DOCUMENT RESUME

ED 475 704 TM 034 866

AUTHOR	Dimitrov, Dimiter M.
TITLE	Expected Values and Reliability of Number-Right Scores for IRT Calibrated Items.
PUB DATE	2003-04-00
NOTE	32p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Item Response Theory; *Reliability; Test Construction; *Test Items; *True Scores; Values
IDENTIFIERS	Binary Scores; *Calibration; *Number Right Scoring

ABSTRACT

This paper provides analytic evaluations of expected (marginal) true-score measures for binary items given their item response theory (IRT) calibration. Under the assumption of normal trait distributions, marginalized true scores, error variance, true score variance, and reliability for norm-referenced and criterion-references 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 for the test 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. Two appendixes contain formula proofs and the Statistical Package for the Social Sciences syntax for evaluation of the marginal true-score measures. (Contains 2 tables and 22 references.) (Author/SLD)

Expected Values and Reliability of Number-Right Scores for IRT Calibrated Items

Dimiter M. Dimitrov
Kent State University
(e-mail: ddimitro@kent.edu)

Paper Presented at the Annual Meeting of the
American Educational Research Association
Chicago IL, April 21-25, 2003

Abstract

This article provides analytic evaluations of expected (marginal) true-score measures for binary items given their IRT calibration. Under the assumption of normal trait distribution, 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 for the test 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.*

Expected Values and Reliability of Number-Right Scores for IRT Calibrated Items

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 and, for example, when pass-fail decisions are made, a cutting score is typically set on the domain-score scale (e.g., Hambleton, Swaminathan, and 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 impact 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 marginal true-score measures and IRT parameters of binary items. Analytic expressions (formulas) 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, θ, that underlies the persons' responses on each item. The IRT conditional error variance at θ, $\sigma^2_{\hat{\theta}|\theta}$, inversely related to the information provided by the test at θ (Birnbaum, 1968), is not to be confused with the conditional raw-score variance at θ, $\sigma^2_{x|\theta}$. The expected value of the latter (when θ varies from -∞ to ∞) is the *error variance for the raw score* (e.g., Lord, 1980), whereas the expected value of the former is referred to as *marginal measurement error variance* (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.; 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 test development and analysis. For example, given a bank of IRT calibrated items, one can select items to develop a test (e.g., for follow-up measurements in longitudinal studies) with 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 θ under an appropriate IRT model: one-parameter (1PL), two-parameter (2PL), or three-parameter (3PL) logistic model (Birnbaum, 1968). As $P_i(\theta)$ is the item true score at θ, the expected marginal number-right (NR) score for a test of *n* binary items is

$$\mu = \sum_{i=1}^{n} \int_{-\infty}^{\infty} P_i(\theta)\varphi(\theta)d\theta, \tag{1}$$

where φ(θ) is the*probability density function* (*pdf*) for the trait distribution. The integration is from -∞ to ∞ since the ability, θ, is not limited in the theoretical framework of IRT. Also, as the product $P_i(\theta)[1 - P_i(\theta)]$ is the conditional error variance for item *i* at θ (Lord, 1980, p. 45), the expected error variance for the NR score on a test of *n* binary items is

$$\sigma_e^2 = \sum_{i=1}^{n} \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)]\varphi(\theta)d\theta. \tag{2}$$

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

$$\sigma_\tau^2 = \int_{-\infty}^{\infty} [n\overline{P}(\theta)]^2 \varphi(\theta)d\theta - \left[\int_{-\infty}^{\infty} n\overline{P}(\theta)\varphi(\theta)d\theta\right]^2, \tag{3}$$

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

Previous research provides limited applications of Equations 1, 2, or 3 using, for example, Gaussian quadrature (Bock & Lieberman, 1970), but analytic solutions are not provided. For example, comparing reliability for NR scores and percentile ranks, May and Nicewander (1993) evaluated the integrals in Equations (2) and (3) using the Simpson's Rule with 100 points on the θ interval from -5 to 5 after approximating the compound binomial distributions of raw scores. This article takes a different approach and provides analytic solutions (formulas) for marginalized true score measures at item level thus making it possible to determine (and control) the contribution of individual items to the values of μ, σ_e^2, σ_τ^2, and reliability indexes at test level. Comments on the advantages of the proposed analytic solutions 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 level. For individual items, formulas are provided for the *item score* (π_i), *item error variance* [$\sigma^2(e_i)$], *item true variance* [$\sigma^2(\tau_i)$], and *item reliability* (ρ_{ii}). At test level, formulas are provided for the population mean of *NR scores* (μ), *domain score* (π), *error variance* (σ_e^2), *true score variance* (σ_τ^2), *reliability* (ρ_{xx}), and *dependability index* [$\Phi(\lambda)$] for criterion-referenced interpretations based on a cutting domain score, λ. For items calibrated with the 2PLM, π_i and $\sigma^2(e_i)$ are evaluated through approximation formulas (with a negligible approximation error). All other true-score measures at item and test level are represented (explicitly or implicitly) as exact analytic functions of π_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, Inc., 2002) provided in Appendix B.

Formulas for Binary Items Calibrated with the 2PLM

With the 2PLM, the probability of a correct answer on a binary item *i* for a person with a trait level θ is determined with

$$P_i(\theta) = \frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]}, \qquad (4)$$

where: a_i is the *item discrimination*, b_i is the *item difficulty*, and *D* is a *scaling factor*; (with $D =$ 1.7, the normal-ogive and logistic item-characteristic functions are almost identical).

Item Score.

The marginal probability of correct responses on item *i* is referred to here as *item score*, π_i In classical test theory, the empirical estimate of π_i is referred to as *item difficulty* (although it is, in fact, the *easiness* of the item.) As proven in Appendix A, π_i can be represented as a function of the IRT item parameters (a_i and b_i) as

$$\pi_i = \frac{1 - erf(X_i)}{2}, \qquad (5)$$

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 an 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:

$$\mathrm{erf(X)} = 1 - \left(1 + \mathrm{a_1X} + \mathrm{a_2X^2} + \mathrm{a_3X^3} + \mathrm{a_4X^4}\right)^{-4}, \qquad (6)$$

where $a_1 = .278393$; $a_2 = .230389$; $a_3 = .000972$; $a_4 = .078108$. When $X < 0$, one can use that $erf(-X) = -erf(X)$. 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). Figure 1 represents the values of π_i (calculated with Formula 5) as a function of the item parameters a_i and b_i.

Item error variance.

As one can see from Equation 2, the marginal error variance for an item *i* can be obtained through the evaluation of the integral

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)]\varphi(\theta)d\theta \qquad (7)$$

With φ(θ) for the standard normal distribution and $D = 1.7$ with the 2PLM, Equation 7 becomes

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} \frac{\exp[1.7a_i(\theta - b_i)]}{\left(1 + \exp[1.7a_i(\theta - b_i)]\right)^2}\left(\frac{1}{\sqrt{2\pi}}\exp(-.5\theta^2)\right)d\theta \qquad (8)$$

Since a closed form evaluation of the integral in Equation 8 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 Inc., 1998). The resulting approximation formula is

$$\sigma^2(e_i) = m_i\exp[-0.5(b_i / d_i)^2], \qquad (9)$$

where b_i is the item difficulty, whereas m_i and d_i depend on the item discrimination (a_i):

$m_i = 0.2646 - 0.118a_i + 0.0187a_i^2$;

$d_i = 0.7427 + 0.7081/a_i + 0.0074/a_i^2$.

Depending on the values a_i and b_i , the error of approximation with Formula 9 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 9 (graphical illustration in Figure 2), 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 2 also shows, larger errors occur with average difficulty items and smaller errors occur with easy or difficult items. It should be noted also that $\sigma^2(e)$ represents an additive *error variance component* of the (total) error variance for the *NR* score, σ_e^2.

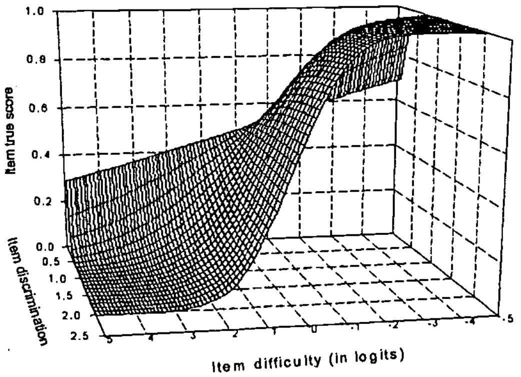


Figure 1. Marginal item score for binary items as a function of their discrimination (a_i) and difficulty (b_i) parameters with the 2PLM

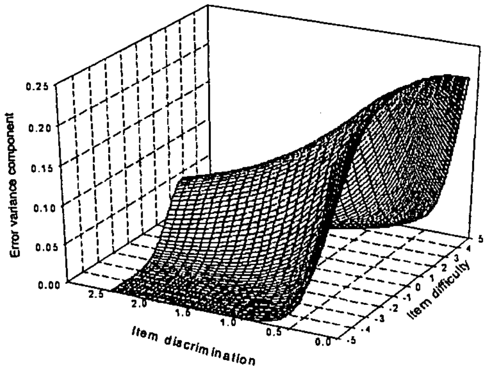


Figure 2. Error variance for binary items as a function of their discrimination (a_i) and difficulty (b_i) parameters with the 2PLM

Item True Variance.

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

$$\sigma^2(\tau_i) = \pi_i(1 - \pi_i) - \sigma^2(e_i). \qquad (10)$$

It should be noted also that the derivation of Formula 10 is the same with *any* IRT model (1PLM, 2PLM, or 3PLM) and *any* (not necessarily normal) trait distribution (see Appendix A).

Item Reliability

In classical test theory, the reliability of item i is empirically estimated with the product $s_i r_{iX}$, where s_i is the item-score standard deviation and r_{iX} 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 reliability* (ρ_{ii}). Thus, given the IRT calibration of binary items, the marginal reliability of an item can be evaluated with

$$\rho_{ii} = \frac{\sigma^2(\tau_i)}{\sigma^2(\tau_i) + \sigma^2(e_i)}, \tag{11}$$

where $\sigma^2(e_i)$ and $\sigma^2(\tau_i)$ are obtained with Formulas 9 and 10, respectively. Information about the reliability of individual items 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).

Marginal NR Score.

Given the item score, π_i, of each item in a test of n binary items, the marginal NR score is

$$\mu = \sum_{i=1}^{n} \pi_i . \tag{12}$$

Error Variance for the NR Score.

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

$$\sigma_e^2 = \sum_{i=1}^{n} \sigma^2(e_i). \tag{13}$$

True Score Variance for the NR Score.

As proven in Appendix A, the marginal true score variance for a test of n binary items is

$$\sigma_\tau^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{[\pi_i(1-\pi_i)-\sigma^2(e_i)][\pi_j(1-\pi_j)-\sigma^2(e_j)]}, \qquad (14)$$

where π_i and $\sigma^2(e_i)$ [or π_j and $\sigma^2(e_j)$] are obtained with Formulas 5 and 7, respectively.

Reliability.

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

$$\rho_{xx} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2}. \qquad (15)$$

In this article, the theoretical value of ρ_{xx} is evaluated by replacing σ_e^2 and σ_τ^2 in Formula 15 with their evaluations obtained through Formulas 13 and 14, 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)

$$\Phi(\lambda) = \frac{\sigma^2(\mathrm{p}) + (\pi-\lambda)^2}{\sigma^2(\mathrm{p}) + (\pi-\lambda)^2 + \sigma^2(\Delta)}, \qquad (16)$$

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

As the parameters in Formula 16 are in the metric of proportion of items correct, their translation in the framework of this article is (a) $\sigma^2(p) = \sigma_\tau^2/n^2$, where σ_τ^2 is the true variance for the NR score, (b) $\sigma^2(\Delta) = \sigma_e^2/n^2 + \sigma^2(\pi_i)/n$, where σ_e^2 is the error variance for the NR score and $\sigma^2(\pi_i)$ is the variance of π_i values ($i = 1, ..., n$), (c) $\sigma^2(i) = \sigma^2(\pi_i)$, and (d) $\sigma^2(pi,e) = \sigma_e^2/n$. With this, the dependability index $\Phi(\lambda)$ translates into

$$\Phi(\lambda) = \frac{\sigma_\tau^2 + n^2(\pi - \lambda)^2}{\sigma_\tau^2 + n^2(\pi - \lambda)^2 + \sigma_e^2 + n\sigma^2(\pi_i)}. \qquad (17)$$

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

$$\Phi = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2 + n\sigma^2(\pi_i)}. \qquad (18)$$

The comparison of Formulas 15 and 18 shows that the dependability index Φ does not exceed the reliability coefficient ρ_{xx}. Intuitively this also makes sense because, as Feldt and Brennan (1993) note, "criterion-referenced interpretations of 'absolute' scores are more stringent than norm-referenced interpretations of 'relative' scores ... Φ can also be interpreted as a chance-corrected index of dependability for criterion-referenced interpretations with squared-error loss" (p. 141). It should be stressed that, while the evaluation of ρ_{xx}, $\Phi(\lambda)$, and Φ in the framework of GT requires sample data (e.g., binary scores), Formulas 15, 17, and 18 in the framework of this article do not require such data as long as the IRT item parameters are 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 θ [denoted here as $P_i^*(\theta)$] is provided with

$$P_i^*(\theta) = c_i + (1 - c_i) / \left\{1 + \exp[-1.7a_i(\theta - b_i)]\right\}, \qquad (19)$$

where c_i is the pseudo-chance level ("guessing") parameter of the model. In order to distinguish true-score measures for items calibrated with the 2PLM from their counterparts with the 3PLM, we star the latter (e.g., π_i^*). Clearly, Equation 19 can be written as

$$P_i^*(\theta) = c_i + (1 - c_i)P_i(\theta), \qquad (20)$$

where $P_i(\theta)$ is with the 2PLM (see Equation 4).

Item Score.

The item score for calibrations with the 3PLM is

$$\pi_i^* = c_i + (1 - c_i)\pi_i, \qquad (21)$$

where π_i is obtained through Formula 5 for calibrations with the 2PLM. The proof follows directly from multiplying on both sides of Equation 20 by $\varphi(\theta)$ and integrating each side from $-\infty$ to ∞.

Item Error Variance.

The item error variance for calibrations with the 3PLM is

$$\sigma^2(e_i^*) = c_i(1 - c_i)(1 - \pi_i) + (1 - c_i)^2\sigma^2(e_i), \qquad (22)$$

where π_i and $\sigma^2(e_i)$ are obtained trough Formulas 5 and 9, respectively, for calibrations with the 2PLM; (proof in Appendix A). Figure 3 graphically represents values of the item error variance (calculated with Formula 22) as a function of the item parameters a_i and b_i for a fixed value of the pseudo-chance level parameter ($c_i = 0.2$).

Item True Variance.

The item true variance for calibrations with the 3PLM is

$$\sigma^2(\tau_i^*) = \pi_i^*(1 - \pi_i^*) - \sigma^2(e_i^*), \qquad (23)$$

where π_i^* and $\sigma^2(e_i^*)$ are obtained with Formulas 21 and 22, respectively. Formula 23 follows directly from Formula 10 because the derivation of the latter does not depend on which model is used for item calibration (1PLM, 2PLM, or 3PLM).

Item Reliability

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

$$\rho_{ii}^* = \frac{\sigma^2(\tau_i^*)}{\sigma^2(\tau_i^*) + \sigma^2(e_i^*)}, \tag{24}$$

where $\sigma^2(e_i^*)$ and $\sigma^2(\tau_i^*)$ are obtained with Formulas 22 and 23, respectively.

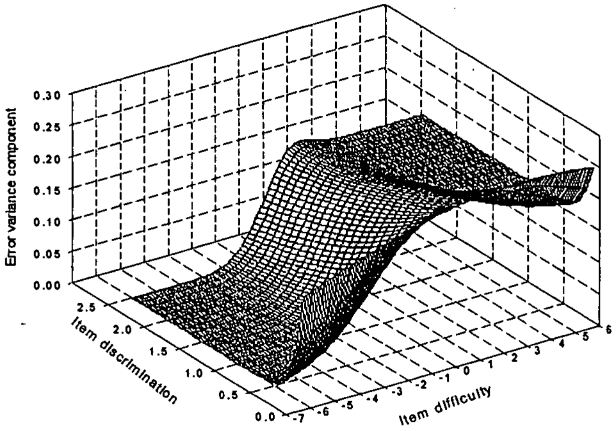


Figure 3. Error variance of binary items as a function of their discrimination (a_i) and difficulty (b_i) parameters for a fixed pseudo-chance level (c_i = 0) with the 3PLM.

True-Score Measures and Reliability at Test Level

Formulas 12, 13, 14, 15, 17, and 18 for true-score measures and reliability at test level with the 2PLM translate directly into their 3PLM counterparts for the *marginal NR score*, *error* *variance for the NR score*, *true score variance*, *reliability*, and dependability (it suffice to use star 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 4 is a constant ($a_i = a$), the 2PLM translates into the 1PLM. With the 1PLM, however, one should know which computational IRT model had been used: logistic (with a scaling constant D = 1.0) or logistic approximation of the normal ogive model (D = 1.7). Both options are provided with some computer programs for calibrations with the 1PLM (e.g., RASCAL; Assessment Systems Corporation, 1995). When the standardization is on the trait scores (D = 1.7), one can use the formulas for true-score measures and reliability (at item and test level) derived here for the 2PLM (a_i= *constant* with the 1PLM). This approach does not work, however, for a "pure" Rasch model (D = 1; Rasch, 1960) in which the standardization is on the item difficulty. For this case, formulas for true-score measures and reliability of binary items are developed by Dimitrov (in press) for 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 purpose of this example is to illustrate the application of the formulas proposed in this article for true-score measures and reliability of binary items calibrated with the 2PLM. The empirical validation of Formulas 5 and 9 [for π_i and $\sigma^2(e_i)$ with the 2PLM] is of particular interest because these two formulas are based on approximations. All other formulas are obtained through exact derivations and represent (explicitly or implicitly) functions of π_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). The assumptions of $\theta \sim N(0,1)$ and model fit with the 2PLM being met with these simulations, the produced binary scores (for 8,000 persons on 20 items) were analyzed using the computer program XCALIBRE (Assessment Systems Corporation, 1995). Using the XCALIBRE estimates of a_i and b_i (given in Table 1) allows us to test the "robustness" of Formulas 5 and 9 when they are used with sample-based (i.e., less than "ideal") estimates of item parameters. The evaluations of true-score measures and reliability in this example were facilitated through the use of the statistical program SPSS (SPSS Inc., 2002). The SPSS syntax developed for this purpose (in Appendix B) works for binary items calibrated with the 3PLM (input variables: a_i, b_i, and c_i), but it also can be used for items calibrated with the 2PLM (with $c_i = 0$) or the 1PLM (with $c_i = 0$ and $\mathrm{a}_i =$ constant). The SPSS run generates the true-score measures and reliability for each item [π_i, $\sigma^2(e_i)$, $\sigma^2(\tau_i)$, and ρ_{ii}] as "new" variables in the SPSS data spreadsheet. At test level, the SPSS printout provides values for the *marginal NR score* (μ), *error variance for the NR score* (σ_e^2), *true variance for the NR score* (σ_τ^2), and *variance of items scores* [$\sigma^2(\pi_i)$].

The results from the SPSS run in this example (with a_i and b_i from Table 1 and $c_i = 0$) are provided in Appendix C. The true-score measures and reliability for individual items (upper panel in Appendix C) are given in Table 1. At test level, the SPSS printout (lower panel in Appendix C) provides the true score variance for the NR score ($\sigma_\tau^2 = 6.315$), the error variance for the NR score ($\sigma_e^2 = 3.719$), the marginal NR score ($\mu = 8.956$), and the variance of π_i values for the 20 items [$\sigma^2(\pi_i) = .045$]. With this, the domain score is $\pi = \mu/n = 8.956/20 = .448$ and the reliability is $\rho_{xx} = .63$ (using Formula 15).

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 π_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 π_i (also in Table 1) is smaller than 0.01 in absolute value. The same is true for the difference between 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 5 and Formula 9 is important because the values of π_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 the item error variance in this example, it is not a surprise then that Cronbach's alpha for the sample of simulated binary scores (N = 8,000) was found to be the same as the theoretically evaluated reliability ($\alpha = \rho_{xx} = .63$). Similarly, the mean and the variance of the empirical item scores in Table 1 [$\overline{p}$ = .448 and $s^2(p_i) = 0.044$] also match their theoretical counterparts [π = 0.448 and $\sigma^2(\pi_i) = 0.045$]. 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.)

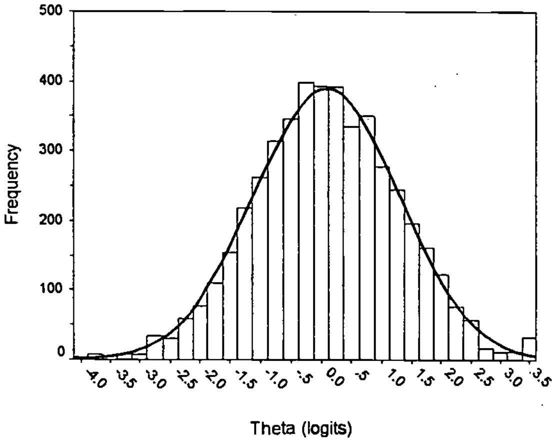
Real Data Example

The data for this example consisted of dichotomously scored responses of 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 is normally distributed and values in excess of 2.0 indicate misfit with a type I error rate of 0.05. In this example, 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 provided in Table 2 (the 24 items are grouped by strands of learning outcomes.)

The normal quantile tests (proportion-proportion and quantile-quantile comparisons of the observed and expected values) were conducted using SPSS with the trait scores, θ, provided by XCALIBRE for the sample data ($N = 4{,}854$). The results indicated a good fit of θ to $N(0,1)$ thus allowing the application of formulas developed in this article (see also Figure 4). The theoretical true-score measures and reliability (at item and test level) 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 17) for values of the cutting score λ on the *proportion of items correct* scale from 0 to 1 with a "step" of 0.01 (see Figure 5). As one can see, 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 is still a good match between theoretical and empirical values for item scores (π_i versus p_i values in Table 2), variance of items scores [$\sigma^2(\pi_i) = .027$ versus $\sigma^2(p_i) = .025$], domain score ($\pi = .585$ versus $\bar{p} = .586$), and reliability ($\rho_{xx} = .789$ versus Cronbach's $\alpha = .801$). Additional comments on ρ_{xx} and its empirical evaluation through Cronbach's α are provided in the discussion part.

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



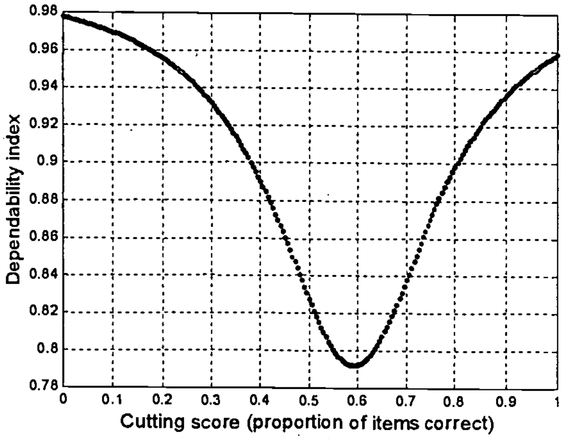


Figure 4. Frequency distribution (with a normal curve fit) of the trait scores for the sample of real data on the OOPT-Reading (N = 4,854).

Figure 5. Dependability index, $\Phi(\lambda)$, as a function of the cutting score, λ, for the OOPT-Reading.

Conclusion

This article provides analytic evaluations (formulas) for marginal 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 the following marginal measures: *item score* (π_i), *item error variance* [$\sigma^2(e_i)$], *item true variance* [$\sigma^2(\tau_i)$], and *item reliability* (ρ_{ii}). At test level, the item true-score measures are "summarized" in formulas for the *marginal NR score* (μ), *domain score* (π), *error variance for the NR score* (σ_e^2), *true variance for the NR score* (σ_τ^2), reliability (ρ_{xx}), and *dependability* [$\Phi(\lambda)$] for criterion-referenced interpretations (e.g., "pass/fail") based on a domain cutting score, λ.

Brief clarifications about the derivation design for the formulas proposed in this article are necessary. For item calibrations with the 2PLM, the formulas for item score (π_i, Formula 5) and item error variance [$\sigma^2(e_i)$, Formula 9] are based on approximations, but the absolute error with these approximations is practically close to zero (less than 0.0005, with Formula 5, and less than 0.005 with Formula 9). All other formulas are obtained through *exact* derivations that (explicitly or implicitly) involve π_i and $\sigma^2(e_i)$ - Formulas 10, 11, 12, 13, 14, 15, 17, 18, 21, 22, and 23. 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 the accuracy. For example, the Simpson's rule for numerical integrations with Equation 4 involves an approximation of the compound binomial distribution of raw scores (e.g., May & Nicewander, 1993) which, in turn, leads to losing accuracy in estimating the true score variance. In contrast, Formula 10 (for item true variance) does not use preliminary approximations. As a reminder, Formula 5 (for π_i) and Formula 9 [for $\sigma^2(e_i)$] use approximations (with an error practically close to zero), whereas all other formulas in this article are based on exact derivations. Along with technical advantages, the formulas provide theoretical relationships that may remain hidden with numerical integrations. Formula 9, for example, shows that the item error variance is an even function of the item difficulty, b_i, for fixed values of the discrimination index, a_i. Also, while Formulas 10 and 23 reveal relationships between true-score measures for item calibrations with the same (e.g., 2PL or 3PL) IRT model, Formulas 21 and 22 connect item true-score measures with the 3PLM to item true-score measures with the 2PLM. The proposed formulas allow researchers to plan (model, predict) true-score measures, whereas the numerical integrations put researchers in a "post-hoc" position. The proposed formulas, therefore, 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 individual binary scores for persons from the target population and (b) provides sample-based estimates which 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 population level without using sample scores (IRT estimates of the item parameters suffice.) It should be emphasized also that Cronbach's alpha is an accurate empirical estimate for reliability (ρ_{xx}) only if there is no correlation among errors and the test components are essentially tau-equivalent (Novick & Lewis, 1967). The evaluation of ρ_{xx} in this article, however, does not require tau-equivalency (the weaker assumption of congeneric items suffice.) 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). It should be also noted that when the tau-equivalency assumption does not hold, Cronbach's alpha underestimates ρ_{xx}. However, Cronbach's alpha may overestimate ρ_{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 ρ_{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, while the *marginal reliability* for IRT trait scores in computerized adaptive testing is evaluated for the population (Green at 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 ρ_{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 of 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, 1995) 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, while "diagnostic" IRT information about trait measures for individuals is valuable, marginal true-score information about the population and the measurement quality of the test is also useful. In a sports analogy, while 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.

References

Allen, J. M., & Yen, W. M. (1979). *Introduction to measurement theory*. Pacific Grove, CA: Brooks/Cole.

Assessment System Corporation (1995). *User's Manual for XCALIBRE marginal maximum-likelihood estimation program (Windows version 1.0)*. St. Paul, MN: Author.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179-197.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*, 277-289.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of a test. *Psychometrika, 16*, 297-334.

Dimitrov, D. M. (in press). Reliability and true-score measures of binary items as a function of their Rasch difficulty parameter. *Journal of Applied Measurement*.

Dimitrov, D. M. (1996, April). *Monte Carlo approach for reliability estimations in generalizability studies*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd. Ed.) Phoenix, AZ: American Council on Education and The Oryx Press (pp. 105-146).

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hambleton, R. K., & Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hastings, C., Jr. (1955). *Approximations for digital computers*. Princeton, NJ: Princeton University Press.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalent and uncorrelated errors on coefficient alpha. *Applied Psychological Measurement, 21*, 337-348.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MathWorks, Inc. (1999). *Learning MATLAB (Version 5.3)*. Natick, MA: Author.

May, K., & Nicewander, W. A. (1993). Reliability and information functions for percentile ranks. *Psychometrika, 58*, 313-325.

National Center for Educational Statistics (1996). *National educational longitudinal study: 1988-94. Data files and electronic codebook system. Base year through third follow-up ECB/CD-ROM.* Washington, DC: Office of Educational Research and Improvement. US Department of Education.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.

Rasch, G. (1960). *Probabilistic models for intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*, 69-76.

Riverside Publishing (1997). *Ohio Off-Grade Proficiency Tests: specifically designed to measure*

Ohio's model course of study. Chicago, IL :Author.

SAS Institute (1985). *SAS user's guide: Version 5 edition*. Cary, NC: Author.

SPSS Inc. (2002). *SPSS Base 11.0 user's guide*. Chicago: Author.

SPSS Inc. (1998). *SigmaPlot 5.0 User's Guide*. Chicago: Author.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.) *Computerized adaptive testing: A primer* (pp. 161-186). Hillsdale, NJ: Lawrence Erlbaum.

Appendix A

Proof of Formula 5.

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

$$\pi_i = \frac{1 - erf(X_i)}{2}, \tag{A1}$$

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

$$erf(X) = (2/\sqrt{\pi})\int_0^X \exp(-u^2)du. \tag{A2}$$

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

$$\pi_i = \frac{1}{\sqrt{2\pi}}\int_{\gamma_i}^{\infty} \exp(-t^2/2)dt, \tag{A3}$$

where $\gamma_i = a_i b_i / \sqrt{1 + a_i^2}$. With the substitution $t = u\sqrt{2}$ (and $\gamma_i = X_i\sqrt{2}$, respectively) we have

$$\pi_i = \frac{1}{\sqrt{\pi}}\int_{X_i}^{\infty} \exp(-u^2)\mathrm{d}u = \frac{1}{2} - \frac{1}{\sqrt{\pi}}\int_0^{X_i} \exp(-u^2)\mathrm{d}u = \frac{1}{2} - \frac{1}{2}erf(X_i),$$

with which the proof is completed.

It should be noted that Formula 5 (or A1) provides an exact theoretical relationship, but it is referred to here as an approximation formula for π_i because the error function, *erf*(X_i), in this formula is evaluated through approximations. With the Hasting's approach (see Equation 6), the approximation error for *erf*(X_i) is smaller than 0.0005 in absolute value. If, however, the *erf*(X_i) is executed through the use of computer programs for mathematics (e.g., MATLAB; MathWorks, Inc.), the absolute approximation error is even smaller than the (practically zero) error of 0.0005.

Proof of Formula 10

Formula 10 represents the item true variance, $\sigma^2(\tau_i)$ as an exact function of the item score, π_i, and item error variance, $\sigma^2(e_i)$. Using the variance expectation rule VAR(X) = E(X^2) - [E(X)]^2 with X = $P_i(\theta)$, we have

$$\sigma^2(\tau_i) = \int_{-\infty}^{\infty} [P_i(\theta)]^2 \varphi(\theta)d\theta - \left(\int_{-\infty}^{\infty} P_i(\theta)\varphi(\theta)d\theta\right)^2$$

$$= \int_{-\infty}^{\infty} \{P_i(\theta) - P_i(\theta)[1 - P_i(\theta)]\}\varphi(\theta)d\theta - \pi_i^2$$

$$= \int_{-\infty}^{\infty} P_i(\theta)\varphi(\theta)d\theta - \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)]\varphi(\theta)d\theta - \pi_i^2$$

$$= \pi_i - \sigma_{e_i}^2 - \pi_i^2 = \pi_i(1 - \pi_i) - \sigma^2(e_i),$$

with which the proof is completed.

Proof of Formula 14

Formula 14 represents the true score variance for the NR score, σ_τ^2, as an exact function of the item score, π_i, and item error variance, $\sigma^2(\tau_i)$. For unidimensional tests (which are dealt with in this article), there is a perfect correlation between the congeneric true scores (τ_i and τ_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 τ_i and τ_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, $\tau = \Sigma\,\tau_i$; (i = 1, ..., n), can be represented as

$$\sigma_\tau^2 = \sum_{i=1}^{n}\sum_{j=}^{n} \sigma(\tau_i, \tau_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma(\tau_i)\sigma(\tau_j). \qquad \text{(A4)}$$

Replacing $\sigma(\tau_i)$ and $\sigma(\tau_j)$ in the far right side of Equation A4 with their equivalent expressions in Formula 10, we obtain Formula 14. With this the proof is completed.

Proof of Formula 22

Formula 22 represents the 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, π_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 20), it can be easily seen that

$$P_i^*(\theta)[1 - P_i^*(\theta)] = c_i(1 - c_i)[1 - P_i(\theta)] + (1 - c_i)^2 P_i(\theta)[1 - P_i(\theta)]. \qquad \text{(A5)}$$

Using Equation A5, the proof of Formula 22 is provided with the following integral manipulations

$$\sigma^2(e_i^*) = \int_{-\infty}^{\infty} P_i^*(\theta)[1 - P_i^*(\theta)]\varphi(\theta)\mathrm{d}\theta$$

$$= c_i(1-c_i)\int_{-\infty}^{\infty} \varphi(\theta)\mathrm{d}\theta - c_i(1-c_i)\int_{-\infty}^{\infty} P_i(\theta)\varphi(\theta)\mathrm{d}\theta$$

$$+(1-c_i)^2\int_{-\infty}^{\infty} P_i(\theta)[1-P_i(\theta)]\varphi(\theta)\mathrm{d}\theta$$

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

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

Appendix B

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

Input variables: IRT item parameters (a_i, b_i, and c_i)

```
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 '  '
   /BRE   AK = (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.)

Appendix C

Results from the SPSS syntax run (Input variables: a_i, b_i from Table 1 and c_i = 0)

Item	*a*	*b*	*c*	*pi*	*var_err*	*var_tau*	*roi*
1	.449	-2.554	.000	.852	.120	.006	.050
2	.402	-2.161	.000	.790	.154	.012	.074
3	.232	-1.551	.000	.637	.220	.011	.047
4	.240	-1.226	.000	.612	.226	.012	.050
5	.610	-.127	.000	.526	.199	.050	.201
6	.551	-.855	.000	.660	.188	.036	.161
7	.371	-.568	.000	.578	.219	.025	.104
8	.321	-.277	.000	.534	.228	.021	.085
9	.403	-.017	.000	.502	.220	.030	.120
10	.434	.294	.000	.454	.215	.033	.131
11	.459	.532	.000	.412	.209	.034	.138
12	.410	.773	.000	.385	.209	.027	.116
13	.302	1.004	.000	.386	.219	.018	.074
14	.343	1.250	.000	.342	.206	.019	.086
15	.225	1.562	.000	.366	.222	.010	.044
16	.215	1.385	.000	.385	.227	.010	.040
17	.487	2.312	.000	.156	.123	.008	.062
18	.608	2.650	.000	.084	.078	.000	.000
19	.341	2.712	.000	.191	.146	.009	.058
20	.465	3.000	.000	.103	.091	.001	.013

Note. pi = π_i ; *var_err* = $\sigma^2(e_i)$; *var_tau* = $\sigma^2(\tau_i)$; *roi* = ρ_{ii}

Report:

True score variance:	6.315
Error variance:	3.719
Marginal NR score:	8.956

Descriptive Statistics

	N	Variance
item score	20	.045

Table 1

True-Score Measures and Reliability for Simulated Binary Items Calibrated with the 2PLM

Item	a_i	b_i	π_i $(p_i)^a$	$\sigma^2(e_i)$	$\sigma^2(\tau_i)$	ρ_{ii}	$p_i - \pi_i$
1	.449	-2.554	.852 (.849)	.120	.006	.050	-.003
2	.402	-2.161	.790 (.785)	.154	.012	.074	-.005
3	.232	-1.551	.637 (.644)	.220	.011	.047	.007
4	.240	-1.226	.612 (.618)	.226	.012	.050	.006
5	.610	-.127	.526 (.526)	.199	.050	.201	-.001
6	.551	-.855	.660 (.653)	.188	.036	.161	-.007
7	.371	-.568	.578 (.577)	.219	.025	.104	-.001
8	.321	-.277	.534 (.534)	.228	.021	.085	.000
9	.403	-.017	.502 (.503)	.220	.030	.120	.001
10	.434	.294	.454 (.456)	.215	.033	.131	.002
11	.459	.532	.412 (.416)	.209	.034	.138	004
12	.410	.773	.385 (.389)	.209	.027	.116	004
13	.302	1.004	.386 (.384)	.219	.018	.074	-.002
14	.343	1.250	.342 (.345)	.206	.019	.086	003
15	.225	1.562	.366 (.360)	.222	.010	.044	-.006
16	.215	1.385	.385 (.379)	.227	.010	.040	-.006
17	.487	2.312	.156 (.163)	.123	.008	.062	.007
18	.608	2.650	.084 (.092)	.078	.000	.000	.008
19	.341	2.712	.191 (.192)	.146	.009	.058	.001
20	.465	3.000	.103 (.099)	.091	.001	.013	-.004

[a] Observed item score (proportion correct responses) for the simulated data ($N = 8{,}000$).

Table 2

True-Score Measures and Reliability by Strands of Learning Outcomes with the OOPT-Reading

Strand Item	a_i	b_i	c_i	π_i $(p_i)^a$	$\sigma^2(e_i)$	$\sigma^2(\tau_i)$	ρ_{ii}	π	σ_e^2	σ_τ^2	ρ_{xx}
Poetry - constructing meaning (*n* = 10)								.664	1.772	3.293	.650
1	1.089	-.732	.209	.767 (.759)	.135	.044	.244				
2	.948	-.418	.220	.698 (.696)	.166	.045	.214				
5	.494	.900	.226	.493 (.495)	.231	.019	.076				
6	.494	.885	.234	.500 (.503)	.231	.019	.075				
7	.905	-.672	.185	.734 (.727)	.154	.041	.212				
8	1.165	-1.144	.205	.847 (.838)	.099	.031	.238				
20	.594	-.412	.209	.670 (.670)	.192	.029	.131				
21	.716	.475	.237	.536 (.542)	.217	.032	.129				
22	.703	-.492	.204	.691 (.689)	.179	.034	.160				
23	.841	-.504	.194	.700 (.696)	.169	.042	.198				
Poetry - extending meaning (*n* = 4)								.596	0.722	0.517	.417
3	1.169	.468	.159	.463 (.470)	.187	.062	.248				
4	.724	-1.541	.211	.855 (.848)	.110	.014	.112				
9	.554	-.042	.197	.605 (.605)	.210	.029	.122				
24	.706	.698	.177	.460 (.463)	.215	.033	.134				
Nonfiction - constructing meaning (*n* = 5)								.529	1.025	0.655	.390
10	.795	-.226	.194	.642 (.641)	.187	.043	.187				
11	.506	1.581	.218	.404 (.406)	.228	.012	.052				
16	.809	-.154	.192	.627 (.626)	.190	.044	.190				
17	.499	2.076	.220	.358 (.362)	.223	.007	.030				
18	.839	.075	.261	.616 (.622)	.198	.039	.164				
Nonfiction - extending meaning (*n* = 5)								.475	0.952	0.520	.353
12	.709	2.238	.190	.269 (.276)	.194	.002	.013				
13	.863	-.727	.221	.753 (.748)	.151	.035	.189				
14	.686	.375	.215	.541 (.545)	.215	.034	.136				
15	.795	.219	.180	.546 (.547)	.203	.044	.179				
19	.812	1.874	.170	.268 (.276)	.188	.008	.041				
Total (*n* = 24)								.585	4.471	16.520	.789

[a] Observed item score (proportion correct responses) for the real data (*N* = 4,854).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:
Expected Values and Reliability of Number-Right Scores for IRT Calibrated Items

Author(s): Dimiter M. Dimitrov

Corporate Source:
George Mason University

Publication Date:
April 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Dimitrov

Printed Name/Position/Title: Dimiter M. Dimitrov, Assoc. Professor

Organization/Address:
George Mason University
Graduate School of Education
4400 Universy Drive, MS 4B3
Fairfax, Virginia 22030-4444

Telephone: 703-993-3842

FAX: 703-993-2013

E-Mail Address: ddimitro@gmu.edu

Date: 09/30/03

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: info@ericfac.piccard.csc.com
WWW: http://ericfacility.org