Applied Psychological Measurement

http://apm.sagepub.com

Using Classical Test Theory in Combination with Item Response Theory

Timo M. Bechger, Gunter Maris, Huub H. F. M. Verstralen and Anton A. Béguin

Applied Psychological Measurement 2003; 27; 319
DOI: 10.1177/0146621603257518

Using Classical Test Theory in Combination With Item Response Theory

Timo M. Bechger, Gunter Maris, Huub H. F. M. Verstralen, and Anton A. Béguin, Cito, Arnhem, The Netherlands

This study is about relations between classical test theory (CTT) and item response theory (IRT). It is shown that CTT is based on the assumption that measures are exchangeable, whereas IRT is based on conditional independence. Thus, IRT is presented as an extension of CTT, and concepts from both theories are related to one another. Furthermore, it is demonstrated that IRT can be used to provide CTT statistics in situations where CTT fails. Reliability, for instance, can be determined even though a test was not administered to the intended population.

Index terms: classical test theory, item response theory, item reliability, true test score, item total correlation, item true score, item response function.

Introduction

Notwithstanding the many developments in item response theory (IRT), classical test theory (CTT) continues to be an important framework for test construction. It is therefore useful to have a clear notion of the relations between IRT and CTT. This should improve the appreciation of both theories and facilitate communication with researchers and item writers who are frequently more familiar with CTT than with IRT. In this article, the relations between CTT and IRT are summarized, and novel applications of CTT that are feasible using IRT are discussed.

This article consists of a theoretical and a practical part. The second section provides a brief outline of CTT and its relation to IRT. In the third section, the CTT concept of reliability is applied in an IRT context, including reliability of estimated latent trait values and reliability of classifications using a test score. In the fourth section, four applications are discussed: (a) reliability estimation from a single administration of a test; (b) relations between test characteristics, the population of test takers, and test scores; (c) the correlation between latent traits measured by different tests; and (d) the selection of items from a pilot test when the pilot test could not be administered to the intended population. The article is concluded in the fifth section.

This article is written in the spirit of work by Verstralen (1997b), Lord (1983), Nicewander (1993), Thissen (1990), Mellenbergh (1994, 1996), and Steyer and Eid (1993), and there is some overlap between these studies and the present article. Naturally, Lord and Novick (1968) are frequently referred to as well.

Classical Test Theory From an IRT Point of View

Classical Test Theory

Let an "item" be a means to produce a measurement X. It is assumed that the respondent's behavior is determined by the value on a vector variable θ, which represents what the item intends

Applied Psychological Measurement, Vol. 27 No. 5, September 2003, 319–334
DOI: 10.1177/0146621603257518

to measure. This variable may be continuous or discrete. The measurement X is defined as a discrete random variable that represents the credit assigned to each response. The function that defines X is called *the scoring rule*. Realizations of X are called "responses" in IRT and "scores" in CTT.

The *true score* of any person is defined as the expectation $E[X|\theta]$ of the distribution of X over subjects with the same ability. The deviations $X - E[X|\theta]$ represent random measurement error, that is, uncontrolled environmental variables that influence the response (Lord & Novick, 1968, pp. 38-39). The distribution of the measurement errors has zero mean and variance $Var(X|\theta)$. Although the measurement error varies across persons with the same θ, the true score is a fixed parameter characterizing the combination of a θ and an item.

Taking the expectation of $E[X|\theta]$ over the distribution of θ in the population of interest gives the expected response to item i. The *reliability* of X in the reference population, ρ_X^2, is defined as the proportion of true variation. Specifically, provided that $Var(X) > 0$,

$$\begin{aligned}\rho_X^2 &\equiv \frac{Var(E[X|\theta])}{Var(X)} \\ &= 1 - \frac{E[Var(X|\theta)]}{Var(E[X|\theta]) + E[Var(X|\theta)]},\end{aligned} \tag{1}$$

where $Var(E[X|\theta])$ denotes the true score variance, and

$$E[Var(X|\theta)] \equiv E(E[(X - E[X|\theta])^2|\theta]) = E[(X - E[X|\theta])^2],$$

is the measurement error variance in the population. It is customary to denote the reliability as a square because ρ_X^2 equals the square of the correlation between the true score and the observed score (Lord & Novick, 1968, p. 57). For this reason, reliability is sometimes denoted by ρ_{XT}^2. A correlation is not invariant under nonlinear transformations, and reliability depends on the scoring rule. That is, some scoring rules give higher reliability than others. Equation (1) also shows that item reliability depends on the ability distribution in the population.

Lord and Novick (1968) consider the following experiment, albeit in different wording: Draw a θ from the population and generate two independent responses x and x^* to the same item. The joint distribution of these responses is

$$\Pr(X, X^*) = \int \Pr(X = x|\theta) \Pr(X^* = x^*|\theta) g(\theta) d\theta, \tag{2}$$

where

$$\Pr(X = x|\theta) = \Pr(X^* = x|\theta).$$

Equation (2) states that the response variables are exchangeable, and henceforth they will be called *exchangeable replications* to indicate that they are independent conditional on θ, but not marginally. Item reliability equals the correlation between exchangeable replications and is sometimes denoted as ρ_{XX^*}. This can be seen using the *covariance decomposition formula*:

$$Cov(X, X^*) = Cov(E[X|\theta], E[X^*|\theta]) + E[Cov(X, X^*|\theta)], \tag{3}$$

where $E[Cov(X, X^*|\theta)] = 0$, and $Cov(E[X|\theta], E[X^*|\theta]) = Var(E[X|\theta])$ by assumption. Dividing $Cov(X, X^*)$ by $\sqrt{Var(X)Var(X^*)} = Var(X)$ gives (1).

Now, consider a test consisting of $I > 1$ items.[1] It is customary to consider a linear combination $Y \equiv \sum_{i=1}^{I} w_i X_i$ of the item responses as a test score, where the w_i are constant weights. The true

[1] The distinction between an item and a test is convenient but unnecessary for classical test theory (CTT).

test score is given by $E[Y|\theta] = \sum_{i=1}^{I} w_i E[X_i|\theta]$; in IRT, this function of θ is known as the *test characteristic curve*. The reliability of the test score in the reference population is given by

$$\rho_Y^2 = \frac{Var(E[Y|\theta])}{Var(E[Y|\theta]) + E[Var(Y|\theta)]}. \tag{4}$$

It follows from exchangeability that the measurement errors on different items are independent given θ, and the error variance of the test score is given by $E[Var(Y|\theta)] = E[Var(X|\theta)] \sum_{i=1}^{I} w_i^2$. Test reliability is of interest because its square root, called “the index of reliability,” provides an upper bound to the validity of the test score with respect to any criterion, that is, the correlation of the test score with any criterion (Lord & Novick, 1968, p. 72).

Another important statistic in CTT is the *item total correlation* (ITC)—the correlation of the score on item i with the score on the test, including the item. By definition, the ITC is equal to

$$ITC_i = \frac{Cov(E[Y|\theta], E[X_i|\theta]) + E[Cov(Y, X_i|\theta)]}{\sqrt{Var(Y)Var(X_i)}}, \tag{5}$$

where the numerator follows from the covariance decomposition formula (3). In CTT, this correlation is interpreted as an item discrimination index because it indicates to what extent the item differentiates between persons with high scores on the test and persons with low scores on the test.

Because the total score on the proposed test is calculated with the score on item i, the ITC is spuriously high. To correct the ITC, it is customary to calculate the *item rest correlation* (IRC), which is the correlation between the score on an item and the total score on the proposed test, excluding the item. Specifically, the IRC of item i equals the corresponding ITC with w_i fixed to zero. The following proposition suggests that the IRC may be interpreted as an approximation to the square root of the item reliability.

Proposition 1 *Assume exchangeability. Then,*

$$\lim_{I\to\infty} IRC_i = \sqrt{\rho_{X_i}^2},$$

where I denotes the number of items in the test.
Proof. *Let Y_{-i} denote the rest score. By definition,*

$$\begin{aligned}
IRC_i &= \frac{Cov(Y_{-i}, X_i)}{\sqrt{Var(Y_{-i})}\sqrt{Var(X_i)}} \\
&= \frac{Cov(E[Y_{-i}|\theta], E[X_i|\theta])}{\sqrt{Var(E[Y_{-i}|\theta])}\sqrt{Var(X_i)}}\sqrt{\frac{Var(E[Y_{-i}|\theta])}{Var(Y_{-i})}} \\
&= \sqrt{\frac{Var(E[X_i|\theta])}{Var(X_i)}}\frac{Cov(E[Y_{-i}|\theta], E[X_i|\theta])}{\sqrt{Var(E[Y_{-i}|\theta])}\sqrt{Var(E[X_i|\theta])}}\sqrt{\rho_{Y_{-i}}^2} \\
&= Corr\left(E\left[Y_{-i}|\theta\right], E[X_i|\theta]\right)\sqrt{\rho_{X_i}^2\rho_{Y_{-i}}^2}.
\end{aligned}$$

Under exchangeability, $Corr\left(E\left[Y_{-i}|\theta\right], E[X_i|\theta]\right) = Corr((I-1)E[X_i|\theta], E[X_i|\theta]) = 1$. It follows that

$$\begin{aligned}
\lim_{I\to\infty}\rho_{Y_{-i}}^2 &= \lim_{I\to\infty}\frac{(I-1)^2 Var(E[X|\theta])}{(I-1)^2 Var(E[X|\theta]) + E[Var(X|\theta)](I-1)} \\
&= \lim_{I\to\infty}\frac{Var(E[X|\theta])}{Var(E[X|\theta]) + E[Var(X|\theta)](I-1)^{-1}} \\
&= 1
\end{aligned}$$

so that $\lim_{I\to\infty} IRC_i = \sqrt{\rho^2_{X_i}}$*. The same holds true for the ITC, which becomes equal to the IRC when the number of items increases.* ■

It is seen that the IRC is positive and dependent on the relation of the true rest score and the item true score.

Item Response Theory and Classical Test Theory

In practice, it is assumed that the responses to different items are exchangeable so that item reliability can be estimated by their correlations. In CTT, such measures are called "equivalent" (Lord & Novick, 1968). This assumption is unrealistic, especially because different items will not frequently have the same conditional distribution. It is, therefore, opportune to relax the assumption of exchangeability and require that responses to different items be independent conditional on θ but not necessarily identically distributed. In IRT, this is called *conditional independence* (CI). For two items, CI is equivalent to

$$\Pr(X_i, X_j) = \int \Pr(X_i = x_i|\theta)\Pr(X_j = x_j|\theta)g(\theta)d\theta, \tag{6}$$

where $\Pr(X_i = x_i|\theta)$, which will be denoted by $P_{ix_i}(\theta)$, is called the *item response function* (IRF). Suppes and Zanotti (1981) show that there always (i.e., for every joint distribution) exists a scalar-valued θ such that CI holds. This means that CI by itself is not a restriction on the data, and additional assumptions are needed on the IRFs. Together with CI, these additional restrictions define an IRT model.

Here, it is assumed that θ is scalar valued, and the item true score, $E[X_i|\theta] = \sum_{x_i} x_i P_{ix_i}(\theta)$, is a monotone increasing function of θ so that the true score is a one-to-one transformation of θ. Together with CI, these assumptions define the family of unidimensional monotone IRT models that encompasses most existing IRT models used for ability measurement.[2]

The Case of Binary, Equivalent Rasch Items

In this section, the items are assumed exchangeable measures, and an IRT model is introduced that is formally equivalent to CTT. All items are binary with $X_i = 1$ if the answer is correct and $X_i = 0$ otherwise. Subscript i will be deleted because all items are equivalent.

Without loss of generality, the IRFs are modeled by the Rasch model (Rasch, 1960); that is,

$$P_1(\theta) = \frac{\exp(\theta - \delta)}{1 + \exp(\theta - \delta)}, \tag{7}$$

where the parameter $\delta \in \mathbb{R}$ is considered known, and θ is a scalar ability. The population distribution is unrestricted. The assumption that $P_1(\theta)$ is modeled by the Rasch model implies no loss in generality because θ can always be transformed such that the IRFs assume any other functional form. The difficulty parameter δ is the value of θ, where $P_1(\theta) = 1 - P_1(\theta) = 0.5$.

With binary items, the item true score equals the probability of a correct response, given θ. An illustration is given in Figure 1. The conditional measurement error variance of the score for each item is equal to $P_1(\theta)(1 - P_1(\theta))$. Using the formulas in the previous section, the following is found:

$$Var(X) = E[P_1(\theta)](1 - E[P_1(\theta)]), \tag{8}$$

[2] Note that given conditional independence (CI), $E[Cov(Y, X_i|\theta)] = w_i E[Var(X_i|\theta)]$, and (5) reduces to a more manageable expression.

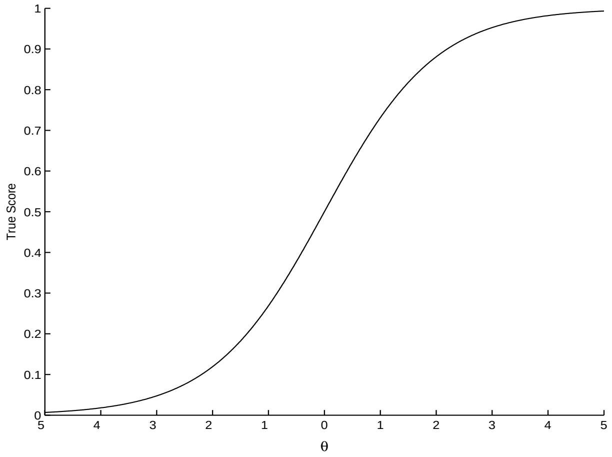
where $E[P_1(\theta)]$—that is, the expected percentage correct—is known as the difficulty of the item. Because the term *item difficulty* has a different meaning in IRT, the term *expected percentage correct* will be used to denote $E[P_1(\theta)]$. The true score variance for any item is given by

$$Var(E[X|\theta]) = E[(P_1(\theta))^2] - E[P_1(\theta)]^2, \tag{9}$$

which equals $Var(P_1(\theta))$, the variance of the proportion correct in the reference population. Note that $E[(P_1(\theta))^2] = \Pr(X_i = 1, X_j = 1)$, and $Var(E[X|\theta]) = \Pr(X_i = 1, X_j = 1) - E[P_1(\theta)]^2$, when i and j index two equivalent binary items. The item reliability follows from substitution of (8) and (9) in (1). Under the present assumptions, item reliability equals Loevinger's (1948) H-coefficient, which is used in Mokken scale analysis (Mokken, 1971, p. 150).

Figure 1

The Item Response Function (IRF) for a Rasch Item With $\delta = 1$



Test reliability can be calculated using the well-known Spearman-Brown formula or Cronbach's alpha. Both are easy to derive under the present assumptions (see Bechger, Maris, Béguin, & Verstralen, 2003).

A plot of the ITC against the expected percentage correct of any of the items in Figure 2 shows that the relation is quadratic. This reveals that, in the given circumstances, the ITC is not a well-defined measure of "item discrimination power" because it depends on the item difficulty, on the expectation and the dispersion of θ, and on the number of items in the test (see also Steyer & Eid, 1993, pp. 137-138). This is also true under more general circumstances when the items are not equivalent. One should therefore be careful to give general rules of thumb for the selection of items based on the ITC (e.g., Ebel & Frisbie, 1986).

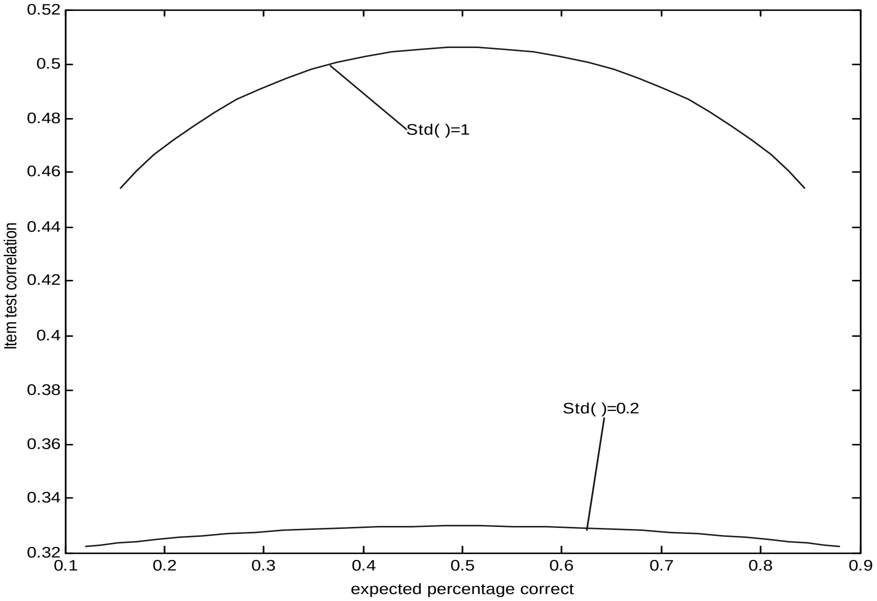
Reliability in Item Response Theory

Reliability of Estimated Abilities

The correlation between Y and $E[Y|\theta]$ is not equal to the correlation between Y and θ unless the latter is a linear transformation of $E[Y|\theta]$ as in the binomial model (Rost, 1996, pp. 113-119). In most applications, the relation between Y and θ is postulated to be nonlinear, however. When estimates of θ are reported and used, it is therefore appropriate to provide the reliability of the estimated ability values $\hat{\theta}$.

Figure 2

Relation Between the Expected Percentage Correct and the Item Test Correlation Assuming 10 Equivalent Rasch Items



Note. The distribution of θ is assumed standard normal.

To derive this reliability, note that

$$\hat{\theta} = E[\hat{\theta}|\theta] + e, \tag{10}$$

where $e \equiv \hat{\theta} - E[\hat{\theta}|\theta]$ can be interpreted as measurement error, and $E[\hat{\theta}|\theta]$ can be interpreted as a true score. Reliability is defined as the proportion of true variance in the reference population, and hence

$$\begin{aligned} \rho^2_{\hat{\theta}} &= \frac{Var(E[\hat{\theta}|\theta])}{Var(\hat{\theta})} \\ &= 1 - \frac{E[Var(\hat{\theta}|\theta)]}{Var(E[\hat{\theta}|\theta]) + E[Var(\hat{\theta}|\theta)]}, \end{aligned} \tag{11}$$

where $Var(\hat{\theta}|\theta)$ denotes the variance of the estimated values, given θ. It follows from the previous discussion that $\rho^2_{\hat{\theta}}$ may be interpreted as a measure of linear association between exchangeable replicates of $\hat{\theta}$. This means that $\rho^2_{\hat{\theta}}$ changes if $\hat{\theta}$ is nonlinearly transformed, and its value depends on the parameterization of the IRT model.

If $\hat{\theta}$ is an unbiased estimator, $Var(E[\hat{\theta}|\theta]) = Var(\theta)$, and $\rho^2_{\hat{\theta}}$ is equal to the square of correlation between $\hat{\theta}$ and θ, which was proposed by Gustafsson (1977) as a measure of "subject separability."

This is also true when $E[\hat{\theta}|\theta] = \alpha_1\theta + \alpha_2$, $(\alpha_1, \alpha_2 \in \mathbb{R})$ because $\hat{\theta}$ is then a linear function of an estimator that is unbiased, and the correlation between exchangeable replicates is invariant under linear transformations. In general, the correlation between θ and $\hat{\theta}$ is equal to

$$\begin{aligned} Corr(\theta, \hat{\theta}) &= \frac{Cov(\theta, \theta + Bias(\theta))}{\sqrt{Var(\hat{\theta})Var(\theta)}} \\ &= \sqrt{\frac{Var(\tilde{\theta})}{Var(\hat{\theta})}}\sqrt{\frac{Var(\theta)}{Var(\tilde{\theta})}} + Corr(Bias(\theta), \theta)\sqrt{\frac{Var(Bias(\theta))}{Var(\hat{\theta})}}, \end{aligned} \tag{12}$$

where $\tilde{\theta}$ denotes an unbiased estimator, and $\hat{\theta}$ is a biased estimator. The ratio $Var(\theta)/Var(\tilde{\theta})$ is the reliability of an unbiased estimator.

There are at least two ways to calculate the reliabilities: The first procedure requires that the estimated θ is a one-to-one function of the test score. That is, Y is minimally sufficient for θ, as in the Rasch model. The IRT model gives the distribution of the test score Y, given θ; $\Pr(Y = y|\theta)$, where y are the values taken by Y. Each value y gives an estimated ability $\hat{\theta}(y)$ and $\Pr(Y = y|\theta) = \Pr(\hat{\theta} = \hat{\theta}(y)|\theta)$—the distribution of the estimated abilities, given θ. The variance of $\hat{\theta}$, given θ, may now be calculated as

$$\begin{aligned} Var(\hat{\theta}|\theta) &= E[\hat{\theta}^2|\theta] - E[\hat{\theta}|\theta]^2 \\ &= \sum_y \hat{\theta}^2(y)\Pr(\hat{\theta} = \hat{\theta}(y)|\theta) - \left(\sum_y \hat{\theta}(y)\Pr(\hat{\theta} = \hat{\theta}(y)|\theta)\right)^2, \end{aligned} \tag{13}$$

and

$$Var(E[\hat{\theta}|\theta]) = E[E[\hat{\theta}|\theta]^2] - E[E[\hat{\theta}|\theta]]^2. \tag{14}$$

This means that the second expression in Equation (11) can be used to compute the reliability.

The second method uses the fact that when the parameters are estimated by the method of marginal maximum likelihood, the variance of θ is typically a parameter that is estimated, whereas the variance of the estimated θs can be computed directly. This implies that the first expression in Equation (11) can be used to compute the reliability.

Thissen (1990; Mellenbergh, 1994, Equation (22); Samejima, 1994, Equation (21)) gives an approximation to the reliability (see Equation (15)). It will now be shown that this approximation gives an upper limit to the reliability. First, it is well known that the ML estimator has a limiting normal distribution with expectation θ and variance equal to the inverse of the information $I(\theta)$:

$$\sqrt{I}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta)),$$

where

$$I(\theta) = \frac{1}{I}\sum_i -E\left[\frac{\partial^2 \ln(P_i(\theta))}{\partial\theta^2}\right].$$

Using this result, the reliability may be expressed as follows:

$$\rho^2_{\hat{\theta}} = \frac{Var(\theta)}{Var(\theta) + Var(E[\hat{\theta}|\theta])}.$$

Second, it follows from Jensen's inequality that $E[I^{-1}(\theta)] \geq E[I(\theta)]^{-1}$. Hence, if bias is ignored (i.e., $Var(E[\hat{\theta}|\theta]) = Var(\theta)$), the following upper limit for the reliability is obtained:

$$\rho^2_{\hat{\theta}} \leq \frac{E[I(\theta)]Var(\theta)}{1 + E[I(\theta)]Var(\theta)}. \tag{15}$$

All the required integrals can be calculated using numerical integration, if necessary. Alternative approximations to $\rho^2_{\hat{\theta}}$ are discussed by Verhelst, Glas, and Verstralen (1995, p. 64) and Rost (1996, pp. 353-354).

The Reliability of Classifications

Suppose that a test score is used to classify examinees in two mutually exclusive categories on the basis of a predetermined observed score cut point c, preferably derived using some sort of standard-setting scheme. The observed cut point may also be a score corresponding to a latent cut point. Thus, persons with test scores less than c will fail the test, and persons with a score equal to c or over c will pass. Now, let I_p denote whether students pass. Then, assuming CI, the conditional probability of passing is equal to

$$\Pr(I_p = 1|\theta) = \sum_{y=c}^{\max(Y)} \Pr(Y = y|\theta) \tag{16}$$

$$= \sum_{y=c}^{\max(Y)} \left[\sum_{\mathbf{x}:\sum_i w_i x_i = y} \prod_i \Pr(X_i = x_i|\theta) \right]. \tag{17}$$

It is seen from (17) that $\Pr(Y = y|\theta)$ is an elementary symmetric function. It may be calculated recursively, as discussed by Lord and Wingersky (1984); Thissen, Pommerich, Billeaud, and Williams (1995); and Bechger et al. (2003, appendix). The marginal probability of passing equals $\Pr(I_p = 1) = E\left[\Pr(I_p = 1|\theta)\right]$.

Equation (1) provides a definition for the reliability of the classification; that is,

$$\rho^2_{Clas} = \frac{Var(E[I_p|\theta])}{Var(E[I_p|\theta]) + E[Var(I_p|\theta)]} \tag{18}$$

$$= \frac{E[\Pr(I_p = 1|\theta)^2] - E[\Pr(I_p = 1|\theta)]^2}{E[\Pr(I_p = 1|\theta)] - E[\Pr(I_p = 1|\theta)]^2}.$$

Because ρ^2_{Clas} is a reliability, it equals the correlation between classifications across two exchangeable administrations of the test. It can also be shown that classification reliability equals Cohen's kappa (Cohen, 1960) when it is computed using two exchangeable administrations of the same test. For later reference, this is stated as a proposition:

Proposition 2 *Assuming exchangeability, classification reliability equals Cohen's kappa (Cohen, 1960).*

Proof. *Let $I_p^{(r)}$ denote passing on the* r*th administration. Cohen's kappa is equal to*

$$\kappa = \frac{P_o - P_c}{1 - P_c}, \tag{19}$$

where $P_o = E[P_o(\theta)]$ *denotes the observed agreement, and* $P_c = \Pr(I_p^{(1)} = 1)\Pr(I_p^{(2)} = 1) + \Pr(I_p^{(1)} = 0)\Pr(I_p^{(2)} = 0)$ *denotes the agreement observed by chance. Under exchangeability,*

$$\begin{aligned}
P_o &= E[P_o(\theta)] \\
&= E[\Pr(I_p^{(1)} = 1, I_p^{(2)} = 1|\theta)] + E[\Pr(I_p^{(1)} = 0, I_p^{(2)} = 0|\theta)] \\
&= E[\Pr(I_p^{(1)} = 1|\theta)\Pr(I_p^{(2)} = 1|\theta)] + E[\Pr(I_p^{(1)} = 0|\theta)\Pr(I_p^{(2)} = 0|\theta)] \\
&= E[\Pr(I_p = 1|\theta)^2] + E[(1 - \Pr(I_p = 1|\theta))^2].
\end{aligned}$$

The last equality follows because $\Pr(I_p^{(1)} = 1|\theta) = \Pr(I_p^{(2)} = 1|\theta)$*, by assumption. In the same way,*

$$P_c = \left(E\left[\Pr(I_p = 1|\theta)\right]\right)^2 + \left(1 - E\left[\Pr(I_p = 1|\theta)\right]\right)^2.$$

If P_o *and* P_c *are expanded and substituted in Equation (19), then*

$$\begin{aligned}
\kappa &= \frac{2Var(\Pr(I_p = 1|\theta))}{-2E[\Pr(I_p = 1|\theta)]^2 + 2E[\Pr(I_p = 1|\theta)]} \\
&= \frac{Var(\Pr(I_p = 1|\theta))}{E[\Pr(I_p = 1|\theta)] - E[\Pr(I_p = 1|\theta)]^2} \\
&= \frac{Var(\Pr(I_p = 1|\theta))}{Var(I_p)} = \frac{Var(E[I_p|\theta])}{Var(I_p)}.
\end{aligned}$$

This ends the proof. ■

As seen in Proposition 2, exchangeability implies that kappa cannot be negative. If it is found to be negative, this is a sign that exchangeability is violated. Note that the weighted kappa coefficient (Cohen, 1968) may serve as a general index for the reliability of classification when there are more than two categories (see Bechger et al., 2003).

Imagine two exchangeable administrations of the same examination. The probability of consistent classification, given θ, equals

$$\begin{aligned}
P_o(\theta) &= \Pr(I_p^{(1)} = 1, I_p^{(2)} = 1|\theta)] + E[\Pr(I_p^{(1)} = 0, I_p^{(2)} = 0|\theta) && (20) \\
&= \left[\sum_{y=c}^{\max(Y)} \Pr(Y = y|\theta)\right]^2 + \left[\sum_{y=0}^{c-1} \Pr(Y = y|\theta)\right]^2. && (21)
\end{aligned}$$

This function is called the *test characteristic decision curve* (TCDC). The probability of inconsistent classification is $1 - P_o(\theta)$. When the TCDC is integrated over the reference population, the probability of consistent classification when the test is applied to the reference population using $y = c$ as a cutoff is obtained. This quantity may prove to be useful in view of the current trend to demand that testing organizations publish procedures and provide formal justification for the quality of their examinations.

Alternative ways to quantify and investigate the quality of classifications are discussed by Livingston and Lewis (1995), Verstralen (1997a), Sluijter (1998), and Spray and Reckase (1994). Lee, Hanson, and Brennan (2002) also consider Cohen's kappa as an index for the quality of classification.

Applications

Calculating Reliability With a Single Administration of a Test

The easiest application is to use the formulas that were given earlier to calculate reliability using a single test administration.[3] To illustrate this possibility, the "KFT data" that are listed on pages 99 and 100 in the book by Jürgen Rost (1996) are used.[4] The data consist of responses to five items by 300 students. The items were found to conform to a theory-based restriction of the Rasch model called the linear logistic test model (Fischer, 1995). A report of the IRT analysis can be found in Rost (1996, p. 248) or Bechger, Verstralen, and Verhelst (2002, section 6). This illustrates that an IRT analysis may provide information about the items that would not be available if one is confined to classical item analysis. Marginal maximum likelihood estimation was used to obtain estimates of population parameters; the population distribution was assumed to be normal, and the item parameters were restricted to sum to zero to achieve identification of the model.

The population mean was estimated to be -0.158 and the standard deviation 1.950. The trapezoidal rule (Davis & Rabinowitz, 1984, chap. 2, section 3.4) was used to approximate the expectations and calculate the values in the following table.

	$\rho^2_{X_i}$	$E[X_i]$	IRC_i
Item 1	0.37	0.63	0.59
Item 2	0.38	0.56	0.60
Item 3	0.38	0.49	0.61
Item 4	0.38	0.42	0.60
Item 5	0.36	0.28	0.56
	$\rho^2_{\hat{\theta}} = 0.74$	$\rho^2_Y = 0.75$	

The items are nearly equivalent so that Cronbach's alpha is, in this case, found to be only slightly lower than the estimated test reliability. The reliability of the unweighted test score almost equals that of the estimated θs. This means that the relation between the unweighted scores and the estimated θs is approximately linear.

Graphical Methods to Investigate Relations Between Test Characteristics, the Population of Test Takers, and Test Scores

Plots are often useful to illustrate relations between test characteristics as determined by an IRT model, the population of test takers, and test scores, especially when such relations cannot easily be described analytically or communicated to test developers. For example, Lord (1953; Lord & Novick, 1968, Figs. 16.14.1-16.14.6) uses plots of the relation between θ and the true score to illustrate how the distribution of the true score depends on the discrimination power of the test. Using numerical integration to calculate expectations, if necessary, the formulas presented here may be used to produce such plots.

In a previous section, this approach was illustrated when the relation between the ITC and the item difficulty was investigated (see Figure 2). Two further illustrations are presented here of the usefulness of the plots.

A first illustration is provided in Figure 3, which shows the relation between the expected information of a test of two items, as well as the difference between the difficulty parameters of the items. Although the mean percentage correct is always a half, it is seen that the expected test

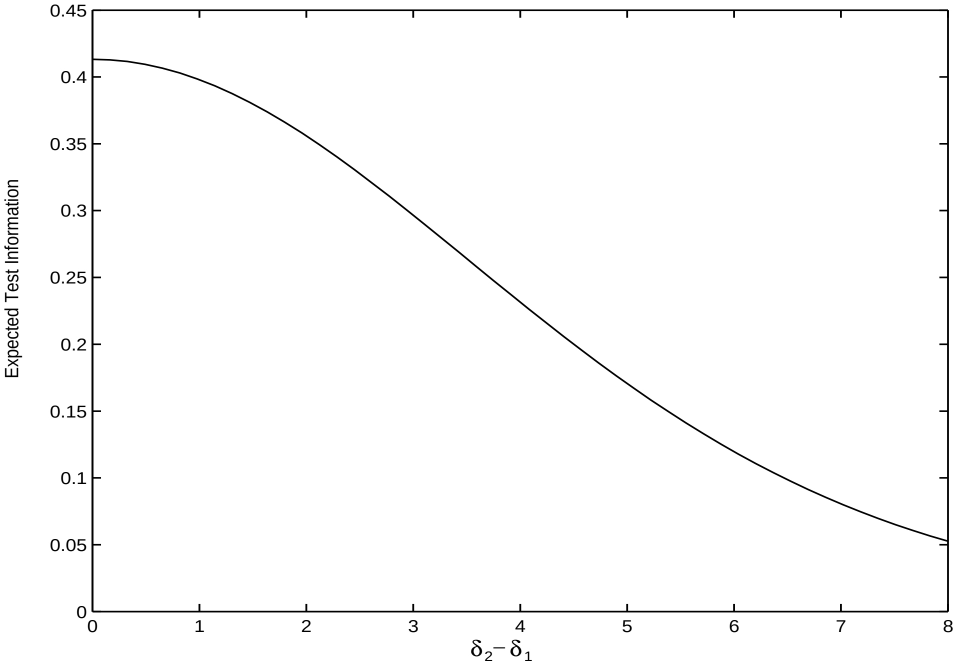
[3] Some of this could be done with the OPTAL program (Verstralen, 1997b), which is part of the OPLM software.

[4] The complete data set with 15 variables comes with the (excellent) WINMIRA software (von Davier, 1994). The present items are the first five items.

information diminishes as one item becomes more difficult and the other more easy (see also Muraki, 1993). Moreover, in the limiting case, every subject solves the easy item and fails the difficult one. That is to say, all true score variance vanishes, which leaves the reliability undefined. From this example, it may be concluded that focusing exclusively on expected percentage correct may lead test developers to a test that fails to distinguish between persons.

Figure 3

Expected Information for a Test With Two Rasch Items,
With Mean Difficulty Equal to Zéro, Plotted Against
the Difference Between the Difficulty Parameters



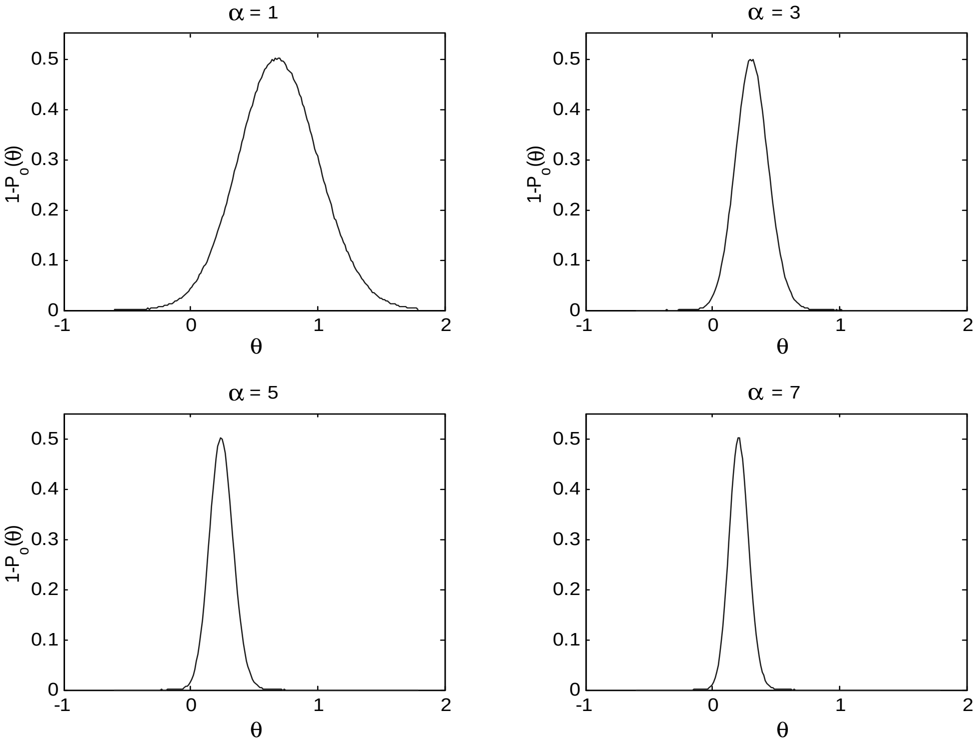
The second illustration is provided in Figure 4, which shows the effect of the item discrimination on the TCDC. In IRT, the item discrimination corresponds to the slope of the IRF at the point $\theta = \delta_i$. The Rasch model can be extended to include an item discrimination parameter α_i, and this model is called the *two-parameter logistic* (2PL) model (Birnbaum, 1968). It is seen that the TCDC becomes more concentrated when the discrimination parameter increases. This illustrates that the quality of a decision increases when items discriminate better, especially when items are located close to the cutoff. This implies that test constructors are advised to select highly discriminating items close to the cutoff.

Calculating the Correlation Between Latent Traits Measured by Different Tests

Suppose one test measures a latent trait θ, and another test measures a latent trait ξ. Let $Corr(\hat{\theta}, \hat{\xi})$ denote the correlation between the estimates of θ and ξ. The following theorem relates $Corr(\hat{\theta}, \hat{\xi})$ to the correlation between θ and ξ.

Figure 4

Plots Illustrating the Effect of the Discrimination Parameters on the Test Characteristic Decision Curve (TCDC) Under the Two-Parameter Logistic (2PL) Model



THEOREM 3 *If both estimates are unbiased, and* $Cov(\hat{\theta}, \hat{\xi}|\theta, \xi) = 0$,

$$Corr(\theta, \xi) = Corr(\hat{\theta}, \hat{\xi})/\sqrt{\rho^2_{\hat{\theta}}\rho^2_{\hat{\xi}}},$$

where $Corr(\theta, \xi)$ *denotes the correlation between* θ *and* ξ.

Proof. *First, the covariance decomposition formula implies that*

$$Cov(\hat{\theta}, \hat{\xi}) = Cov\left(E\left[\hat{\theta}|\theta, \xi\right], E\left[\hat{\xi}|\theta, \xi\right]\right) + E\left[Cov(\hat{\theta}, \hat{\xi}|\theta, \xi)\right].$$

It follows that $Cov(\hat{\theta}, \hat{\xi}) = Cov\,(\theta, \xi)$ *if both estimators are unbiased, and* $Cov(\hat{\theta}, \hat{\xi}|\theta, \xi) = 0$. *Hence,*

$$\begin{aligned}\frac{Cov\,(\theta, \xi)}{\sqrt{Var(\theta)Var(\xi)}} &= \frac{Cov(\hat{\theta}, \hat{\xi})}{\sqrt{Var(\theta)Var(\xi)}} \\ &= \frac{Cov(\hat{\theta}, \hat{\xi})}{\sqrt{Var(\hat{\theta})Var(\hat{\xi})}}\frac{\sqrt{Var(\hat{\theta})Var(\hat{\xi})}}{\sqrt{Var(\theta)Var(\xi)}}\end{aligned}$$

$$= \frac{Cov(\hat{\theta}, \hat{\xi})}{\sqrt{Var(\hat{\theta})Var(\hat{\xi})}}\sqrt{\frac{Var(\hat{\theta})}{Var(\theta)}\frac{Var(\hat{\xi})}{Var(\xi)}}$$

$$= \frac{Cov(\hat{\theta}, \hat{\xi})}{\sqrt{Var(\hat{\theta})Var(\hat{\xi})}}\frac{1}{\sqrt{\rho^2_{\hat{\theta}}\rho^2_{\hat{\xi}}}}.$$

This ends the proof. ■

Theorem (3) shows that $Corr(\hat{\theta}, \hat{\xi})$ may be much lower than $Corr(\theta, \xi)$ due to unreliability in the estimates. It has been shown how the reliability can be computed, and $Corr(\hat{\theta}, \hat{\xi})$ is easily estimated from the data.

Note that the assumption that $Cov(\hat{\theta}, \hat{\xi}|\theta, \xi) = 0$ is violated when responses to items in one test are dependent on the responses to items in the other test, conditional on θ and ξ. When $Cov(\hat{\theta}, \hat{\xi}|\theta, \xi) > 0$, the correlation between the latent traits is overestimated, but if $Cov(\hat{\theta}, \hat{\xi}|\theta, \xi) < 0$, the correlation between the latent traits is underestimated.

Selecting Items From a Pilot Test

This application is discussed in the context of a real example. The state examination of Dutch as a second language is a large-scale examination of the ability to use the Dutch language in practical situations. There are separate examinations for listening, speaking, writing, and reading. An IRT model is used to scale the data and equate an examination to a reference examination to ensure that the ability required to pass the examination stays the same over years. Estimated achievement is transformed to a convenient scale and serves as examination marks.

The construction of a new examination is preceded by a pilot study that entails the administration of new items to a sample of immigrants who participate in a language course. The purpose of the pilot study is to select the items for the coming examinations. After the data have been collected, they are added to a large incomplete data set that contains the data obtained from previous pilot studies and examinations. This data set is called the *data bank*. The *reference examination* is a subset of the items in the data bank. This reference test was chosen by the examination committee and used to set the cutoff. The *reference population* is the population of examinees who are generally more able than the persons who participate in the pilot study.

The analysis of the pilot data consists of three stages. The IRT model used is the generalized partial credit model (GPCM) (Muraki, 1992). The GPCM is fit to the data using all relevant parts of the data bank. Items are discarded that do not conform to the model. In the second stage, the test developers are given two additional pieces of information. First, the expected proportions correct in the reference population are provided. The examination committee desires the expected proportions correct between 0.50 and 0.70. Second, IRC_is are provided using the score on the reference examination as a rest score. These IRCs may be interpreted as a measure of the fit of an item to the reference examination. With this information, and under strict surveillance by the examination committee, the developers compose a new examination. Once an examination has been constructed, an estimate of the reliability of the estimated θs is provided. This is the third stage of the analysis. It is convenient to use the common statistics from CTT, which are well understood by the developers. Developers, in turn, find this language convenient to explain matters to the examination committee.

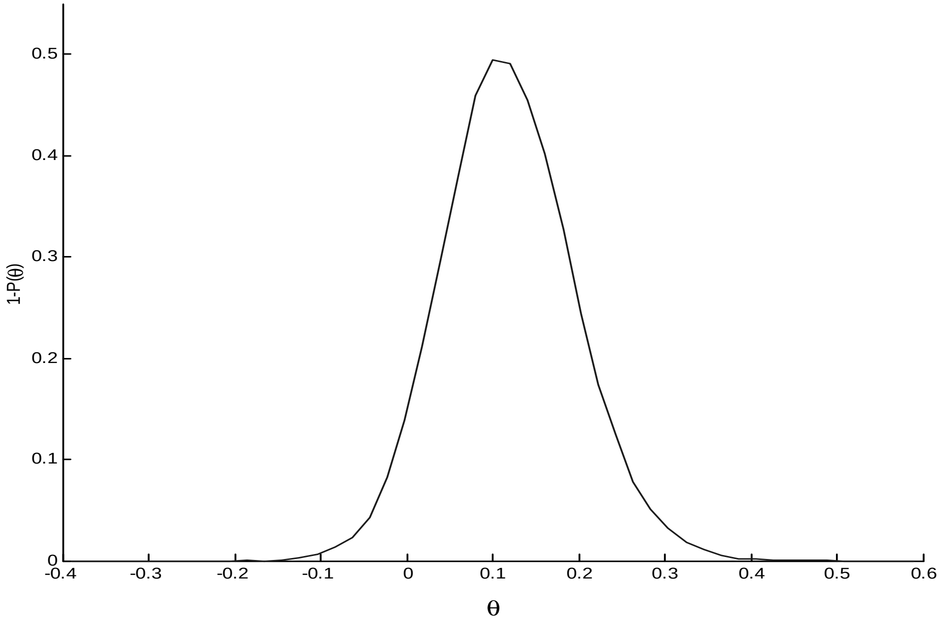
The expected proportion correct has been reported to the developers for some years now, and it appears possible to successfully predict those found in the actual examinations. For instance, of the past nine examinations of listening, the realized expected proportion correct ranged between 0.63 and 0.68, as intended.

Determine the Reliability of Classifications

The tests discussed in the previous section are high-stakes examinations. To gain insight in the quality of the decision made with these tests, $1 - P_o(\theta)$ is presented in Figure 5 for one of the examinations. As one might expect, $1 - P_o(\theta)$ increases to 0.5 when θ becomes close to θ_c, corresponding to the cutoff.

Figure 5

A Plot of 1 Minus the Probability of Consistent Classification, Given θ, for a Test With 40 Binary Items



Note. The generalized partial credit model (GPCM) was estimated with 2, 500 examinees using the method of marginal ML (Muraki, 1992).

It is found that $0.25 \leq 1 - P_o(\theta) \leq 0.50$ for about 16% of the examinees.

This percentage is dependent on the postulated population distribution. In this case, it can be argued that the distribution is unlikely to be normal as the examinees constitute a mixture of immigrants from many different countries. The R_0 test, incorporated in the OPLM software (Glas & Verhelst, 1995), and histograms of estimated θ confirm this argument. When the distribution of estimated θs is considered, the mentioned percentage rises from 16% to 35%. This percentage appears quite high for a high-stakes examination.

Discussion

The aim of this article has been to clarify relations between CTT and IRT, generalize concepts from CTT to IRT, and demonstrate that, when an appropriate IRT model is found, one is able to calculate and use classical indices for properties of items and tests in situations when CTT could normally not be applied. A number of applications have been described ranging from issues in test construction to analysis of examination data. Other applications of this kind have been discussed by Mellenbergh (1994, pp. 227-229), who explains how an IRT model can be used to select a test of items that are parallel in the CTT sense; Kolen, Zeng, and Hanson (1996) also use IRT to estimate the standard errors of scale scores. Monotone, unidimensional IRT models have been considered, but this was not essential. General formulas have been presented here precisely with the aim to facilitate the derivation of reliability and so forth using any IRT model.

It must be noted that all calculations are predicated on the validity of the IRT model, as well as the availability of good estimates of the distribution of the population of interest. Furthermore, it is necessary that the model is sufficiently parameterized but, at the same time, simple enough to admit (approximate) calculation of moments in the population of interest. To assess IRT model fit, most software packages provide a myriad of goodness-of-fit indices, and ways to test IRT models are continuously being developed.

When an IRT model is found appropriate, an impression of the sample variance involved in this study's calculations can be obtained by varying the values of the parameters. For example, if the population distribution is assumed normal with mean μ and variance σ_θ^2, an approximate 95% interval of uncertainty may be constructed by varying σ_θ between $\sigma_\theta^{(\text{low})} = \sigma_\theta - 1.64SE$ and $\sigma_\theta^{(\text{high})} = \sigma_\theta + 1.64SE$, where SE denotes the standard error of the standard deviation. In this case, it is opportune to vary σ_θ because it is estimated with much less precision than the mean and is the main determinant of CTT indices. In the analysis discussed in the fourth section, $SE = 0.149$, which provides the following interval: $\rho_Y^2 \in [0.70 - 0.78]$.

References

Bechger, T. M., Maris, G., Béguin, A., & Verstralen, H. H. F. M. (2003). *Combining classical test theory and item response theory*. R&D Report 2003-4, Cito, Arnhem, The Netherlands.

Bechger, T. M., Verstralen, H. H. F. M., & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika*, *67*, 123-136.

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. 395-479). Reading, MA: Addison-Wesley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213-220.

Davis, P. J., & Rabinowitz, P. (1984). *Methods of numerical integration* (2nd ed.). New York: Academic Press.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Glas, C. A. W., & Verhelst, N. D. (1995). Tests of fit for polytomous Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Gustafsson, J. E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program*. Rep. no. 85, Institute of Education, University of Göteborg.

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.

Lee, W.-C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*, 412-432.

Livingstone, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179-197.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, *45*, 507-530.

Lord, F. M. (1953). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, *17*, 181-194.

Lord, F. M. (1983). *Applications of item response theory to practical testing problems*. Englewood Cliffs, NJ: Lawrence Erlbaum.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, *8*, 453-461.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223-236.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293-299.

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. The Hague, The Netherlands: Mouton.

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

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*, 351-363.

Nicewander, W. A. (1993). Some relationships between the information function of IRT and the signal/noise ratio and reliability coefficient of classical test theory. *Psychometrika*, *58*, 139-141.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen, Denmark: Nissen and Lydicke.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion* [Textbook for test theory and test construction]. Bern, Switzerland: Hans Huber.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, *18*, 229-244.

Sluijter, C. (1998).*Toetsen en beslissen. Toetsing bij doorstroombeslissingen in het voorgezet onderwijs* [Tests and decision making: Making placement decisions in secondary education]. Unpublished doctoral dissertation, Cito, Arnhem, The Netherlands.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the national meeting of the National Council on Measurement in Education, New Orleans, LA.

Steyer, R., & Eid, M. (1993). *Messen und testen* [Measuring and testing]. Berlin: Springer-Verlag.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, *48*, 191-199.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. L. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 161-186). Hillsdale, NJ: Lawrence Erlbaum.

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.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One parameter logistic model OPLM* (Computer software manual). Arnhem, The Netherlands: Cito.

Verstralen, H. H. F. M. (1997a). *A logistic latent class model for multiple choice items*. R&D report, Cito, Arnhem, The Netherlands.

Verstralen, H. H. F. M. (1997b). *OPTAL: Inverse OPLAT and item and test characteristics in populations*. Arnhem, The Netherlands: Cito.

von Davier, M. (1994). *WINMIRA: A Windows 3.x program for analysis with the Rasch model, with the latent class model, and with the mixed Rasch model*. Kiel, Germany: Institute for Science Education (IPN).

Author’s Address

Address correspondence to Timo M. Bechger, Cito, P.O. Box 1034, 6801 MG, Arnhem, The Netherlands; e-mail: timo.bechger@citogroep.nl.