Asia Pacific Educ. Rev. (2010) 11:179–188
DOI 10.1007/s12564-009-9062-8

The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics

Seonghoon Kim · Leonard S. Feldt

Received: 16 February 2009 / Revised: 10 October 2009 / Accepted: 30 November 2009 / Published online: 28 January 2010

Abstract The primary purpose of this study is to investigate the mathematical characteristics of the test reliability coefficient $\rho_{XX'}$ as a function of item response theory (IRT) parameters and present the lower and upper bounds of the coefficient. Another purpose is to examine relative performances of the IRT reliability statistics and two classical test theory (CTT) reliability statistics (Cronbach's alpha and Feldt–Gilmer congeneric coefficients) under various testing conditions that result from manipulating large-scale real data. For the first purpose, two alternative ways of exactly quantifying $\rho_{XX'}$ are compared in terms of computational efficiency and statistical usefulness. In addition, the lower and upper bounds for $\rho_{XX'}$ are presented in line with the assumptions of essential tau-equivalence and congeneric similarity, respectively. Empirical studies conducted for the second purpose showed across all testing conditions that (1) the IRT reliability coefficient was higher than the CTT reliability statistics; (2) the IRT reliability coefficient was closer to the Feldt–Gilmer coefficient than to the Cronbach's alpha coefficient; and (3) the alpha coefficient was close to the lower bound of IRT reliability. Some advantages of the IRT approach to estimating test-score reliability over the CTT approaches are discussed in the end.

Keywords Test reliability · Item response theory (IRT) · Lower and upper bounds of reliability coefficient · Test score metric versus ability score metric

S. Kim (✉)
Department of Education, Keimyung University,
2800 Dalgubeoldaero, Dalseo-Gu, Daegu 704-701, South Korea
e-mail: seonghoonkim@kmu.ac.kr

L. S. Feldt
The University of Iowa, Iowa City, IA, USA
e-mail: leonard-feldt@uiowa.edu

Even when a test form is developed using item response theory (IRT), practitioners often use the test score (X) metric as in classical test theory (CTT) rather than the ability score (θ) metric as a basis for reporting examinees' scores. This preference is related to the practical problem that the θ metric is often not easily understood by examinees. In this paper, the test score metric refers to the one that is developed by summing item scores, regardless of whether dichotomous or polytomous.

Although not all assumed this context, many studies (e.g., Dimitrov 2003; Kolen et al. 1996; Lord 1980, p. 52; May and Nicewander 1994; Shojima and Toyoda 2002) have presented formulas for computing the IRT counterpart of test score reliability in CTT as a function of known item parameters and some distribution of ability. However, as described in detail later, the formulas shown for quantifying IRT-based test score reliability (simply referred to as IRT test reliability) are not based on the same assumptions and thus do not lead to the same reliability coefficient. In fact, some of the formulas (e.g., Kolen et al. 1996; May and Nicewander 1994) are for exactly quantifying the IRT test reliability coefficient (denoted $\rho_{XX'}$) and others (e.g., Dimitrov 2003; Shojima and Toyoda 2002) are for approximating the exact coefficient.

The primary purpose of this study is to investigate the mathematical characteristics of the IRT test reliability coefficient and present its lower and upper bounds. For this, two alternative ways of exactly quantifying $\rho_{XX'}$ are compared in terms of computational efficiency and statistical usefulness. In addition, the lower and upper bounds for $\rho_{XX'}$ are derived under the assumptions of essential tau-equivalence and congeneric similarity, respectively. In this presentation, it is shown that IRT test unidimensionality does not necessarily imply tau-equivalence or congeneric similarity in true scores across items. Another purpose is to

examine relative performances of the IRT reliability statistics and two CTT-based reliability statistics (Cronbach's alpha and Feldt-Gilmer congeneric coefficients) under various testing conditions that result from manipulating large-scale real data of mixed-format tests. Before studying the relative performances, a theoretical study is conducted to show how IRT item- and test-level reliability statistics vary depending on the configuration of the discrimination, difficulty, and guessing parameters of dichotomously scored items.

So far, most of the formulas for exactly quantifying or approximating $\rho_{XX'}$ have been presented under dichotomous IRT models. For generality, this study presents the formulas assuming polytomous IRT models.

The IRT test reliability coefficient

Let X_i be the observed score variable for a response on item i in an n-item test. The observed test score considered in this paper is defined as $X = \sum_{i=1}^{n} X_i$. In addition, denote as $P_{ik}(\theta)$ the IRT response function (i.e., probability) for category k of item i conditional on ability θ. It is assumed throughout this paper that (a) increasing score weights W_{ik} are given to the $m_i(\geq 2)$ response categories of item i and that (b) the category index k ranges from 1 to m_i. The following notation is adopted: (a) E(•) and Var(•) are used as the expectation and variance operators, (b) μ_Z and σ_Z^2 indicate the values of mean and variance for variable Z, and (c) σ_{YZ} and ρ_{YZ} refer to the values of covariance and correlation between variables Y and Z. Lastly, define the item response function (IRF) for item i, $T_i(\theta)$, as

$$T_i(\theta) \equiv \mathrm{E}(X_i|\theta) = \sum_{k=1}^{m_i} W_{ik} P_{ik}(\theta). \tag{1}$$

Definition of test reliability

Conceptually, test reliability is based on the fundamental measurement model of $X = T + e$, where T and e are the true and error scores (Feldt and Brennan 1989). This model can be applied to item scores such as $X_i = T_i + e_i$, and thus

$$X = T + e = \sum_{i=1}^{n} T_i + \sum_{i=1}^{n} e_i. \tag{2}$$

In CTT, the true score T is often defined as E(X), where X has a known or assumed frequency distribution (Haertel 2006; Lord 1980). The true score T is a constant for a given examinee but becomes a variable for a population of examinees (Feldt and Brennan 1989). This definition for T may be applied to the true item score T_i in the same manner. In IRT, the ability θ is an alternative expression of T_i or T and can be nonlinearly transformed into the true item score $T_i(\theta)$ or the true test score $T(\theta)$ by the IRF or the test response function (TRF). Based on Eq. 1, the TRF is expressed, conditional on θ, as

$$\begin{aligned} T(\theta) = \mathrm{E}(X|\theta) &= \mathrm{E}\left(\sum_{i=1}^{n} X_i|\theta\right) = \sum_{i=1}^{n} T_i(\theta) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{m_i} W_{ik} P_{ik}(\theta). \end{aligned} \tag{3}$$

To indicate that T and θ are the same thing expressed on different metrics of measurement (Lord 1980), Eq. 3 can be more explicitly expressed as $T(\theta) = \mathrm{E}[X|\theta, T(\theta)] = \mathrm{E}[X|T(\theta)] = \mathrm{E}(X|\theta)$.

The definition of true scores leads to two important statistical relations for each of the test and item scores. First, in expectations, the following are derived by the definition:

$$\mu_T \equiv \mathrm{E}(T) = \mathrm{E}[\mathrm{E}(X|\theta)] = \mathrm{E}(X) \equiv \mu_X,$$
$$\mu_{T_i} \equiv \mathrm{E}(T_i) = \mathrm{E}[\mathrm{E}(X_i|\theta)] = \mathrm{E}(X_i) \equiv \mu_{X_i}.$$

Further, it can be shown by the definition that, over examinees, the observed score variance equals the sum of true score variance and error variance (see Holland and Hoskens 2003; Lord 1980, p. 4):

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2, \tag{4}$$
$$\sigma_{X_i}^2 = \sigma_{T_i}^2 + \sigma_{e_i}^2. \tag{5}$$

The test reliability coefficient, then, is defined as the ratio of true-score variance to observed-score variance, σ_T^2/σ_X^2, which is equivalent to the squared correlation between T and X, ρ_{TX}^2. From the perspective of nonlinear regression, the ratio can also be interpreted as the correlation ratio of test score X on ability θ (Lord 1980, p. 52). This interpretation reflects the facts that (a) $\rho_{XX'} = \sigma_T^2/\sigma_X^2 = 1 - \sigma_e^2/\sigma_X^2$, (b) the conditional error variance $\sigma_{e|\theta}^2$, given θ, equals the conditional test score variance $\sigma_{X|\theta}^2$ and (c) the test error variance σ_e^2 is the average over an ability distribution (strictly speaking, density function) $g(\theta)$ of the conditional error variances:

$$\begin{aligned} \sigma_e^2 &= \mathrm{E}[\mathrm{Var}(e|\theta)] + \mathrm{Var}[\mathrm{E}(e|\theta)] = \mathrm{E}[\mathrm{Var}(e|\theta)] \\ &= \int_{-\infty}^{+\infty} \sigma_{e|\theta}^2 \, g(\theta)\, \mathrm{d}\theta, \end{aligned} \tag{6}$$

where $\mathrm{E}(e|\theta) = 0$ for all θ by definition and thus $\mathrm{Var}[\mathrm{E}(e|\theta)] = 0$ (Holland and Hoskens 2003). Based on Eqs. 2 and 5, the variances σ_X^2 and σ_T^2 can be decomposed of item-level variances and covariances, as follows:

$$\sigma_X^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{X_iX_j} = \sum_{i=1}^{n}\sigma_{X_i}^2 + \sum_{i\neq j}^{n}\sum^{n}\sigma_{X_iX_j} \tag{7}$$

$$\sigma_T^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{T_iT_j} = \sum_{i=1}^{n}\sigma_{T_i}^2 + \sum_{i\neq j}^{n}\sum^{n}\sigma_{T_iT_j} \tag{8}$$

Therefore, the test reliability coefficient is expressed as

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2} = \frac{\sum_{i=1}^{n}\sigma_{T_i}^2 + \sum_{i\neq j}^{n}\sum^{n}\sigma_{T_iT_j}}{\sum_{i=1}^{n}\sigma_{X_i}^2 + \sum_{i\neq j}^{n}\sum^{n}\sigma_{X_iX_j}}. \tag{9}$$

Two approaches to concretizing the IRT test reliability coefficient

The definition of $\rho_{XX'}$ in Eq. 9 remains only a conceptual notion unless σ_T^2 and σ_e^2 are concretized in a reasonable manner. Unlike CTT, IRT has a framework for quantifying σ_T^2, σ_e^2, and σ_X^2 as a function of item parameters and some ability distribution $g(\theta)$, based on the assumptions of model-data fit and local independence. It should be noted that the local independence assumption in IRT replaces the uncorrelated errors assumption in CTT. Because of the relation in Eq. 4, computation of any two of the three variances is enough for computing $\rho_{XX'}$. This paper considers two approaches to computing $\rho_{XX'}$; one is based on quantifying σ_T^2 and σ_e^2 (e.g., May and Nicewander 1994), and the other concerns quantifying σ_X^2 and σ_e^2 (e.g., Kolen et al. 1996). In this paper, the terms true score variance approach and observed score variance approach are used to refer to the first and second approaches, respectively.

For the true score variance approach, σ_e^2 is computed as follows. By local independence, the two conditional covariances between item i and j, $\sigma_{T_ie_j|\theta}$ and $\sigma_{e_ie_j|\theta}$, are equal to zero (which leads to $\sigma_{T_ie_j} = \sigma_{e_ie_j} = 0$). Therefore, the conditional error variance given θ is

$$\sigma_{e|\theta}^2 = \sum_{i=1}^{n}\sigma_{e_i|\theta}^2 = \sum_{i=1}^{n}\sigma_{X_i|\theta}^2, \tag{10}$$

where

$$\sigma_{X_i|\theta}^2 = \sum_{k=1}^{m_i} W_{ik}^2 P_{ik}(\theta) - [T_i(\theta)]^2.$$

Note that for dichotomously scored (0 or 1) items, the conditional error variance in Eq. 10 is simply expressed as

$$\sigma_{e|\theta}^2 = \sum_{i=1}^{n} P_{i2}(\theta)[1 - P_{i2}(\theta)],$$

where $P_{i2}(\theta)$ is the IRF for the correct response (i.e., $k = 2$) and is equal to $T_i(\theta)$. Then, σ_e^2 is obtained by incorporating Eq. 10 into Eq. 6. Alternatively, due to $\sigma_{e_ie_j} = 0$ derived by local independence, σ_e^2 can be obtained by the sum of item error variances:

$$\sigma_e^2 = \sum_{i=1}^{n}\sigma_{e_i}^2, \tag{11}$$

where

$$\sigma_{e_i}^2 = \int_{-\infty}^{+\infty}\sigma_{X_i|\theta}^2 g(\theta)\,\mathrm{d}\theta. \tag{12}$$

On the other hand, σ_T^2 is computed using the statistical definition of variance based on the true score distribution; that is,

$$\sigma_T^2 = \mathrm{E}(T^2) - [\mathrm{E}(T)]^2 = \int_{-\infty}^{+\infty} T^2(\theta)g(\theta)\,\mathrm{d}\theta - \mu_T^2, \tag{13}$$

where

$$\mu_T = \int_{-\infty}^{+\infty} T(\theta)g(\theta)\,\mathrm{d}\theta.$$

By Eq. 8, σ_T^2 can be also computed as the sum of item-level variances ($\sigma_{T_i}^2$) and covariances ($\sigma_{T_iT_j}$), where

$$\sigma_{T_i}^2 = \mathrm{E}(T_i^2) - [\mathrm{E}(T_i)]^2 = \int_{-\infty}^{+\infty} T_i^2(\theta)g(\theta)\,\mathrm{d}\theta - \mu_{T_i}^2, \tag{14}$$

$$\sigma_{T_iT_j} = \mathrm{E}(T_iT_j) - \mathrm{E}(T_i)\mathrm{E}(T_j) = \int_{-\infty}^{+\infty} T_i(\theta)T_j(\theta)g(\theta)\,\mathrm{d}\theta - \mu_{T_i}\mu_{T_j}. \tag{15}$$

The coefficient $\rho_{XX'}$ is then computed as $\sigma_T^2/(\sigma_T^2 + \sigma_e^2)$.

The observed score variance approach computes the conditional frequency distribution of X given θ, $f(x|\theta)$, to obtain the test error variance σ_e^2. By the assumptions made previously, the values of X range from $\sum_i^n W_{i1}$ to $\sum_i^n W_{im_i}$. From $f(x|\theta)$, one can calculate the conditional mean ($\mu_{X|\theta}$ = true score) and variance ($\sigma_{X|\theta}^2$) of X. Because $\sigma_{e|\theta}^2 = \sigma_{X|\theta}^2$, as described earlier, σ_e^2 is obtained by averaging the conditional variances over the ability distribution by Eq. 6. The observed score variance σ_X^2 is obtained using the (marginal) frequency distribution of X, $f(x)$, which is calculated by integrating $f(x|\theta)$ over the ability distribution:

$$f(x) = \int_{-\infty}^{+\infty} f(x|\theta)g(\theta)\,\mathrm{d}\theta. \tag{16}$$

With the σ_e^2 and σ_X^2 obtained, the coefficient $\rho_{XX'}$ is computed as $1 - \sigma_e^2 / \sigma_X^2$.

The critical task in the observed score variance approach is to compute $f(x|\theta)$. As noted in the literature, $f(x|\theta)$ follows a compound multinomial distribution, under the assumption that IRFs vary across items. Unfortunately, the probability generating function for a compound multinomial distribution has not been expressed in a tractable form, and thus there is no simple way to calculate $f(x|\theta)$ using the generating function. Instead, the conditional distribution can be calculated iteratively using the recursive algorithm described by, for example, Thissen et al. (1995) and Kolen and Brennan (2004, p. 219). Although not detailed here, the recursive algorithm basically does, at a given ability, the following: (a) across items, it finds all possible combinations of item categories that lead to a specific score x, (b) it calculates the product of values of IRFs for each combination by local independence, and then (c) it sums all the products across combinations.

Relative merits

Theoretically, the two approaches to computing $\rho_{XX'}$ lead to the same coefficient. However, each has relative merits over the other, either in terms of computational efficiency or statistical usefulness of intermediate results from the computation.

The true score variance approach would be more efficient than the observed score variance approach because the latter involves a computationally intensive iterative algorithm to obtain the conditional frequency distribution $f(x|\theta)$. This is almost always true if the true score variance σ_T^2 is computed using Eq. 13. Another advantage of the true score variance approach is that it allows one to compute the IRT-based item score reliability coefficient (simply, IRT item reliability coefficient)

$$\rho_{X_iX_i'} = \frac{\sigma_{T_i}^2}{\sigma_{X_i}^2} = \frac{\sigma_{T_i}^2}{\sigma_{T_i}^2 + \sigma_{e_i}^2} \tag{17}$$

for each item, when σ_T^2 is computed as the sum of item-level variances ($\sigma_{T_i}^2$) and covariances ($\sigma_{T_iT_j}$), based on Eqs. 14 and 15. Of course, this advantage is obtained at the cost of computational efficiency associated with Eq. 13.

The observed score variance approach would be more appealing to practitioners, because it deals with observed score distributions (fitted by IRT models). Once the conditional frequency distribution $f(x|\theta)$ is obtained, one can compute any statistical value of interest for the raw score X or its transformed score S(X). For example, a practitioner can compute the mean and variance of X to obtain the test true score $T(\theta)$ and conditional error variance $\sigma_{X|\theta}^2$ at a given value of θ. Such computation may also be applied to S(X) (see Kolen et al. 1996, for more applications). In addition, consideration of $f(x|\theta)$ over the θ continuum clarifies that $T(\theta)$ is the regression of X on θ (that is, $\mu_{X|\theta}$), and the test reliability coefficient is the proportion of the test score variance (nonlinearly) accounted for by θ. Another merit of the observed score variance approach is the simplicity in computation of σ_X^2 based on $f(x)$. Of course, these advantages are secured at the cost of intensive computation of both $f(x|\theta)$ and $f(x)$.

The lower and upper bounds of the IRT test reliability coefficient

The IRT test reliability coefficient can be determined by either approach described earlier, but determination of the lower and upper bounds of the coefficient adds to our understanding of it. In fact, the sources for the lower and upper bounds are found in Shojima and Toyoda (2002) and Dimitrov (2003), respectively, although the researchers did not explicitly state or recognize it. The details are described next.

Shojima and Toyoda (2002) presented the IRT counterpart of Cronbach's (1951) coefficient alpha under the assumption of tau-equivalence across items. Using the notation used in this paper, the IRT-based alpha coefficient is expressed as

$$_{\alpha}\rho_{XX'} = \frac{[n/(n-1)]\sum_{i\neq j}^{n}\sum^{n}\sigma_{X_iX_j}}{\sigma_X^2}, \tag{18}$$

where, by local independence and the equation $\mu_{X_i} = \mu_{T_i}$, $\sigma_{X_iX_j}$ ($i \neq j$) equals $\sigma_{T_iT_j}$ in Eq. 15. Designate the numerator in Eq. 18 by $_{\alpha}\sigma_T^2$, which is the true score variance approximated by the tau-equivalence assumption. Then, the error score variance is derived by $_{\alpha}\sigma_e^2 = \sigma_X^2 - {}_{\alpha}\sigma_T^2$. Note that Eq. 18 is in form identical to the classical equation for Cronbach's alpha. Thus, as well documented in the literature (e.g., Novick and Lewis 1967),

$$\rho_{XX'} \geq {}_{\alpha}\rho_{XX'}, \tag{19}$$

because

$$\sum_{i}^{n}\sigma_{T_i}^2 + \sum_{i\neq j}^{n}\sum^{n}\sigma_{T_iT_j} \geq [1/(n-1)]\sum_{i\neq j}^{n}\sum^{n}\sigma_{T_iT_j} + \sum_{i\neq j}^{n}\sum^{n}\sigma_{T_iT_j} = [n/(n-1)]\sum_{i\neq j}^{n}\sum^{n}\sigma_{X_iX_j}.$$

The lower bound $_{\alpha}\rho_{XX'}$ equals $\rho_{XX'}$ when the tau-equivalent relation $T_i(\theta) = T_j(\theta) + C_{ij}$ holds for every pair of items i and j, where the constant C_{ij} depends only on the item pair. In IRT, the equality holds only when all items have the same IRF, which is rarely the case in practice.

Dimitrov (2003, p. 455, Eq. A4) adopted the congeneric similarity assumption to compute σ_T^2. Specifically, he regarded

IRT test unidimensionality as congeneric similarity in true item scores and thus equalized the covariance of T_i and T_j as

$$\sigma_{T_iT_j} = \sigma_{T_i}\sigma_{T_j}. \tag{20}$$

As a result, the congeneric similarity-based coefficient ${}_C\rho_{XX'}$ may be expressed as

$$\begin{aligned} {}_C\rho_{XX'} &= \frac{\sum_{i=1}^{n}\sigma_{T_i}^2 + \sum_{i\neq j}^{n}\sum^{n}\sigma_{T_iT_j}}{\sum_{i=1}^{n}\sigma_{T_i}^2 + \sum_{i\neq j}^{n}\sum^{n}\sigma_{T_iT_j} + \sigma_e^2} = \frac{\sum_{i}^{n}\sum_{j}^{n}\sigma_{T_i}\sigma_{T_j}}{\sum_{i}^{n}\sum_{j}^{n}\sigma_{T_i}\sigma_{T_j} + \sigma_e^2} \\ &\equiv \frac{{}_C\sigma_T^2}{{}_C\sigma_X^2}, \end{aligned} \tag{21}$$

where ${}_C\sigma_T^2\left(=\sum\sum\sigma_{T_i}\sigma_{T_j}\right)$ is the true score variance approximated by the congeneric similarity assumption and ${}_C\sigma_X^2(={}_C\sigma_T^2+\sigma_e^2)$ is the observed score variance defined by that assumption, as in Dimitrov (2003). Equation 20 implies that the true item scores are linearly related such that $T_i = \lambda_{ij}T_j + D_{ij}$, where the constants λ_{ij} and D_{ij} depend only on the item pair. However, this is rarely true under IRT test unidimensionality unless the two items have the same IRF, because $T_i(\theta)$ and $T_j(\theta)$ are usually nonlinearly related (see Meredith 1965). In essence, IRT test unidimensionality does not necessarily suggest congeneric similarity in true item scores. We conclude that Dimitrov (2003) over-approximated σ_T^2. This leads to over-approximation of σ_X^2, as determined under the exact IRT approach. Therefore, Eq. 20 should be corrected into

$$\sigma_{T_iT_j} = \rho_{T_iT_j}\sigma_{T_i}\sigma_{T_j} \leq \sigma_{T_i}\sigma_{T_j}. \tag{22}$$

It follows from Eq. 22 that

$$\rho_{XX'} \leq {}_C\rho_{XX'}, \tag{23}$$

because the relationship $\sigma_T^2 \leq {}_C\sigma_T^2$ implies that

$$\frac{\sigma_T^2}{\sigma_T^2+\sigma_e^2} \leq \frac{{}_C\sigma_T^2}{{}_C\sigma_T^2+\sigma_e^2}.$$

To summarize, ${}_\alpha\rho_{XX'}$ is the lower bound of $\rho_{XX'}$ because ${}_\alpha\sigma_T^2 \leq \sigma_T^2$ given the same σ_X^2, whereas ${}_C\rho_{XX'}$ is the upper bound of $\rho_{XX'}$ because $\sigma_T^2 \leq {}_C\sigma_T^2$ given the same error variance σ_e^2. Note that the error variance for ${}_\alpha\rho_{XX'}$ is derived by subtracting ${}_\alpha\sigma_T^2$ from the exact observed score variance σ_X^2 for $\rho_{XX'}$ and the observed score variance for ${}_C\rho_{XX'}$ is obtained by adding ${}_C\sigma_T^2$ to the exact error variance σ_e^2 for $\rho_{XX'}$.

Practical issues

The preceding sections assume implicitly that IRT test reliability is evaluated in a population. That is, the formulas and discussion for IRT test reliability have been presented assuming that item parameters and an underlying ability distribution are known. In practice, however, the item parameters are usually unknown. Therefore, they must be estimated from sample data using, for example, marginal maximum likelihood estimation via the EM algorithm (Bock and Aitkin 1981; Harwell et al. 1988; Thissen 1982; Woodruff and Hanson 1996) or Bayes modal estimation via the EM algorithm (Harwell and Baker 1991; Mislevy 1986; Tsutakawa and Lin 1986). The ability distribution is often taken a priori as a unit normal distribution [denoted N(0, 1)], but can be estimated from sample data parametrically (Mislevy 1984) or discretely (Bock and Aitkin 1981; Woodruff and Hanson 1996) using the EM algorithm. The maximum likelihood estimates of relative weights (i.e., probabilities) at discrete ability points are often labeled "posterior" weights in the literature (Mislevy and Bock 1990).

When the estimates of item parameters and ability distribution are used for calculating IRT test reliability, the resulting coefficient becomes an estimate ($\hat{\rho}_{XX'}$) of the population reliability coefficient $\rho_{XX'}$. This IRT-based estimation is analogous to estimation in CTT of a reliability coefficient using sample variances and covariances. The IRT reliability coefficient estimates would fluctuate from sample to sample and yield a sampling distribution. But little is known about the sampling distribution.

The variances and covariances that are used for evaluating IRT test reliability involve computation of integrals over a continuous ability distribution. In general, the integrals cannot be computed analytically but are estimated numerically. Although many numerical integration techniques have been suggested (see, e.g., Press et al. 1992), Gaussian quadrature using a set of discrete abscissas (points) and weights would suffice for the computation of variances and covariances. This is true because the integrands have a form of $q(\theta)W(\theta)$, where $W(\theta)$ corresponds to the underlying ability distribution $g(\theta)$. When the ability distribution is assumed as a N(0, 1), Gauss-Hermite quadrature should be suitable for highly precise approximation of the integrals. Incidentally, the ability distribution often is presented using a set of (equally) spaced points and weights. In this case, integration is simply replaced with sum to compute variances and covariances.

Theoretical and practical studies

Two studies were conducted to investigate theoretical characteristics of the IRT item- and test-level reliability statistics presented earlier and to examine relative performances of the IRT and CTT test reliability coefficients using real data. The first study concerned dichotomously scored multiple choice (MC) item tests and used the three-parameter logistic (3PL) model (Birnbaum 1968) to

manipulate the degrees of discrimination, difficulty, and guessing for the MC items. The second study used the 3PL model and the generalized partial credit (GPC) model (Muraki 1992) to deal with mixed-format tests that contained both MC and polytomously scored constructed response (CR) items. For simplicity, the MC and CR parts (i.e., subtests) of a mixed-format test (analyzed by the 3PL and GPC models, respectively) were referred to as the 3PL and GPC tests, respectively. The mixed-format test, as a whole, was referred to as the 3PL + GPC test. A computer program, called IRT_REL, was developed to compute IRT and CTT reliability statistics used for both studies. The computer program is available upon request from the authors. To analyze the MC and CR items used for the second study, the computer program PARSCALE (Muraki and Bock 2003) was employed. Note that PARSCALE can be effectively used not only for calibration of a single-format test (such as the 3PL and GPC tests) but also for calibration of a mixed-format test (such as the 3PL + GPC test), because the program allows for specifying distinct IRT models across items in a test.

Table 1 IRT reliability statistics for unit items and IRT reliability for parallel-items tests

a	b	c	Unit item statistic				$\rho_{XX'}$ of parallel items		
			$\sigma^2_{T_i}$	$\sigma^2_{e_i}$	$\sigma^2_{X_i}$	$\rho_{X_iX'_i}$	$n = 10$	30	50
0.5	−1.0	0.0	.027	.191	.219	.125	.589	.811	.877
		0.2	.018	.174	.192	.091	.502	.751	.834
	0.0	0.0	.034	.216	.250	.136	.612	.825	.887
		0.2	.022	.218	.240	.091	.500	.750	.833
	1.0	0.0	.027	.191	.219	.125	.589	.811	.877
		0.2	.018	.231	.248	.071	.432	.695	.792
1.0	−1.0	0.0	.055	.128	.183	.301	.812	.928	.956
		0.2	.035	.120	.156	.227	.746	.898	.936
	0.0	0.0	.084	.166	.250	.336	.835	.938	.962
		0.2	.054	.186	.240	.224	.743	.897	.935
	1.0	0.0	.055	.128	.183	.301	.812	.928	.956
		0.2	.035	.203	.238	.148	.634	.839	.897
1.5	−1.0	0.0	.072	.091	.163	.440	.887	.959	.975
		0.2	.046	.091	.137	.335	.834	.938	.962
	0.0	0.0	.121	.129	.250	.483	.903	.966	.979
		0.2	.077	.163	.240	.322	.826	.934	.960
	1.0	0.0	.072	.091	.163	.440	.887	.959	.975
		0.2	.046	.186	.231	.198	.712	.881	.925

Study 1

Method

The purpose of this study was to show how the IRT item- and test-level reliability coefficients vary depending on the configuration of 3PL item parameters (a, b, and c) given a N(0, 1) ability distribution. Considering a variety of 3PL items in terms of the a, b, and c parameters, 18 types of unit MC items were created by fully crossing three a's (.5, 1, 1.5), three b's (−1, 0, 1), and two c's (0, .2). For each of the 18 unit MC items, the item-level IRT statistics of true score variance ($\sigma^2_{T_i}$), error variance ($\sigma^2_{e_i}$), observed score variance ($\sigma^2_{X_i}$), and reliability coefficient ($\rho_{X_iX'_i}$) were computed. To investigate the effects of test length on IRT test reliability coefficients, each unit item was replicated 10, 30, and 50 times, so that 10-, 30-, and 50-parallel-item MC tests might be constructed. All the IRT statistics for each item/test were computed using Gauss–Hermite quadrature with 51 points so that results with high precision might be obtained under the N(0, 1) ability distribution.

Results and discussion

Table 1 shows that the IRT item- and test-level reliability statistics are functions of item parameters, given the N(0, 1) ability distribution. For the 18 types of unit items, interesting results are found. For given values of b and c, the IRT item reliability coefficient $\rho_{X_iX'_i}$ becomes higher (due to an increase in $\sigma^2_{T_i}$ against a decrease in $\sigma^2_{e_i}$) as the value of a increases. For example, with $b = -1$ and $c = 0$, $\rho_{X_iX'_i}$ changes from .125 to .301 to .440 as a increases from .5 to 1 to 1.5. For given values of a and b, $\rho_{X_iX'_i}$ becomes lower as the value of c increases. The reason for $\rho_{X_iX'_i}$ becoming lower is differential depending on the value of b; with $b = -1$, $\sigma^2_{T_i}$ decreases by much more than does $\sigma^2_{e_i}$, whereas with both $b = 0$ and $b = 1$, $\sigma^2_{T_i}$ decreases but $\sigma^2_{e_i}$ increases. The decrease of $\sigma^2_{T_i}$ is related to the observation that the range of item true scores (T_i) becomes shorter as the value of c gets higher. For a given value of a and $c = 0$, $\rho_{X_iX'_i}$ becomes higher as the distance between the value of b and the mean (i.e., zero) of ability distribution gets closer. Thus, for example, with $a = .5$ and $c = 0$, $\rho_{X_iX'_i}$ is the highest at $b = 0$ and is the same both at $b = -1$ and $b = 1$. In contrast, for a given value of a and $c = .2$, $\rho_{X_iX'_i}$ is the highest at $b = -1$ and the lowest at $b = 1$. This finding suggests that nonzero values of c interact, somewhat unpredictably, with both the item and ability-distribution locations to result in $\rho_{X_iX'_i}$. However, it is noteworthy here that $\sigma^2_{T_i}$ is the same both at $b = -1$ and $b = 1$ even when $c = .2$ (for example, with $a = .5$ and $c = .2$, $\sigma^2_{T_i} = .018$ at $b = -1$ and $b = 1$). This equality may be explained using the following relations; at ability points $-\theta_q$ and $+\theta_q$, which are symmetric around mean ability, $g(-\theta_q) = g(\theta_q)$ and

$$\left| P_i(-\theta_q|a, b = 1, c = .2) - .5(1 + c) \right| = \left| P_i(+\theta_q|a, b = -1, c = .2) - .5(1 + c) \right|,$$

where $.5(1 + c)$ is the midpoint for a true-score interval $[c, 1]$.

As can be seen at the right columns of Table 1, the IRT test reliability coefficient $\rho_{XX'}$ increases as the test length increases. For example, with parallel items that are all of $a = .5, b = -1$, and $c = 0$, $\rho_{XX'}$ changes from .589 to .811 to .877 as the test length increases from 10 to 30 to 50. This trend is precisely what one would predict from the Spearman–Brown formula. It is noteworthy that for the unit item with $a = 1.5, b = 0$, and $c = 0$, $\rho_{XX'}$ is greater than .9 even when the test length is 10. As expected, the relative sizes of test reliability coefficients among the "18 types" of tests are exactly the same as those of item reliability coefficients among the 18 types of unit items.

Overall, the results from study 1 confirm the conventional wisdom associated with CTT. These results suggest that to enhance test score reliability test developers should use MC items that have high discrimination, are affected little by guessing, and have difficulty matching to the average ability level of a target examinee group.

Study 2

Method

This study was conducted to examine relative performances of IRT and CTT approaches to estimating test reliability coefficients using random samples from real data sets of a large-scale test battery for middle school students. For this study, the test data for three subject areas were considered: English, Mathematics, and Science. The English, Mathematics, and Science tests each consisted of both MC and CR items (i.e., mixed-format tests in nature) and were administered to about 25,000 examinees. The English and Science tests each had 30 MC and 5 CR items and the Mathematics test had 25 MC and 4 CR items. All the MC items were scored 0 or 1 (incorrect or correct), whereas all the CR items were scored 0, 1, …, or $K - 1$ (the number of categories minus 1). The numbers of response categories for the 5 English CR items, in vector expression, were (3, 3, 3, 3, 5), those for the 4 Mathematics CR items were (4, 5, 5, 5), and those for the 5 Science CR items were (4, 4, 4, 4, 4).

To manipulate the degree of heterogeneity among items, each of the three tests was divided into three subtests: (1) a 3PL test consisting only of MC items, (2) a GPC test consisting only of CR items, and (3) a 3PL + GPC test including both MC and CR items. The qualifiers, 3PL, GPC, and 3PL + GPC, for the three subtests were simply adopted to indicate which models were used for the analysis of item response data, as in Study 1. To investigate effects of sample size on estimates of reliability statistics, sample size conditions of 200, 1,000, and 5,000 were considered. Crossing the 3 sample-size conditions with the 9 subtests (3 subtests per subject area) resulted in a total of 27 testing conditions. At each testing condition, 30 samples of item response data were created by randomly sampling from the total data set of the about 25,000 examinees, so as to take into account sampling fluctuations of reliability coefficient estimates.

With each of the 810 random sample data sets, both IRT and CTT approaches to estimating test reliability statistics were performed. For the IRT approach, the (exact) IRT test reliability coefficient and its lower and upper bounds were estimated for each set of sample data. The reliability analysis was based on the estimates of item parameters and ability distribution for 41 quadrature points that were produced via PARSCALE. On the other hand, for the CTT approach, Cronbach's coefficient alpha and Gilmer and Feldt's (1983) congeneric coefficient (labeled Feldt–Gilmer coefficient) were evaluated with each sample data set. The Feldt–Gilmer coefficient is obtained through the estimation of the effective or functional length λ_i of item i. Congeneric similarity assumes that item and test true scores are related as $T_i = \lambda_i T + C_i$, where the constant C_i depends only on the item. Based on the estimated values of λ_i, the Feldt–Gilmer coefficient is computed as

$$_{FG}\rho_{XX'} = \frac{1}{1 - \sum_{i=1}^{n} \hat{\lambda}_i^2} \left(1 - \frac{\sum_{i=1}^{n} \sigma_{X_i}^2}{\sigma_X^2} \right)$$

(see Feldt 2002; Gilmer and Feldt 1983, for the derivation and details of the formula).

Results and discussion

Table 2 presents means of IRT and CTT reliability statistics over 30 random samples from the large-scale assessment real data by subject area, test type, and sample size. For the sample estimates ($\hat{\rho}_{XX'}$) of the exact IRT test reliability coefficient, their standard deviation (SD) is also provided at each testing condition. As dictated by theory, it is verified in Table 2 that an IRT reliability coefficient is greater than its lower bound and is less than its upper bound, over all testing conditions. As expected by the sampling theory, the SD of $\hat{\rho}_{XX'}$ becomes smaller as the sample size (N) increases, although the means do not differ much among the sample sizes, particularly between $N = 1{,}000$ and $N = 5{,}000$. For example, for the English 3PL test, the SD decreases from .007 to .004 to .001 as the sample size increases from 200 to 1,000 to 5,000. Note that the SD of $\hat{\rho}_{XX'}$ is not very large

Table 2 Average values of IRT and CTT reliability statistics over 30 random samples from large-scale assessment real data by subject, test type, and sample size (SS)

Subject	Test (*n*)	SS	IRT test reliability			Cronbach alpha	Feldt-Gilmer
			Lower B.	Exact (SD)	Upper B.		
English	3PL (30)	200	.902	.906 (.007)	.912	.904	.905
		1,000	.902	.905 (.004)	.912	.902	.903
		5,000	.902	.906 (.001)	.912	.902	.903
	GPC (5)	200	.711	.773 (.024)	.790	.697	.735
		1,000	.697	.755 (.015)	.771	.694	.731
		5,000	.693	.751 (.006)	.765	.692	.730
	3PL + GPC (35)	200	.915	.926 (.005)	.932	.915	.922
		1,000	.913	.923 (.003)	.929	.913	.920
		5,000	.913	.923 (.001)	.929	.913	.920
Mathematics	3PL (25)	200	.884	.889 (.009)	.897	.884	.886
		1,000	.884	.889 (.005)	.899	.884	.886
		5,000	.884	.889 (.002)	.899	.884	.886
	GPC (4)	200	.826	.856 (.017)	.864	.820	.831
		1,000	.821	.852 (.008)	.860	.819	.830
		5,000	.820	.850 (.003)	.858	.818	.828
	3PL + GPC (29)	200	.897	.930 (.006)	.935	.896	.924
		1,000	.896	.929 (.003)	.934	.896	.924
		5,000	.896	.928 (.001)	.934	.896	.923
Science	3PL (30)	200	.867	.871 (.011)	.879	.865	.867
		1,000	.865	.869 (.005)	.876	.865	.867
		5,000	.866	.869 (.002)	.877	.866	.867
	GPC (5)	200	.787	.791 (.019)	.792	.776	.778
		1,000	.769	.772 (.010)	.774	.766	.767
		5,000	.769	.771 (.005)	.773	.767	.768
	3PL + GPC (35)	200	.899	.910 (.008)	.914	.896	.906
		1,000	.896	.906 (.003)	.910	.895	.904
		5,000	.896	.906 (.002)	.910	.896	.905

even when the sample size is 200. A comparison among sample sizes suggests that the SD of $\hat{\rho}_{XX'}$ is inversely related to the square root of sample size, with other conditions being equal.

It is interesting to see for all subject areas that for the 3PL test, the reliability coefficient is closer to the lower bound, whereas for the GPC and 3PL + GPC tests, it is closer to the upper bound. It is also interesting to find that IRT reliability lower bound estimates are, in most cases, close to Cronbach's alpha coefficients. A notable finding, which is verified over all testing conditions, is that the IRT reliability coefficient is higher than either of the two CTT reliability statistics. However, the IRT reliability coefficient is closer to the Feldt–Gilmer congeneric coefficient than to the Cronbach's alpha coefficient. Note for the English GPC test that the IRT reliability coefficient estimates (about .76 on average) appear to be substantially higher than the Cronbach's alpha coefficients (about .69). This result is not unexpected, because a congeneric relationship is a more reasonable assumption than tau-equivalence for items with varying IRFs. If all items had similar IRFs, there would be a small difference between the alpha and Feldt–Gilmer coefficients.

In summary, it appears that the alpha coefficient is slightly less than the Feldt–Gilmer coefficient for the all 3PL tests and the Science GPC test, but the former is relatively much less than the latter for the other (GPC and 3PL + GPC) tests. This result is reasonable because the tau-equivalence assumption is likely to be less violated in a test that consists of the same types of items (for which the number of response categories is the same and the IRT model applied is the same). In other words, the comparative results suggest that a congeneric coefficient is more preferable to coefficient alpha when variations in IRF are substantial across items, as in the case of the 3PL + GPC tests.

General discussion and conclusion

The results from these two empirical studies suggest the answers to two practical questions: (1) What kind of items (identified by their discrimination, difficulty, and/or guessing parameters) should a test include to enhance the reliability of test scores? (2) How much may the IRT reliability coefficient differ from the CTT-based reliability statistics (Cronbach's alpha and Feldt–Gilmer congeneric coefficient) when they are applied to the data of real tests? With regard to the first question, one may draw the following conclusion: to enhance the reliability of item and test scores, test developers should use MC and CR items that have high discrimination and have difficulty that matches the average ability of the target examinee group. For the MC items that are dichotomously scored, the influence of guessing should be minimized so that the item true-score variance and in turn the IRT item reliability coefficient may be enhanced. The fact that high discrimination (identified by IRT models) is associated with high item-score reliability is helpful in understanding the rationale of the definition of the item-reliability index (Gulliksen 1950). This index is defined as the item-test point biserial correlation multiplied by the item-score SD and is taken as a measure of item discrimination in CTT. Therefore, the item-reliability index can be reasonably interpreted as reflecting item reliability and the contribution of an item to test score reliability. The fact that matching IRT item location with ability location leads to high item-score reliability also supports the definition of the item-reliability index. When item difficulties are properly matched with the ability levels of examinees, the observed item scores would vary much among examinees. Thus, it seems reasonable to incorporate the item-score SD into the item-reliability index.

With regard to the second question, the answer depends on the degree to which the tau-equivalence or congeneric similarity assumption among items is violated in practice. In theory, tau-equivalence is a special case of congeneric similarity; with practical test data, the former is harder to strictly hold than the latter. Thus, the Feldt–Gilmer congeneric approach to estimating test-score reliability is the one better fitting the data in practice than the Cronbach's (essential) tau-equivalence approach. In particular, the congeneric approach is more reasonable than the tau-equivalence approach for use with a mixed-format test having heterogeneous types of items. By contrast, the IRT approach does not require even the congeneric similarity assumption, although the assumptions of model-data fit and local independence are met to a satisfactory degree when the approach is applied to test data. Of course, for the IRT approach, item parameters need to be successfully estimated and the underlying ability distribution should be properly specified or estimated under the IRT assumptions. Consequently, when the IRT assumptions and the CTT congeneric relationship are satisfactorily met across items, the IRT reliability coefficient should be close to the Feldt–Gilmer congeneric coefficient. Recall that such closeness was verified for the (single-format) 3PL tests in the second study. Otherwise, as suggested by the second study results, the "exact" IRT reliability coefficient could be slightly higher than the Feldt–Gilmer congeneric coefficient but much higher than the Cronbach's alpha coefficient. However, in either condition, the alpha coefficient should be close to the lower bound of IRT test reliability, as suggested by the second study results.

Despite the possible difference in performance between the IRT- and CTT-based reliability statistics, one might reasonably ask: what is the advantage of the IRT approach over the CTT approaches when the test-score metric is used as a basis for reporting and interpreting examinees' scores? Why not simply compute test-score reliability coefficients using either of the Cronbach's alpha and Feldt–Gilmer coefficients, since these CTT-based reliability statistics would not be very different from the IRT reliability coefficient in a practical sense? The foregoing answers to these questions can be defended on the basis of our empirical and theoretical research whenever the chosen IRT model fits the item and test score data being analyzed.

However, the CTT-based approaches are not preferable or have limitations for the following reasons. First, it would not be natural or plausible to adopt the CTT approaches to estimating test-score reliability when test development and item analysis are implemented in the framework of IRT. As long as the IRT parameters (or their estimates) used for computing reliability are available, it is reasonable to use the IRT-based approach. In this situation, the IRT approach is expected to result in a more accurate estimate of $\rho_{XX'}$ than the CTT approaches. Second, the CTT-based approaches do not provide facilities to project the test-score reliability for other examinee groups of interest beyond the group tested. That is, the CTT reliability statistics are both test- and group-dependent. In contrast, by the invariance property of IRT modeling (Lord 1980), the IRT test reliability coefficient can be estimated for populations beyond the present group. Of course, a reasonable assumption must be made about the ability distribution for the target group or an acceptable estimate obtained of this distribution, and the same item parameters must be assumed to apply. In this sense, the IRT reliability coefficient may be said to be test-dependent only. Third, the CTT-based approaches are feasible only for estimating test-level reliability (i.e., they cannot be used for a one-item test), but the IRT-based approach is applicable for estimating both item- and test-level reliability. Thus, one may better understand the contribution of each item to test-score reliability by using the IRT approach than by using the CTT approaches.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

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

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27*, 440–458.

Feldt, L. S. (2002). Estimating the internal consistency reliability of tests composed of testlets varying in length. *Applied Measurement in Education, 15*, 33–48.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika, 48*, 99–111.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15*, 375–389.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*, 243–271.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68*, 123–149.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*, 129–140.

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

May, K., & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Journal of Educational Measurement, 31*, 313–325.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika, 30*, 419–440.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177–195.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software International.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data*. Lincolnwood, IL: Scientific Software International.

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). New York, NY: Cambridge University Press.

Shojima, K., & Toyoda, H. (2002). Estimation of Cronbach's alpha coefficient in the context of item response theory. *The Japanese Journal of Psychology, 73*, 227–233. (In Japanese).

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39–49.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51*, 251–267.

Woodruff, D. J., & Hanson, B. A. (1996). *Estimation of item response models using the EM algorithm for finite mixtures*. Iowa City, IA: ACT, Inc. (ACT Research Report 96–6).