variance for the number-correct score (e.g., Lord, 1980), whereas the expected value of the former is referred to as marginal measurement error variance in IRT (Green, Bock, Humphreys, Linn, & Reckase, 1984). The marginal reliability in IRT is used, for example, as an overall index of precision in computerized adaptive testing for comparison with the classical internal-consistency reliability estimated for paper-and-pencil forms (Green et al., 1984; Thissen, 1990). Such comparisons, however, require more accurate evaluations of the population reliability for paper-and-pencil forms than those provided by sample-based empirical indexes such as Cronbach’s coefficient alpha (Cronbach, 1951). Some additional comments on this issue are provided in the Discussion section.

The formulas proposed in this article, derived under the assumption of normal trait distribution, can be very useful in comparing and bridging IRT and classical measures in test development and analysis. For example, given a bank of IRT calibrated binary items, one can develop a test (e.g., for follow-up measurements in longitudinal studies) with marginal true-score characteristics and reliability known prior to data collection.

Theoretical framework

Let $P_i(\theta)$ be the probability for correct response on item $i$ for a person with a trait score $\theta$ under an appropriate IRT model: one-parameter (1PLM), two-parameter (2PLM), or three-parameter (3PLM) logistic model (Birnbaum, 1968). Specifically, with the 2PLM,

$$P_i(\theta) = \frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]},$$

where $a_i$ is the item discrimination, $b_i$ is the item difficulty, and $D$ is a scaling factor (with $D = 1.7$, and values of $P_i(\theta)$ for the 2PLM and the two-parameter normal ogive model differ in absolute value by less than 0.01 for any value of $\theta$). With $a_i = 1$, equation (1) generates $P_i(\theta)$ values under the 1PLM. The equation for $P_i(\theta)$ with the 3PLM is provided later in this article.

It should be noted that if $u_i$ is the binary score on item $i$, $P_i(\theta)$ is the expected mean of $u_i$ for a person with a trait score $\theta$; that is, $P_i(\theta) = \mathbb{E}(u_i)$ is the person’s true score on item $i$ at $\theta$. The marginal probability of correct responses on item $i$ is then

$$\pi_i = \int_{-\infty}^{\infty} P_i(\theta) \Phi(\theta) d\theta,$$

where $\Phi(\theta)$ is the probability density function (pdf) for the trait distribution. The integration is from -4 to 4 since the ability, $\theta$, is not limited in the theoretical framework of IRT. Thus, $\pi_i$ is the expected proportion of correct responses on item $i$ for a population of examinees with a trait distribution $\Phi(\theta)$. For a test of $n$ binary items, then, the expected NR score for this population of examinees is

$$\mu = \sum_{i=1}^{n} \pi_i.$$
Also, as \( u_i \) is a binary score and \( P_i(\theta) \) is its expected value at \( \theta \), the binomial variance of \( u_i \) at \( \theta \) is \( \sigma^2(u_i|\theta) = P_i(\theta)[1 - P_i(\theta)] \). Under the assumptions in classical test theory, \( u_i = \tau_i + e_i \), where \( \tau_i \) is the true score on item \( i \) and \( e_i \) is a random error, and \( \sigma^2(u_i) = \sigma^2(e_i) \). Therefore, the conditional item error variance at \( \theta \) is \( \sigma^2(e_i|\theta) = P_i(\theta)[1 - P_i(\theta)] \). The expected item error variance for a population with a trait distribution \( \phi(\theta) \) is then

\[
\sigma^2(e_i) = \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)] \phi(\theta) d\theta. \tag{4}
\]

At the test level, assuming no correlated errors, the expected error variance for the NR score is

\[
\sigma^2_e = \sum_{i=1}^{n} \sigma^2(e_i). \tag{5}
\]

The true score variance for the NR score is usually presented (e.g., May & Nicewander, 1993) as

\[
\sigma^2_{\tau} = \int_{-\infty}^{\infty} [n \bar{P}(\theta)]^2 \phi(\theta) d\theta - \left[ \int_{-\infty}^{\infty} n \bar{P}(\theta) \phi(\theta) d\theta \right]^2, \tag{6}
\]

where \( \bar{P}(\theta) \) is the mean of \( P_i(\theta) \) at \( \theta \) (\( i = 1, ..., n \)).

Previous research provides limited applications of equations (2), (4), or (6) using, for example, Gaussian quadrature (Bock & Lieberman, 1970), but analytic approximations are not provided. For example, comparing reliability for NR scores and percentile ranks, May and Nicewander (1993) evaluated the integrals in equations (4) and (6) using the Simpson's Rule with 100 points on the \( \theta \) interval from -5 to 5 after approximating the compound binomial distributions of the number-correct scores. This article takes a different approach and provides formulas for marginalized true-score measures at the item level thus making it possible to determine (and control) the contribution of individual items to the values of \( \mu \), \( \sigma^2_e \), \( \sigma^2_{\tau} \), and reliability indexes at the test level. Comments on the advantages of the proposed formula over direct brute-force quadrature integrations are provided in the Discussion section.

Given the IRT calibration of binary items, marginalized true-score measures for a normal trait distribution are evaluated in this article at both item and test levels. For individual items, formulas are provided for the expected item score \( (\pi_i) \), item error variance \( (\sigma^2(e_i)) \), item true variance \( (\sigma^2(\tau_i)) \), and item reliability \( (\rho_i) \). At the test level, formulas are provided for the expected NR score \( (\mu) \), domain score \( (\pi) \), error variance \( (\sigma^2_e) \), true score variance \( (\sigma^2_{\tau}) \), reliability \( (\rho) \), and dependability index \( (\Phi(\lambda)) \) for criterion-referenced interpretations based on a cutting domain score, \( \lambda \). For items calibrated with the 2PLM, \( \pi_i \) and \( \sigma^2(e_i) \) are evaluated through approximation formulas (with a negligible approximation error). All other true-score measures at both item and test levels are represented (explicitly or implicitly) as exact analytic functions of \( \pi_i \) and \( \sigma^2(e_i) \). The next sections provide formulas for binary items calibrated with the 2PLM, 3PLM, and 1PLM and two illustrative examples. The mathematical derivations of the formulas are given in Appendix A. The calculations with the proposed formulas are facilitated by the use of a SPSS syntax (SPSS, 2002) provided in Appendix B.

**Formulas for Binary Items Calibrated with the 2PLM**

**Expected Item Score**

The expected item score, \( \pi_i \), is estimated here through an approximate evaluation of the integral in equation (2). In classical test theory, the empirical estimate of \( \pi_i \) is referred to as *item difficulty*.
(although it is, in fact, the easiness of the item). As proven in Appendix A, $\pi_i$ can be represented as a function of the IRT item parameters ($a_i$ and $b_i$):

$$\pi_i = \frac{1 - \text{erf}(X_i)}{2},$$  \hspace{1cm} (7)

where $X_i = a_i b_i / \sqrt{2(1 + a_i^2)}$, and erf is a known mathematics function called the error function. With a relatively simple approximation provided by Hastings (1955, p. 185), the error function (for $X_i > 0$) can be evaluated with an absolute error smaller than 0.0005 as

$$\text{erf}(X) = 1 - \left(1 + m_1 X + m_2 X^2 + m_3 X^3 + m_4 X^4\right)^{-4},$$  \hspace{1cm} (8)

where $m_1 = .278393$, $m_2 = .230389$, $m_3 = .000972$, and $m_4 = .078108$. When $X < 0$, one can use that $\text{erf}(-X) = -\text{erf}(X)$. It should be also noted that the erf(X) is directly executable with computer programs for mathematics (e.g., MATLAB 5.3; MathWorks, Inc., 1999).

**Expected Item Error Variance**

The expected marginal error variance for an item $i$ is estimated through an approximate evaluation of the integral in equation (4). With $\varphi(\theta)$ for the standard normal distribution and $D = 1.7$ with the 2PLM, equation (4) becomes

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} \frac{\exp[1.7a_i(\theta - b_i)]}{\left(1 + \exp[1.7a_i(\theta - b_i)]\right)^2} \left(\frac{1}{\sqrt{2\pi}} \exp(-.5\theta^2)\right) d\theta$$  \hspace{1cm} (9)

Because a closed form evaluation of the integral in equation (9) does not exist, an approximation was developed in two steps. First, using the computer program MATLAB 5.3 (MathWorks, Inc., 1999), quadrature method evaluations were obtained for practically occurring values from 0 to 3 for the item discrimination, $a_i$, and from -6 to 6 for the item difficulty, $b_i$, with a step of 0.01 on the logit scale. Second, the results were tabulated and approximated using the three-parameter Gaussian function with the regression wizard of the computer program SigmaPlot 5.0 (SPSS, 1998). The resulting approximation formula is

$$\sigma^2(e_i) = m_i \exp[-0.5(b_i / d_i)^2],$$  \hspace{1cm} (10)

where $b_i$ is the item difficulty, whereas $m_i$ and $d_i$ depend on the item discrimination as follows:

$$m_i = 0.2646 - 0.118 a_i + 0.0187 a_i^2;$$
$$d_i = 0.7427 + 0.7081/a_i + 0.0074/a_i^2.$$

Depending on the values $a_i$ and $b_i$, the error of approximation with formula (10) varies from 0 to 0.005 in absolute value (with a mean of 0.001 and a standard deviation of 0.001). As one can see from formula (10) (graphical illustration in Figure 1), the item error variance is an even function of $b_i$ for fixed values of $a_i$. In other words, the value of $\sigma^2(e_i)$ is the same for $b_i$ and $-b_i$ when the value of $a_i$ is fixed. As Figure 1 also shows, larger errors occur with average difficulty items, and smaller errors occur with easy or difficult items. It should be noted that $\sigma^2(e_i)$ is an additive error variance component of the expected (marginal) error variance for the NR score, $\sigma_e^2$. 
Expected Error Variance of a Binary Item Calibrated With the Two-Parameter Logistic Model (2PLM), $\sigma^2(e_i)$, as a Function of Its Discrimination ($a_i$) and Difficulty ($b_i$). $\sigma^2(e_i)$ is an (Additive) error Variance component of the expected error variance for the number-right (NR) score, $\sigma^2$. 

Figure 1

Expected Item True Variance

As proven in Appendix A, the expected item true variance can be represented as an exact function of the expected item score and expected item error variance:

$$\sigma^2(\tau_i) = \pi_i (1 - \pi_i) - \sigma^2(e_i).$$

(11)

It should be noted also that the derivation of formula (11) is the same with any IRT model (1PLM, 2PLM, or 3PLM) and any (not necessarily normal) trait distribution (see Appendix A).

Reliability of Item Score

In classical test theory, the score reliability of a binary item $i$ is empirically estimated with the product $s_i r_{ik}$, where $s_i$ is the item score standard deviation, and $r_{ik}$ is the point-biserial correlation between the item score and total test score (e.g., Allen & Yen, 1979, p. 124). This article uses the ratio "item true variance to observed item variance" for the evaluation of item score reliability ($\rho_i$).
Thus, given the IRT calibration of binary items, the reliability of the item score is evaluated with

$$\rho_{ii} = \frac{\sigma^2(\tau_i)}{\sigma^2(\tau_i) + \sigma^2(e_i)},$$

(12)

where $\sigma^2(e_i)$ and $\sigma^2(\tau_i)$ are obtained with formulas (10) and (11), respectively. It should be noted that $s_i r_{xx}$ is an empirical estimate of the reliability of sample scores on item $i$, whereas $\rho_{ii}$ (with equation (12)) is a theoretical evaluation of item score reliability for a population of examinees. Information about the reliability of the item score can be particularly useful when the purpose is to select items that maximize the internal consistency reliability of test scores (e.g., Allen & Yen, 1979, p. 125).

**Expected NR Score**

Given the expected item score, $\pi_i$, of each item in a test of $n$ binary items, the expected NR score on the test is

$$\mu = \sum_{i=1}^{n} \pi_i.$$  

(13)

**Expected Error Variance for the NR Score**

Given the expected item error variance, $\sigma^2(e_i)$, for each item in a test of $n$ binary items, the expected error variance for the NR score on the test is

$$\sigma^2_e = \sum_{i=1}^{n} \sigma^2(e_i).$$  

(14)

**Expected True Score Variance for the NR Score**

As proven in Appendix A, the expected true score variance for the NR score on a test of $n$ binary items is

$$\sigma^2_t = \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{[\pi_i (1 - \pi_i) - \sigma^2(e_i)] [\pi_j (1 - \pi_j) - \sigma^2(e_j)]},$$  

(15)

where $\pi_i$ and $\sigma^2(e_i)$ (or $\pi_j$ and $\sigma^2(e_j)$) are obtained with formulas (7) and (10), respectively.

**Reliability of Test Score**

Under the true-score model (Lord & Novick, 1968), the reliability of test score is

$$\rho_{xx} = \frac{\sigma^2_t}{\sigma^2_t + \sigma^2_e}.$$  

(16)

In this article, the theoretical value of $\rho_{xx}$ for the NR score on a test of $n$ binary items is evaluated by replacing $\sigma^2_t$ and $\sigma^2_e$ in formula (16) with their values obtained through formula (5) (with $\sigma^2(e_i)$ obtained via formula (10)) and formula (15), respectively.
Dependability Index

Brennan and Kane (1977) introduced a dependability index, $\Phi(\lambda)$, for criterion-referenced interpretations in the framework of generalizability theory (GT) (e.g., Brennan, 1983):

$$\Phi(\lambda) = \frac{\sigma^2(p) + (\pi - \lambda)^2}{\sigma^2(p) + (\pi - \lambda)^2 + \sigma^2(\Delta)},$$

(17)

where $\sigma^2(p)$ is the universe-score variance for persons, $\sigma^2(\Delta)$ is the absolute error variance, $\pi$ is the population mean, and $\lambda$ is the cutting score; ($\pi$ and $\lambda$ are in the metric of proportion of items correct). When $\pi = \lambda$, the index $\Phi(\lambda)$ reaches its lower limit referred to also as $\Phi$ in GT.

As Feldt and Brennan (1993) note, the index $\Phi(\lambda)$ characterizes the dependability of decisions based on the testing procedure, whereas the index $\Phi$ characterizes the contribution of the testing procedure to the dependability of such decisions (p. 141). With the “person x item” (p x i) design in GT, the absolute error variance is $\sigma^2(\Delta) = \sigma^2(p_{i,e})/n + \sigma^2(i)/n$.

As the parameters in formula (17) are in the metric of proportion of items correct, their translation in the framework of this article is (a) $\sigma^2(p) = \sigma^2_{\tau}/n^2$, where $\sigma^2_{\tau}$ is the expected true variance for the NR score; (b) $\sigma^2(\Delta) = \sigma^2_{\epsilon}/n^2 + \sigma^2(\pi_i)/n$, where $\sigma^2_{\epsilon}$ is the expected error variance for the NR score and $\sigma^2(\pi_i)$ is the variance of $\pi_i$ values for the test items ($i = 1, ..., n$), (c) $\sigma^2(i) = \sigma^2(\pi_i)$; and (d) $\sigma^2(p_{i,e}) = \sigma^2_{\epsilon}/n$. With this, the dependability index $\Phi(\lambda)$ translates into

$$\Phi(\lambda) = \frac{\sigma^2_{\tau} + n^2(\pi - \lambda)^2}{\sigma^2_{\epsilon} + \sigma^2_{\tau} + n\sigma^2(\pi_i)}.$$  

(18)

Index $\Phi(\lambda)$ achieves its lowest value when $\pi = \lambda$. The resulting dependability index is

$$\Phi = \frac{\sigma^2_{\tau}}{\sigma^2_{\epsilon} + \sigma^2_{\tau} + n\sigma^2(\pi_i)}.$$  

(19)

It should be stressed that although the empirical evaluation of $\rho_{xx}$, $\Phi(\lambda)$, and $\Phi$ in the framework of GT requires sample data (e.g., binary scores), their theoretical evaluation with formulas (16), (18), and (19) does not require such data as long as the IRT calibration of items is available.

Formulas for Binary Items Calibrated with the 3PLM

With the 3PLM (Birnbaum, 1968), the probability for correct response on item $i$ for a person with a trait score $\theta$ (denoted here as $P^*_i(\theta)$) is provided with

$$P^*_i(\theta) = c_i + (1-c_i) / [1 + \exp[-1.7a_i(\theta - b_i)]],$$  

(20)

where $c_i$ is the pseudo-chance level (“guessing”) parameter of the model. True-score measures for items calibrated with the 2PLM are distinguished from their counterparts calibrated with the 3PLM by using asterisks for the latter (e.g., $\pi_i^*$). Clearly, equation (20) can be written as

$$P^*_i(\theta) = c_i + (1-c_i)P_i(\theta),$$  

(21)

where $P_i(\theta)$ is with the 2PLM (see equation (11)).

Expected Item Score

The expected item score for calibrations with the 3PLM is

$$\pi_i^* = c_i + (1-c_i)\pi_i,$$  

(22)
where $\pi_i$ is obtained through formula (7) for calibrations with the 2PLM. The proof follows from multiplying on both sides of equation (21) by $\varphi(\theta)$ and integrating each side from -4 to 4.

**Expected Item Error Variance**

The *expected item error variance* for calibrations with the 3PLM is

$$\sigma^2(e_i^*) = c_i(1 - c_i)(1 - \pi_i) + (1 - c_i)^2 \sigma^2(e_i),$$

where $\pi_i$ and $\sigma^2(e_i)$ are obtained through formulas (7) and (10), respectively, for calibrations with the 2PLM (proof in Appendix A). Figure 2 graphically represents expected values of the item error variance (calculated with formula (23)) as a function of the item parameters $a_i$ and $b_i$ for a fixed value of the pseudo-chance level parameter ($c_i = 0.2$).
Expected Item True Variance.

The *expected item true variance* for calibrations with the 3PLM is

\[ \sigma^2(\tau^*_i) = \pi^*_i(1 - \pi^*_i) - \sigma^2(e^*_i), \]

where \( \pi^*_i \) and \( \sigma^2(e^*_i) \) are obtained with formulas (22) and (23), respectively. Formula (24) follows directly from formula (11) because the derivation of the latter does not depend on the model used for item calibration (1PLM, 2PLM, or 3PLM).

Reliability of Item Score

As with the 2PLM, the reliability of individual binary items calibrated with the 3PLM is

\[ \rho_{ii}^* = \frac{\sigma^2(\tau^*_i)}{\sigma^2(\tau^*_i) + \sigma^2(e^*_i)}, \]

where \( \sigma^2(e^*_i) \) and \( \sigma^2(\tau^*_i) \) are obtained with formulas (23) and (24), respectively.

True-Score Measures and Reliability at the Test Level

Formulas (13) through (16), (18), and (19) for true-score measures and reliability at the test level with the 2PLM translate directly into their 3PLM counterparts for the expected marginal NR score, error variance for the NR score, true score variance, reliability, and dependability (it suffices to use starred notations for the symbols that participate in the right-hand side of each of these six formulas).

Formulas for Binary Items Calibrated with the 1PLM

When the discrimination index in equation (1) is a constant \( a_i = a \), the 2PLM translates into the 1PLM. However, one should know which computational 1PLM is used for the item calibration: logistic (with a scaling constant \( D = 1.0 \)) or logistic approximation of the normal ogive model \( (D = 1.7) \). Both options are available with some computer programs (e.g., RASCAL; Assessment Systems Corporation, 1995a). When the 1PLM item analysis is conducted with standardization on the trait scores (with \( D = 1.7 \)), one can use directly the formulas for true-score measures and reliability derived in this article for the 2PLM (with \( a_i = constant \), as provided with the 1PLM item calibration). For the "pure" Rasch model \( (D = 1, a_i = 1) \) (Rasch, 1960), in which the standardization is on the item difficulty, one can use formulas developed by Dimitrov (2003) for expected marginal true-score measures and reliability of binary items with normal and logistic trait distributions.

Examples

Simulated Data Example

In this example, binary scores for 8,000 persons were simulated to fit the 2PLM with the standard normal distribution for trait scores, \( \theta \sim N(0, 1) \), and fixed values of \( a_i \) and \( b_i \) for 20 items. The empirical validation of formulas (7) and (10) (for \( \pi_i \) and \( \sigma^2(e_i) \) with the 2PLM) is of particular interest because all other formulas represent (explicitly or implicitly) functions of \( \pi_i \) and \( \sigma^2(e_i) \).

The data were generated using a computer program written in SAS (SAS Institute, 1985) for Monte Carlo simulations of binary data that fit IRT models (Dimitrov, 1996). When the assumptions of \( \theta \sim N(0, 1) \) and model fit with the 2PLM were met with these simulations, the produced binary
## Table 1
True-Score Measures and Reliability for Simulated Binary Items Calibrated With the Two-Parameter Logistic Model (2PLM)

<table>
<thead>
<tr>
<th>Item</th>
<th>$a_i$</th>
<th>$b_i$</th>
<th>$\pi_i$</th>
<th>$(p_i)^a$</th>
<th>$\sigma^2(e_i)$</th>
<th>$\sigma^2(\tau_i)$</th>
<th>$\rho_{ii}$</th>
<th>$p_i - \pi_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.449</td>
<td>-2.554</td>
<td>0.852</td>
<td>(0.849)</td>
<td>0.120</td>
<td>0.006</td>
<td>0.050</td>
<td>-0.003</td>
</tr>
<tr>
<td>2</td>
<td>0.402</td>
<td>-2.161</td>
<td>0.790</td>
<td>(0.785)</td>
<td>0.154</td>
<td>0.012</td>
<td>0.074</td>
<td>-0.005</td>
</tr>
<tr>
<td>3</td>
<td>0.232</td>
<td>-1.551</td>
<td>0.637</td>
<td>(0.644)</td>
<td>0.220</td>
<td>0.011</td>
<td>0.047</td>
<td>0.007</td>
</tr>
<tr>
<td>4</td>
<td>0.240</td>
<td>-1.226</td>
<td>0.612</td>
<td>(0.618)</td>
<td>0.226</td>
<td>0.012</td>
<td>0.050</td>
<td>0.006</td>
</tr>
<tr>
<td>5</td>
<td>0.610</td>
<td>-0.127</td>
<td>0.526</td>
<td>(0.526)</td>
<td>0.199</td>
<td>0.050</td>
<td>0.201</td>
<td>-0.001</td>
</tr>
<tr>
<td>6</td>
<td>0.551</td>
<td>-0.855</td>
<td>0.660</td>
<td>(0.653)</td>
<td>0.188</td>
<td>0.036</td>
<td>0.161</td>
<td>-0.007</td>
</tr>
<tr>
<td>7</td>
<td>0.371</td>
<td>-0.568</td>
<td>0.578</td>
<td>(0.577)</td>
<td>0.219</td>
<td>0.025</td>
<td>0.104</td>
<td>-0.001</td>
</tr>
<tr>
<td>8</td>
<td>0.321</td>
<td>-0.277</td>
<td>0.534</td>
<td>(0.534)</td>
<td>0.228</td>
<td>0.021</td>
<td>0.085</td>
<td>0.000</td>
</tr>
<tr>
<td>9</td>
<td>0.403</td>
<td>-0.017</td>
<td>0.502</td>
<td>(0.503)</td>
<td>0.220</td>
<td>0.030</td>
<td>0.120</td>
<td>0.001</td>
</tr>
<tr>
<td>10</td>
<td>0.434</td>
<td>0.294</td>
<td>0.454</td>
<td>(0.456)</td>
<td>0.215</td>
<td>0.033</td>
<td>0.131</td>
<td>0.002</td>
</tr>
<tr>
<td>11</td>
<td>0.459</td>
<td>0.532</td>
<td>0.412</td>
<td>(0.416)</td>
<td>0.209</td>
<td>0.034</td>
<td>0.138</td>
<td>0.004</td>
</tr>
<tr>
<td>12</td>
<td>0.410</td>
<td>0.773</td>
<td>0.385</td>
<td>(0.389)</td>
<td>0.209</td>
<td>0.027</td>
<td>0.116</td>
<td>0.004</td>
</tr>
<tr>
<td>13</td>
<td>0.302</td>
<td>1.004</td>
<td>0.386</td>
<td>(0.384)</td>
<td>0.219</td>
<td>0.018</td>
<td>0.074</td>
<td>-0.002</td>
</tr>
<tr>
<td>14</td>
<td>0.343</td>
<td>1.250</td>
<td>0.342</td>
<td>(0.345)</td>
<td>0.206</td>
<td>0.019</td>
<td>0.086</td>
<td>0.003</td>
</tr>
<tr>
<td>15</td>
<td>0.225</td>
<td>1.562</td>
<td>0.366</td>
<td>(0.360)</td>
<td>0.222</td>
<td>0.010</td>
<td>0.044</td>
<td>-0.006</td>
</tr>
<tr>
<td>16</td>
<td>0.215</td>
<td>1.385</td>
<td>0.385</td>
<td>(0.379)</td>
<td>0.227</td>
<td>0.010</td>
<td>0.040</td>
<td>-0.006</td>
</tr>
<tr>
<td>17</td>
<td>0.487</td>
<td>2.312</td>
<td>0.156</td>
<td>(0.163)</td>
<td>0.123</td>
<td>0.008</td>
<td>0.062</td>
<td>0.007</td>
</tr>
<tr>
<td>18</td>
<td>0.608</td>
<td>2.650</td>
<td>0.084</td>
<td>(0.092)</td>
<td>0.078</td>
<td>0.000</td>
<td>0.000</td>
<td>0.008</td>
</tr>
<tr>
<td>19</td>
<td>0.341</td>
<td>2.712</td>
<td>0.191</td>
<td>(0.192)</td>
<td>0.146</td>
<td>0.009</td>
<td>0.058</td>
<td>0.001</td>
</tr>
<tr>
<td>20</td>
<td>0.465</td>
<td>3.000</td>
<td>0.103</td>
<td>(0.099)</td>
<td>0.091</td>
<td>0.001</td>
<td>0.013</td>
<td>-0.004</td>
</tr>
</tbody>
</table>

*a* Observed item score (proportion correct responses) for the simulated data ($n = 8,000$).

Scores (for 8,000 persons on 20 items) were analyzed using the computer program XCALIBRE (Assessment Systems Corporation, 1995b). Instead of the “ideal” values of $a_i$ and $b_i$ used with the generating 2PLM in this simulation, their XCALIBRE estimates (given in Table 1) were used with the purpose to test the robustness of formulas (7) and (10) for sample-based (i.e., less than “ideal”) estimates of item parameters. The evaluations of true-score measures and reliability in this example were facilitated by using the statistical program SPSS. The SPSS program syntax developed for this purpose (see Appendix B) works for binary items calibrated with the 3PLM (input variables: $a_i$, $b_i$, and $c_i$), but it also works for items calibrated with the 2PLM ($c_i = 0$) or the 1PLM ($c_i = 0$ and $a_i = \text{constant}$). The SPSS run provides expected true-score measures and reliability for each item ($\pi_i$, $\sigma^2(e_i)$, $\sigma^2(\tau_i)$, and $\rho_{ii}$) as “new” variables in the SPSS spreadsheet. At the test level, the SPSS printout provides the expected true score variance for the NR score ($\sigma^2(\pi)$), the variance of expected scores for the test items ($\sigma^2(\pi)$).

In this example, the SPSS syntax in Appendix B was run with the values of $a_i$ and $b_i$ from Table 1 and $c_i = 0$. The results for individual items are provided also in Table 1. At the test level, the SPSS printout provided the expected true score variance for the NR score ($\sigma^2 = 6.315$), the expected error variance for the NR score ($\sigma^2 = 3.719$), the expected NR score ($\mu = 8.956$), and the variance of $\pi_i$ values for the 20 items ($\sigma^2(\pi) = .045$). With this, the domain score is $\pi = \mu / n = 8.956 / 20 = .448$, and the reliability of the NR scores is $\rho_{xx} = .63$ (using formula (16)).
The empirical estimates of true-score measures and reliability for the simulated data were also
determined and compared to their theoretical counterparts. Most importantly, a strong match was
found between the theoretical evaluations of $\pi_i$ and $\sigma^2(e_i)$ and their empirical counterparts
denoted here as $p_i$ and $s_i^2$, respectively. The empirical item scores, $p_i$ (provided by XCALIBRE for the
simulated data) are given in Table 1. The difference between $p_i$ and $\pi_i$ (also in Table 1) is smaller
than 0.01 in absolute value. The same is true for the difference between the theoretical and empirical
item variances: $\sigma^2(e_i) - s_i^2$. One can check this quickly and easily using, for example, the SPSS
spreadsheet for Table 1 and calculating $s_i^2 = p_i (1 - p_i)$.

As noted earlier, the empirical validation of the accuracy of formulas (7) and (10) is important
because the values of $\pi_i$ and $\sigma^2(e_i)$ produced by these two formulas govern the values of other
true-score measures and reliability indexes. Given the strong match between theoretical and
empirical estimates for the item score and item error variance in this example, it is not a surprise
that Cronbach's alpha for the sample of simulated binary scores ($n = 8,000$) was equal (to the
nearest hundredth) to the theoretically evaluated reliability ($\alpha = \rho_{x_\pi} = .63$). Similarly, the empirical
mean and variance of the item scores in Table 1 ($\bar{p} = .448$ and $\sigma^2(p_i) = 0.044$) match their
theoretical counterparts ($\pi = .448$ and $\sigma^2(x_\pi) = 0.045$). Thus, with the assumptions of data fit
and normal trait distribution met, there is a strong match between the theoretical and empirical
values of true-score measures even when the proposed formulas are applied with IRT estimates (not
“ideal” values) of the item parameters for relatively large samples (in this case, $n = 8,000$).

Real Data Example

The data for this example consist of binary scores for 4,854 fifth graders on 24 multiple-choice
items of the Ohio Off-Grade Proficiency Test - Reading (Riverside Publishing, 1997) in a
large urban area in northeastern Ohio. The items capture four strands of learning outcomes defined
by the publisher as (a) examining meaning given a fiction or poetry text, (b) extending meaning
given a fiction or poetry text, (c) examining meaning given a nonfiction text, and (d) extending
meaning given a nonfiction text. The data were analyzed using XCALIBRE with the 3PLM (to
accommodate for “guessing” with the multiple-choice items). For the test of data fit XCALIBRE
reports a standardized residual statistic for each item. This statistic follows (approximately) the
standard normal distribution, and values in excess of 2.0 indicate misfit with a Type I error rate of
0.05. The standardized residuals for the 24 binary items ranged from 0.34 to 1.13 thus indicating
that the data fit the 3PLM. The XCALIBRE estimates of item discrimination ($a_i$), item difficulty
($b_i$), and pseudo-chance level ($c_i$) are given in Table 2 (the items are grouped by strands of learning
outcomes).

The normal quantile tests were conducted using SPSS with the trait scores, $\theta$, provided by
XCALIBRE for the sample data ($n = 4,854$). The results indicated a good fit of $\theta$ to $\mathcal{N}(0,1)$ thus
allowing the application of formulas developed in this article. The theoretical true-score measures
and reliability were evaluated through the use of the SPSS syntax in Appendix B (with the item
parameters $a_i, b_i$, and $c_i$ in Table 2 as “input” SPSS variables). The results are summarized in Table 2
by strands of learning outcomes. In terms of domain score, the highest performance of the target
population of fifth graders is on the learning outcome “poetry — constructing meaning” ($\pi = .664$),
whereas their lowest performance is on the learning outcome “nonfiction — extending meaning”
($\pi = .475$). The dependability index $\Phi(\lambda)$ was also calculated using formula (18) for values of
the cutting score $\lambda$ (proportion of items correct) from 0 to 1 with a “step” of 0.01. As one can see
from Figure 3, for example, the dependability of pass/fail decisions based on a domain cutting score
$\lambda = .8$ (i.e., 80% items correct) is $\Phi(\lambda) = .90$.

With the data in this example (as with any sample of real data), it is not realistic to expect
ideal conditions for the assumptions of model fit and normality of the trait distribution. Yet, there
Table 2
True-Score Measures and Reliability by Strands of Learning Outcomes With the Ohio Off-Grade Proficiency Test (OOGP)–Reading

<table>
<thead>
<tr>
<th>Strand</th>
<th>Strand</th>
<th>Strand</th>
<th>Observed Item Score (Proportion Correct Responses) for the Real Data (n = 4,854)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poetry - constructing meaning (n = 10)</td>
<td>0.664</td>
<td>1.772</td>
<td>3.293</td>
</tr>
<tr>
<td>Item</td>
<td>ai</td>
<td>bi</td>
<td>ci</td>
</tr>
<tr>
<td>1</td>
<td>1.089</td>
<td>-0.732</td>
<td>0.209</td>
</tr>
<tr>
<td>2</td>
<td>0.948</td>
<td>-0.418</td>
<td>0.220</td>
</tr>
<tr>
<td>5</td>
<td>0.494</td>
<td>0.900</td>
<td>0.226</td>
</tr>
<tr>
<td>6</td>
<td>0.494</td>
<td>0.885</td>
<td>0.234</td>
</tr>
<tr>
<td>7</td>
<td>0.905</td>
<td>-0.672</td>
<td>0.185</td>
</tr>
<tr>
<td>8</td>
<td>1.165</td>
<td>-1.144</td>
<td>0.205</td>
</tr>
<tr>
<td>20</td>
<td>0.594</td>
<td>-0.412</td>
<td>0.209</td>
</tr>
<tr>
<td>21</td>
<td>0.716</td>
<td>0.475</td>
<td>0.237</td>
</tr>
<tr>
<td>22</td>
<td>0.703</td>
<td>-0.492</td>
<td>0.204</td>
</tr>
<tr>
<td>23</td>
<td>0.841</td>
<td>-0.504</td>
<td>0.194</td>
</tr>
<tr>
<td>Poetry - extending meaning (n = 4)</td>
<td>0.596</td>
<td>0.722</td>
<td>0.517</td>
</tr>
<tr>
<td>3</td>
<td>1.169</td>
<td>0.468</td>
<td>0.159</td>
</tr>
<tr>
<td>4</td>
<td>0.724</td>
<td>-1.541</td>
<td>0.211</td>
</tr>
<tr>
<td>9</td>
<td>0.554</td>
<td>-0.042</td>
<td>0.197</td>
</tr>
<tr>
<td>24</td>
<td>0.706</td>
<td>0.698</td>
<td>0.177</td>
</tr>
<tr>
<td>Nonfiction - constructing meaning (n = 5)</td>
<td>0.529</td>
<td>1.025</td>
<td>0.655</td>
</tr>
<tr>
<td>10</td>
<td>0.795</td>
<td>-0.226</td>
<td>0.194</td>
</tr>
<tr>
<td>11</td>
<td>0.506</td>
<td>1.581</td>
<td>0.218</td>
</tr>
<tr>
<td>16</td>
<td>0.809</td>
<td>-0.154</td>
<td>0.192</td>
</tr>
<tr>
<td>17</td>
<td>0.499</td>
<td>2.076</td>
<td>0.220</td>
</tr>
<tr>
<td>18</td>
<td>0.839</td>
<td>0.075</td>
<td>0.261</td>
</tr>
<tr>
<td>Nonfiction - extending meaning (n = 5)</td>
<td>0.475</td>
<td>0.952</td>
<td>0.520</td>
</tr>
<tr>
<td>12</td>
<td>0.709</td>
<td>2.238</td>
<td>0.190</td>
</tr>
<tr>
<td>13</td>
<td>0.869</td>
<td>-0.727</td>
<td>0.221</td>
</tr>
<tr>
<td>14</td>
<td>0.686</td>
<td>0.375</td>
<td>0.215</td>
</tr>
<tr>
<td>15</td>
<td>0.795</td>
<td>0.219</td>
<td>0.180</td>
</tr>
<tr>
<td>19</td>
<td>0.812</td>
<td>1.874</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Total (n = 24) | 0.585 | 4.471 | 16.520 | 0.789 |

^ Observed item score (proportion correct responses) for the real data (n = 4,854).

is still a good match between theoretical and empirical values for item scores (\(\pi_i\) vs. \(p_i\) values in Table 2), variance of items scores (\(\sigma^2(\pi) = \frac{.027}{.025}\) domain score (\(\pi = .585\) vs. \(\overline{p} = .586\)), and reliability (\(\rho_{xx} = .789\) vs. Cronbach's \(\alpha = .801\)). Additional comments on \(\rho_{xx}\) and its empirical evaluation through Cronbach's \(\alpha\) are provided in the Discussion section.

In this example the 3PLM estimates of item parameters were determined from sample data, but the procedures remain the same when \(a_i\), \(b_i\), and \(c_i\) are known from previous (or simulated)
calibrations with the 3PLM. Thus, once the items are calibrated, one can determine (without further data collection) the true-score characteristics and reliability for any (sub)set of items. In the context of this example, for instance, one can use the calibration of items with the OOPT-Reading test to develop test booklets for tailored follow-up reading diagnostics (e.g., in different school districts).

Discussion

This article provides analytic evaluations (formulas) for expected true-score measures and reliability of binary items as a function of their IRT parameters. Assuming the normal distribution of trait scores, the formulas can be applied for items calibrated with the 1PLM, 2PLM, or 3PLM without information about binary scores or trait scores of persons from the target population. At item level, the proposed formulas provide evaluation for expected (marginal) values of item score ($\pi_i$), item error variance ($\sigma^2(e_i)$), item true variance ($\sigma^2(\tau_i)$), and item reliability ($\rho_{ii}$). At the test level, the item true-score measures are “summarized” in formulas for the expected values of the NR score ($\mu$), domain score ($\pi$), error variance for the NR score ($\sigma^2(e)$), true variance for the NR score ($\sigma^2(\tau)$), reliability ($\rho_{xx}$), and dependability ($\Phi(\lambda)$) for criterion-referenced interpretations based on a domain cutting score, $\lambda$.

Brief clarifications about the derivation design for the formulas proposed in this article are necessary. For item calibrations with the 2PLM, the formulas for expected item score, $\pi_i$ (formula (7)), and expected item error variance, $\sigma^2(e_i)$ (formula (10)), are based on approximations...
with an absolute error practically close to zero (less than 0.0005 with formula (7) and less than 0.005 with formula (10)). The (negligible) approximation errors with formulas (7) and (10) lurk in the other formulas derived in this article although the expected true-score measures in these formulas represent (explicitly or implicitly) exact functions of \( \pi_i \) and \( \sigma^2(e_i) \) (e.g., see formulas (11) and (12)).

Some arguments in support of using the formulas proposed in this article versus brute-force numerical integrations also seem appropriate. First, the proposed formulas are easy to perform with widely used spreadsheets, statistical programs (e.g., SPSS, see Appendix B), or even regular calculators. Numerical integrations, instead, require computer programming with more complicated analytic expressions (e.g., Gaussian quadratures) thus limiting the range of potential users with studies that involve evaluations at true-score level. Moreover, some methods of numerical integrations involve procedures that may negatively affect accuracy. For example, the Simpson’s rule for numerical integrations with equation (6) involves an approximation of the compound binomial distribution of the number-correct scores (e.g., May & Nicewander, 1993) which, in turn, leads to losing accuracy in estimating the true score variance. In contrast, formula (11) (for expected true-score variance of individual items) does not use preliminary approximations. Along with technical advantages, the formulas provide theoretical relationships that may remain hidden with numerical integrations. Formula (10), for example, shows that the expected item error variance is an even function of the item difficulty, \( b_i \), for fixed values of the discrimination index, \( a_i \). Also, although formulas (11) and (15) reveal relationships between true-score measures for item calibrations with the same (e.g., 2PLM or 3PLM) IRT model, formulas (22) and (23) connect item true-score measures with the 3PLM to item true-score measures with the 2PLM. The proposed formulas can help researchers to plan (model, predict) true-score measures, whereas the numerical integrations put researchers in a post-hoc position. Also, the formulas provide more than just calculations — they capture theoretical relationships between concepts of IRT and true-score theory that may have useful applications in research and instructional settings (e.g., graduate courses in measurement).

The comparison of theoretical true-score measures and reliability with their empirical counterparts for real data also deserves attention. The empirical approach (a) requires information about binary scores for persons from the target population and (b) provides sample-based estimates that may (to a large extent) misrepresent the population parameters for true-score measures and reliability. Conversely, the proposed formulas provide accurate evaluation of true-score measures and reliability at the population level without using sample-based number-correct scores or trait scores (IRT estimates of the item parameters suffice). It should be also noted that Cronbach's alpha is an accurate empirical estimate for reliability (\( \rho_{xx} \)) only if there is no correlation among errors and the test components are essentially tau-equivalent (Novick & Lewis, 1967). The theoretical evaluation of \( \rho_{xx} \) in this article, however, does not require tau-equivalency (the weaker assumption of congeneric items suffices). As a reminder, test items are (a) congeneric if they measure the same trait and (b) tau-equivalent if they measure the same trait and their true scores have equal variances (e.g., Jöreskog, 1971). When the tau-equivalency assumption does not hold, Cronbach’s alpha underestimates \( \rho_{xx} \). However, Cronbach’s alpha may also overestimate \( \rho_{xx} \) when there is a correlation among errors, (e.g., Komaroff, 1997; Raykov, 2001). Correlated errors may occur, for example, with items that relate to a common stimulus (e.g., same paragraph or graph). For example, the fact that (with the real data example in the previous section) Cronbach’s alpha (.801) slightly overestimated the theoretical evaluation of \( \rho_{xx} \) (.789) should not be a surprise as some items in the reading test (OOPT-Reading) relate to the same paragraph (i.e., correlated errors may occur.) From another perspective, although the marginal reliability for IRT trait scores in computerized adaptive testing is evaluated for the population (Green et al., 1984), it is compared to Cronbach’s alpha for alternatively used paper-and-pencil forms. Clearly, it is more appropriate to compare the theoretical
marginal reliability in an IRT system to theoretical evaluations of the classical reliability, \( \rho_{xx} \) (e.g., with formulas provided in this article).

As illustrated with the examples in the previous section, given the IRT calibration of binary items, one can evaluate their true-score measures and reliability for norm-referenced and criterion-referenced interpretations. One can also do this for any combination of items grouped by measurement or substantive characteristics (e.g., by content or learning outcomes) without using (trait or raw-score) data. This can be particularly useful in developing test booklets for follow-up measurements in longitudinal studies using the IRT calibration of items for a base year study. It should be noted that in previous studies (e.g., National Center for Educational Statistics, 1996) test booklets that are developed for follow-up measurements are usually compared on average item difficulty, thus ignoring the effect of the other item parameter(s). With the formulas proposed in this article, true-score measures and reliability are evaluated as functions of all item parameters (with an appropriate IRT model) prior to follow-up data collection. The formulas can also be incorporated into computer programs for simulation studies, thus allowing researchers to generate targeted true-score measures from (hypothetical or real) IRT parameters of binary items.

It is important to emphasize that the formulas proposed in this article deal with marginal true-score measures and reliability and, therefore, do not provide conditional information about scores and their accuracy at separate trait levels. However, although "diagnostic" IRT information about trait measures for separate persons is valuable, marginal true-score information about the population and the measurement quality of test items is also useful. In a sports analogy, although the assessment of individual players is very important, the evaluation of the team as a whole is also important. In conclusion, researchers and practitioners can greatly benefit from combining IRT conditional information about trait and true-score measures (e.g., using a test characteristic curve) with marginal true-score information provided by the proposed formulas.

Appendix A

Proof of Formula (5)

Formula (5) provides an approximation (with an absolute error smaller than 0.0005) for the expected marginal scores of binary items

\[
\pi_i = \frac{1 - \text{erf}(X_i)}{2},
\]  

(A1)

where \( X_i = a_i b_i / \sqrt{2(1 + a_i^2)} \) and \( \text{erf}(X_i) \) is the error function (e.g., Hastings, 1955, p. 185)

\[
\text{erf}(X) = (2 / \sqrt{\pi}) \int_0^X \exp(-u^2)du.
\]  

(A2)

Lord's approximation (Lord & Novick, 1968, p. 377, Equation 16.9.3) for expected item score (marginal probability for correct response on the item) is

\[
\pi_i = \frac{1}{\sqrt{2\pi}} \int_{\gamma_i}^\infty \exp(-t^2 / 2)dt,
\]  

(A3)

where \( \gamma_i = a_i b_i / \sqrt{1+a_i^2} \). With the substitution \( t = u\sqrt{2} \) (and thus \( \gamma_i = X_i\sqrt{2} \)),
\[
\pi_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-u^2)du = \frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_{0}^{X_i} \exp(-u^2)du = \frac{1}{2} - \frac{1}{2} \text{erf}(X_i),
\]

with which the proof is completed.

The benefit from representing the Lord’s integral for \( \pi_i \) through the error function, \( \text{erf}(X_i) \), is that the approximation of \( \text{erf}(X_i) \) with equation (6) is simple and produces a practically negligible error (less than 0.0005 in absolute value). The approximation error is even much smaller when \( \text{erf}(X_i) \) is executed in computer programs for mathematics (e.g., MATLAB; MathWorks, Inc., 1999).

**Proof of Formula (11)**

Formula (11) represents the expected item true variance, \( \sigma^2(\tau_i) \), as an exact function of the expected item score, \( \pi_i \), and expected item error variance, \( \sigma^2(e_i) \). Using the variance expectation rule \( \text{VAR}(X) = E(X^2) - [E(X)]^2 \) with \( X = P_4(\theta) \),

\[\sigma^2(\tau_i) = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} P_i(\theta) \varphi(\theta)d\theta - \left( \int_{-\infty}^{\infty} P_i(\theta) \varphi(\theta)d\theta \right)^2 \right] \varphi(\theta)d\theta = \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)]\varphi(\theta)d\theta - \pi_i^2 = \pi_i^2 - \sigma^2_e - \pi_i^2 = \pi_i(1 - \pi_i) - \sigma^2(e_i),\]

with which the proof is completed.

**Proof of Formula (15)**

Formula (15) represents the expected true score variance for the NR score, \( \sigma^2_{\tau} \), as an exact function of the expected item score, \( \pi_i \), and item error variance, \( \sigma^2(e_i) \). For unidimensional tests (which are dealt with in this article), there is a perfect correlation between the congeneric true scores \( (\tau_i, \tau_j) \) of any two items, \( i \) and \( j \), because of the linear relationship: \( \tau_i = a_{ij} + b_{ij} \tau_j \), where \( b_{ij} \neq 0 \), 1 (e.g., Jöreskog, 1971). Thus, the covariance of \( \tau_i \) and \( \tau_j \) is \( \sigma(\tau_i, \tau_j) = \sigma(\tau_i)\sigma(\tau_j) \). With this, the variance of the true number-right score on a test of \( n \) binary items, \( \tau = \sum \tau_i (i = 1, ..., n) \), can be represented as

\[\sigma^2_{\tau} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma(\tau_i, \tau_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma(\tau_i)\sigma(\tau_j). \] (A4)

By replacing \( \sigma(\tau_i) \) and \( \sigma(\tau_j) \) in the far-right side of equation (A4) with their equivalent expressions in formula (11), we obtain formula (15). With this the proof is completed.

**Proof of Formula (23)**

Formula (23) represents the expected error variance for individual binary items calibrated with the 3PLM, \( \sigma^2(e_i^*) \), as an exact function of the 2PLM evaluations for item score, \( \pi_i \), and item error
variance, $\sigma^2(e_i)$. Given the relationship between $P_i^*(\theta)$ with the 3PLM and $P_i(\theta)$ with the 2PLM (see equation (21)), it can be easily seen that

$$P_i^*(\theta)[1 - P_i^*(\theta)] = c_i (1 - c_i) [1 - P_i(\theta)] + (1 - c_i)^2 P_i(\theta)[1 - P_i(\theta)]. \quad (A5)$$

Using equation (A5), the proof of formula (23) is provided with the following integral manipulations:

$$\begin{align*}
\sigma^2(e_i^*) &= \int_{-\infty}^{\infty} P_i^*(\theta)[1 - P_i^*(\theta)] \varphi(\theta) d\theta \\
&= c_i (1 - c_i) \int_{-\infty}^{\infty} \varphi(\theta) d\theta - c_i (1 - c_i) \int_{-\infty}^{\infty} P_i(\theta) \varphi(\theta) d\theta \\
&+ (1 - c_i)^2 \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)] \varphi(\theta) d\theta \\
&= c_i (1 - c_i) - c_i (1 - c_i) \pi_i + (1 - c_i)^2 \sigma^2(e_i) \\
&= c_i (1 - c_i)(1 - \pi_i) + (1 - c_i)^2 \sigma^2(e_i).
\end{align*}$$

Appendix B

**SPSS Syntax: Evaluation of Marginal True-Score Measures for Binary Items**

**Input Variables: IRT Item Parameters** $(a_i, b_i, \text{ and } c_i)$

```spss
COMPUTE p = .2646 - .118*a + .0187*(a**2). 
COMPUTE s = .7427 + .7081/a + .0074/(a**2). 
COMPUTE ve = p*exp(-.5*((b/s)**2)). 
COMPUTE X = (a*b)/sqrt(2*(1+a**2)). 
COMPUTE erf = (1+.278393*abs(X) + .230389*X**2 + .000972*(abs(X))**3 + .078108*X**4)**4. 
COMPUTE erf = 1 - 1/erf. 
IF(X < 0) erf = -erf. 
COMPUTE pi = (1-erf)/2. 
COMPUTE vt = pi*(1 - pi) - ve. 
IF(vt < 0) vt = 0. 
COMPUTE ve = c*(1-c)*(1-pi) + ve*((1-c)**2). 
COMPUTE pi = c + (1-c)*pi. 
COMPUTE vt = pi*(1 - pi) - ve. 
IF(vt < 0) vt = 0. 
SET FORMAT = F8.3 ERRORS = NONE RESULTS OFF HEATHER NO. 
FLIP VARIABLES a b c pi ve vt. 
VECTOR V = VAR001 TO VAR020. 
COMPUTE Y = 0. 
LOOP #I = 1 TO 20.
```
LOOP #J = 1 TO 20.
   COMPUTE Y = Y + SQRT(V(#I)*V(#J)).
END LOOP.
END LOOP.
FLIP VAR001 TO VAR020 Y.
COMPUTE roi = vt/(vt + ve).
SET RESULTS ON.
REPORT FORMAT = AUTOMATIC
   /VARIABLES = pi '  ' ve '   ' vt '  '
   /BREAK = (TOTAL)
   /SUMMARY = MAX(vt) 'True score variance:'
   /SUMMARY = SUBTRACT(SUM(ve) MAX(ve)) (vt (COMMA) (3)) 'Error variance:'
   /SUMMARY = SUBTRACT(SUM(pi) MAX(pi)) (vt (COMMA) (3)) 'Marginal NR score:'
SELECT IF(CASE_LBL ~= 'Y') .
RENAME VARIABLES (CASE_LBL = ITEM) (ve=var_err) (vt=var_tau).
VARIABLE LABELS pi 'item score'.
DESCRIPTIVES
   VARIABLES = pi
   /STATISTICS = VAR.

Note. The user should specify the number of items (in this example, 20) in the syntax. With 50 items, for example, change 20 to 50 and VAR020 to VAR050 (see the bold notations in the four syntax lines.)

References


**Acknowledgements**

I would like to thank the editor and the anonymous reviewers for their helpful comments on earlier versions of this manuscript.

**Author’s Address**

Address correspondence to Dimiter M. Dimitrov, Graduate School of Education, George Mason University, 4400 University Drive, MS 4B3, Fairfax, Virginia 22030-4444; e-mail: ddimitro@gmu.edu