Measurement and Research Department Reports 2003-4

Combining Classical Test Theory And Item Response Theory

Timo Bechger
Anton Béguin
Gunter Maris
Huub Verstralen

Cito groep

Measurement and Research Department Reports 2003-4

COMBINING CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY

Timo Bechger Gunter Maris

Anton Béguin Huub Verstralen

CITO, NATIONAL INSTITUTE FOR EDUCATIONAL MEASUREMENT

ARNHEM

Citogroep
Arnhem, March 3, 2003

Abstract

The present paper is about relations between Classical test theory (CTT) and item response theory (IRT). It is demonstrated that IRT can be used to provide CTT statistics in situations where CTT fails.

1. 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 idea about the relations between IRT and CTT. This should improve our appreciation of both theories and facilitate communication to researchers and item writers who are frequently more familiar with CTT than with IRT. In this paper we survey some of the relations between CTT and IRT, and discuss novel applications of CTT that are feasible using IRT.

This paper is structured as follows: Section 2 provides a brief outline of CTT and its relation to IRT. In Section 3, the CTT concept of reliability is applied in an IRT context. We discuss reliability of estimated latent trait values, and reliability of classifications using a test score. In Section 4, five applications are discussed: (1) to illustrate that reliability can be determined from a single administration of a test; (2) to demonstrate how relations between test characteristics, the population of test takers, and test scores may be explored; (3) to demonstrate how the correlation between latent traits measured by different tests can be calculated; (4) discuss the selection of items from a pilot test when the pilot test could not be administered to the intended population; (5) to describe IRT-based test equating. The paper is concluded in Section 5.

This paper is written in the spirit of work by Verstralen (1997a), Lord (1983), Nicewander (1993), Thissen (1990), Mellenbergh (1994; 1996), and Steyer and Eid (1993) and there is some overlap between these papers and this paper. Naturally, we shall often refer to the time-honoured work by Lord and Novick (1968), which will be abbreviated to L&N just like a friend is often known by a shortened name.

2. Classical Test Theory From an IRT Point of View

2.1. General Introduction

Let an "item" be a means to produce a measurement X. It is assumed that the respondent 's behavior is determined by his value on a vector variable θ which represents what the item intends to measure. This variable may be continuous or discrete but is assumed to be some sort of ability. For ease of presentation, θ is referred to as the subject's ability. 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 and we will use both names interchangeably.

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 (L&N, pp. 38-39). The distribution of the measurement errors has zero mean and variance $Var(X|\theta)$. While the measurement error varies across subjects with the same ability, the true score is a fixed parameter characterizing the combination of an ability and an item.

Taking the expectation of $E[X|\theta]$ over the distribution of θ in the population of interest gives us 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 $Var(X) > 0$,

$$\rho_X^2 \equiv \frac{Var(E[X|\theta])}{Var(X)} \tag{1}$$

$$= 1 - \frac{E[Var(X|\theta)]}{Var(E[X|\theta]) + E[Var(X|\theta)]}, \tag{2}$$

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

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

that is, the measurement error variance in the population. It is customary to denote the reliability as a square since ρ_X^2 equals the square of the correlation between the true score and the observed score (L&N, p. 57). A correlation is not invariant under non-linear transformations and reliability depends on the scoring rule; some scoring rules give higher reliability than others. Equation (1) also shows that item reliability depends on the ability distribution in the population.

L&N 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

$$\Pr(X, X^*) = \int \Pr(X = x|\theta) \Pr(X^* = x^*|\theta) g(\theta) d\theta \quad , \tag{3}$$

where

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

Equation (3) states that the response variables are exchangeable and we shall henceforth call them *exchangeable replications* to indicate that they are independent conditional upon θ, but not marginally. Item reliability equals the correlation between exchangeable replications. This can be seen using *the covariance decomposition formula*:

$$Cov(X, X^*) = Cov(E[X|\theta], E[X^*|\theta]) + E[Cov(X, X^*|\theta)], \tag{4}$$

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

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

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

$$\rho_Y^2 = \frac{Var(E[Y|\theta])}{Var(E[Y|\theta]) + E[Var(Y|\theta)]}. \tag{5}$$

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

$$E[Var(Y|\theta)] = E[Var(X|\theta)] \sum_{i=1}^{I} w_i^2 \quad .$$

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 (L&N, p. 72). Furthermore, if measurement error is assumed to be normally distributed, $Y \pm 1.96\sqrt{E[Var(Y|\theta)]}$ provides an approximate 95% confidence interval for the observed test score.

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

$$ITC_i = \frac{Cov(E[Y|\theta], E[X_i|\theta]) + E[Cov(Y, X_i|\theta)]}{\sqrt{Var(Y)Var(X_i)}}, \tag{6}$$

where the numerator follows from the covariance decomposition formula (4). In CTT, this correlation is interpreted as an item discrimination index because it indicates to what extent the item differentiates between subjects with high scores on the test and subjects with low scores on the test. We will comment upon the interpretation of the ITC shortly.

Since 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 provides an interpretation of the IRC.

Proposition 1. *Assume that the item responses are independent given θ. Let Y_{-i} denote the rest-score. The IRC of item i is equal to*

$$Corr\left(E\left[Y_{-i}|\theta\right], E[X_i|\theta]\right)\sqrt{\rho^2_{X_i}\rho^2_{Y_{-i}}}$$

Proof. By definition:

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

□

It is seen that the IRC is positive and dependent upon the relation of the true rest score and the item true score, which will usually be non-linear. Under exchangeability, $Corr\left(E\left[Y_{-i}|\theta\right], E[X_i|\theta]\right) = Corr\left((I-1)E[X_i|\theta], E[X_i|\theta]\right) = 1$ and

$$\begin{aligned}
\lim_{I\to\infty}\rho^2_{Y_{-i}} &= \lim_{I\to\infty}\frac{(I-1)^2Var(E[X|\theta])}{(I-1)^2Var(E[X|\theta]) + E[Var(X|\theta)](I-1)} \\
&= \lim_{I\to\infty}\frac{Var(E[X|\theta])}{Var(E[X|\theta]) + E[Var(X|\theta)](I-1)^{-1}} \\
&= 1
\end{aligned} \tag{7}$$

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.

2.2. IRT as an Extension of CTT

In practice, it is assumed that the responses to different items are exchangeable so that item reliability can be estimated by their correlation. In CTT, such measures are called "parallel." 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 upon θ, but not necessarily identically distributed. In IRT, this is called *conditional independence (CI)*. For two items, CI is equivalent to

$$\Pr(X_i, X_j) = \int P_{ix_i}(\theta) P_{jx_j}(\theta) g(\theta) d\theta, \tag{8}$$

where $P_{ix_i}(\theta) \equiv \Pr(X_i = x_i|\theta)$ is called the *item category response function (ICRF)*. 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 ICRFs. 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 ability so that the true score is a one-to-one transformation of the ability. Together with CI, these assumptions define the family of unidimensional monotone IRT models which encompasses most existing IRT models.[2]

[2]Note that given CI, $E[Cov(Y, X_i|\theta)] = w_i E[Var(X_i|\theta)]$ and (6) reduces to a more manageable expression.

2.3. The Case of Binary, Equivalent 2PL Items

In this section, we assume that the items are exchangeable measures and introduce an IRT model 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 since all items are equivalent. Without loss in generality, we assume that the ICRFs are modelled by a two-parameter logistic model (2PL); that is,

$$P_1(\theta) = \frac{\exp(\alpha(\theta - \delta))}{1 + \exp(\alpha(\theta - \delta))}, \tag{9}$$

where the parameters $\alpha, \delta \in \mathbb{R}$ are considered known and θ is a scalar ability. The population distribution is unrestricted. The assumption that $P_1(\theta)$ is modelled by the 2PL implies no loss in generality because we can always transform θ such that the ICRFs assume any other functional form. The value of the α-parameter governs the slope of the ICRFs and is therefore interpreted as a *discrimination parameter*, while the category parameter δ is the value of θ where $P_1(\theta) = 1 - P_1(\theta) = 0.5$. When the discrimination parameters are unity we obtain the ubiquitous Rasch model (Rasch, 1960). For the purpose of illustration we have drawn the ICRFs and the true score for a 2PL in Figure (1).

With binary items, the item true score equals the probability of a correct response, given θ. The conditional measurement error variance of the score for each item is equal to $P_1(\theta)(1 - P_1(\theta))$. Using the formulae in the previous section we find that

$$Var(X) = E[P_1(\theta)](1 - E[P_1(\theta)]), \tag{10}$$

where $E[P_1(\theta\delta)]$, that is, the expected percentage correct, is known as the difficulty of the item. The true-score variance for any item is given by

$$Var(E[X|\theta]) = E[(P_1(\theta))^2] - E[P_1(\theta)]^2, \tag{11}$$

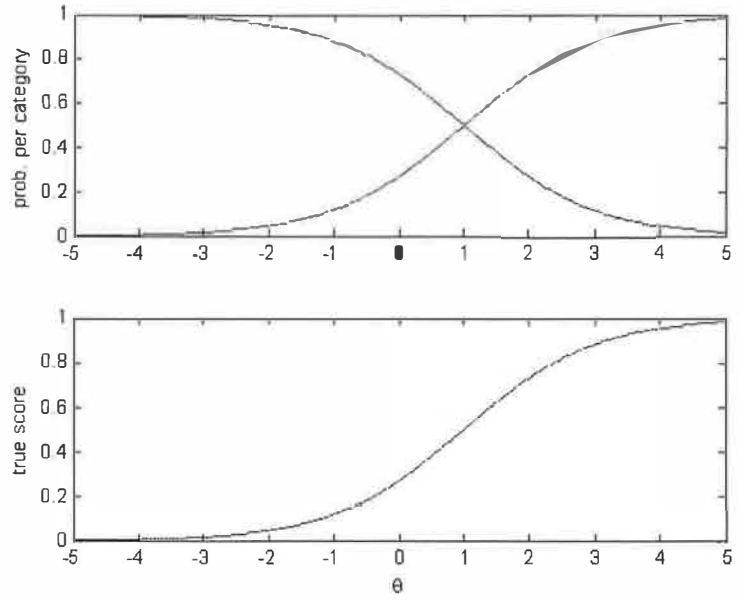

FIGURE 1.

The upper figure shows ICRFs for a GPCM item with $\delta = 1$, and $\alpha = 2$. The lower figure shows the true score as a function of theta.

which equals $Var(P_1(\theta))$; the variance of the proportion correct in the reference population. Note that $E[(P_1(\theta))^2] = \Pr(X_i = 1, X_j = 1)$ when i and j index two equivalent binary items. The item reliability follows from substitution of (10) and (11) in (1). It is seen that under the present assumptions, item reliability equals Loevinger's (1948) H-coefficient which is used in Mokken scale analysis (Mokken, 1971, p. 150).

The expected unweighted sum score on a test with I equivalent items, given θ, equals $I \times P_1(\theta)$. The reliability of the test score is given by

$$\begin{aligned} \rho_Y^2 &= \frac{I^2 Var(P_1(\theta))}{I\,[(I-1)Var(P_1(\theta)) + Var(X)]} \\ &= \frac{I\rho_X^2}{(I-1)\rho_X^2 + 1}. \end{aligned} \tag{12}$$

The equation is well-known as the *Spearman-Brown (SB) formula.* If $I = 1$, for

instance, $\rho_Y^2 = \rho_X^2$ the reliability of a single item score. For $I = I^* + Z$, we obtain the reliability of a test with I^* items when it is lengthened by adding Z equivalent items. The SB formula shows that the reliability of the test score goes to 1 if I becomes large.

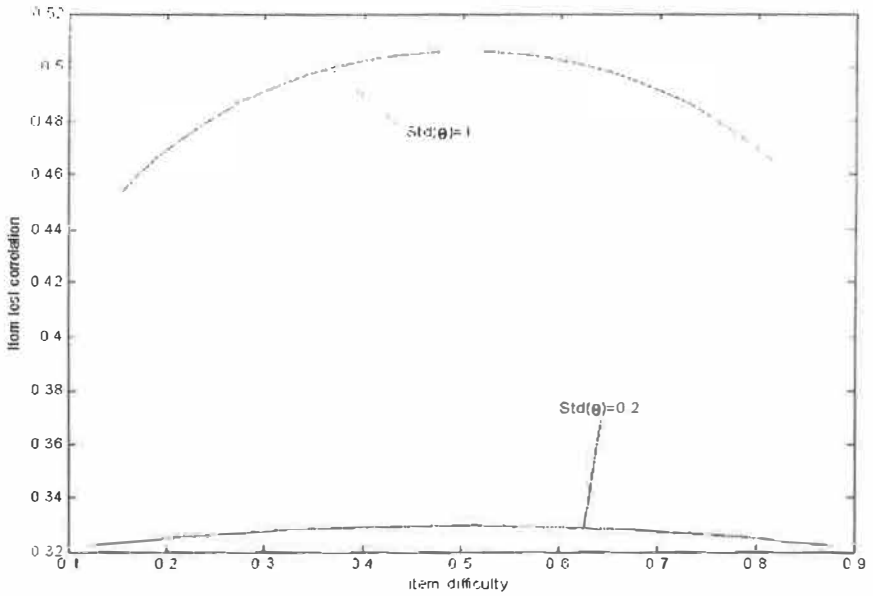

FIGURE 2.

Relation between item difficulty and the item test correlation assuming 10 equivalent GPCM items ($\alpha = 1$). We assume that the distribution of θ is standard normal.

Remark 1. *Let Y denote the unweighted score. Assume that $Var(E[Y|\theta]) = I^2 Var(P_1(\theta))$ and $Cov(X_i, X_j) = I(I-1)Var(P_1(\theta))$ so that*

$$Var(Y) = \sum_i Var(X_i) + \sum_i \sum_{j \neq i} Cov(X_i, X_j) = \sum_i Var(X_i) + I(I-1)Var(P_1(\theta))$$

Then,

$$\begin{aligned}
\rho_Y^2 &= \frac{Var(E[Y|\theta])}{Var(Y)} \\
&= \frac{I^2 Var(P_1(\theta))}{Var(Y)} \\
&= \left[\frac{I(I-1)Var(P_1(\theta))}{Var(Y)}\right]\frac{I}{I-1} \\
&= \left[\frac{Var(Y)-\sum_i Var(X_i)}{Var(Y)}\right]\frac{I}{I-1} \\
&= \left[1-\frac{\sum_i Var(X_i)}{Var(Y)}\right]\frac{I}{I-1}.
\end{aligned}$$

This equation is known as Cronbach's alpha *(Cronbach, 1951) and it is widely used to estimate reliability. Let N denote the number of respondents and p_i the percentage of them that have answered correctly to item i. In practice, $Var(X_i)$ is estimated by $Np_i(1-p_i)$ and $Var(Y)$ by the observed variance of the sum-scores. It can be shown that Cronbach 's alpha provides an underestimate of reliability if the assumptions do not hold. Alternative estimates are surveyed by Verhelst (1998).*

A bit of algebra shows that, under the present assumptions, Equation (6) simplifies to

$$ITC_i = \sqrt{\frac{(I-1)}{I}\rho_X^2 + \frac{1}{I}} \Leftrightarrow \tag{13}$$

$$\rho_X^2 = \frac{ITC_i^2 I - 1}{I-1}. \tag{14}$$

As expected from Proposition (1) $ITC_i \approx \sqrt{\rho_X^2}$ when the number of items is large.

A plot of the ITC against the difficulty 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 dispersion of θ, as well as on the number of items in the test (see also Steyer and Eid, 1993, p. 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).

3. Reliability in IRT

3.1. Item and Test Information

Consider a poly(cho)tomous item i, with $J_i + 1$ response categories indexed $0, 1, ..., J_i$. The scoring rule is that X_i takes the value of the index of the category that is chosen; i.e., $X_i = 2$ if category 2 was chosen. This scoring rule makes sense if category j reveals more ability than category $j + 1$. The extension of the 2PL for polytomous items called *the generalized partial credit model (GPCM)* (Muraki, 1992). The GPCM implies that

$$P_{ix_i}(\theta) = \frac{1}{D_i} \exp(\alpha_i \sum_{p=1}^{x_i} (\theta - \delta_{ip})), \tag{15}$$

where $\sum_{p=1}^{0}(\theta - \delta_{ip}) \equiv 0$, $\delta_i = (\delta_{i1}, .., \delta_{iJ_i})$, and D_i is a constant that is added to make sure that $\sum_{x_i=0}^{J_i} P_{ix_i}(\theta) = 1$. The *category parameters*, δ_{ip}, are the values of θ where the ICRFs of adjacent categories are equal. It is assumed that $\delta_{i1} < \delta_{i2} < \cdots < \delta_{iJ_i}$. For illustration purposes, we have drawn the ICRFs of a GPCM item with four categories in Figure (3). Figure (3) also shows a plot of the true score as a function of θ, and the information function of the item which will be defined shortly.

Let $L(\theta | X_i = x_i) = P_{ix_i}(\theta)$ denote the likelihood function of θ given the observed

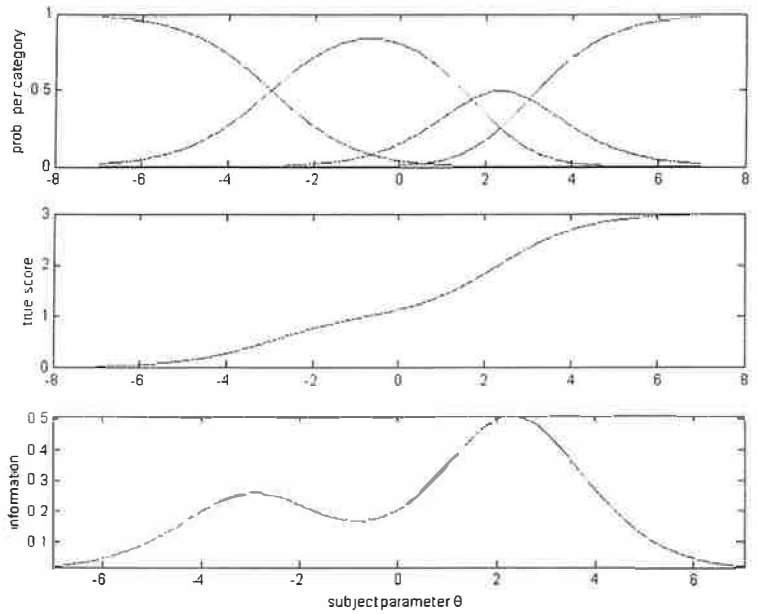

FIGURE 3.

ICRFs, true score and information for an GPCM item with four categories; $\delta_i = (-3, 1.67, 3)$, and $\alpha_i = 1$. Note that the information function is not unimodal (cf. Akkermans, and Muraki, 1997)

response. The *item (Fisher) information function* is defined as

$$\begin{aligned} Inf_{X_i}(\theta) &\equiv E\left[\left(\frac{\partial}{\partial\theta}\ln L(\theta|X_i)\right)^2|\theta\right] \\ &= P_{i0}(\theta)\left(\frac{\partial}{\partial\theta}\ln P_{i0}(\theta)\right)^2 + \cdots + P_{iJ_i}(\theta)\left(\frac{\partial}{\partial\theta}\ln P_{iJ_i}(\theta)\right)^2 \\ &= \sum_{x_i=0}^{J_i}\frac{\left[\frac{\partial}{\partial\theta}P_{ix_i}(\theta)\right]^2}{P_{ix_i}(\theta)}. \end{aligned} \tag{16}$$

This shows that, in general, the item information depends on the combined rate of change in the ICRFs. The item information function of the GPCM is found to be equal to $\alpha_i^2 Var(X_i|\theta) = Var(\alpha_i X_i + b_i|\theta)$ for some constant b_i. Thus, in the GPCM the item information equals the conditional measurement error variance of $\alpha_i X_i + b_i$ and is consequently dependent upon the scoring rule. It is interesting therefore to investigate the effect of different scoring rules on the information function.

When the item responses are independent given θ, the *test information function* is the sum of the item information functions, i.e., $TInf(\theta) \equiv \sum_{i=1}^{I} Inf_{X_i}(\theta)$, and

$$E[TInf(\theta)] = \sum_{i=1}^{I} E[Inf_{X_i}(\theta)]. \tag{17}$$

In the next paragraph, the expected test information will be related to the reliability of estimated abilities.

3.2. 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 non-linear, however. When estimates of θ are reported and used it is therefore appropriate to provide the reliability of the estimated ability values $\hat{\theta}$.

To derive this reliability we first note that

$$\hat{\theta}_v = E[\hat{\theta}|\theta = \theta_v] + e, \tag{18}$$

where $e \equiv \hat{\theta}_v - E[\hat{\theta}|\theta = \theta_v]$ can be interpreted as "measurement error", and $E[\hat{\theta}|\theta = \theta_v]$ as a "true score." Subscript v denotes a generic subject. Reliability is defined as the proportion of true variance in the reference population and we find that

$$\rho^2_{\hat{\theta}} = \frac{Var(E[\hat{\theta}|\theta])}{Var(\hat{\theta})} \tag{19}$$

$$= 1 - \frac{E[Var(\hat{\theta}|\theta)]}{Var(E[\hat{\theta}|\theta]) + E[Var(\hat{\theta}|\theta)]}, \tag{20}$$

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 a measure of linear association between exchangeable replicates of $\hat{\theta}$. This means that $\rho^2_{\hat{\theta}}$ changes if $\hat{\theta}$ is non-linearly 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 $E[\hat{\theta}|\theta] = \alpha_1\theta + \alpha_2$, $(\alpha_1, \alpha_2 \in \mathbb{R})$ since $\hat{\theta}$ 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

$$Corr(\theta, \hat{\theta}) = \frac{Cov(\theta, \theta + Bias(\theta))}{\sqrt{Var(\hat{\theta})Var(\theta)}} \tag{21}$$
$$= \sqrt{\frac{Var(\tilde{\theta})}{Var(\hat{\theta})}}\sqrt{\frac{Var(\theta)}{Var(\tilde{\theta})}} + Corr(Bias(\theta), \theta)\sqrt{\frac{Var(Bias(\theta))}{Var(\hat{\theta})}},$$

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

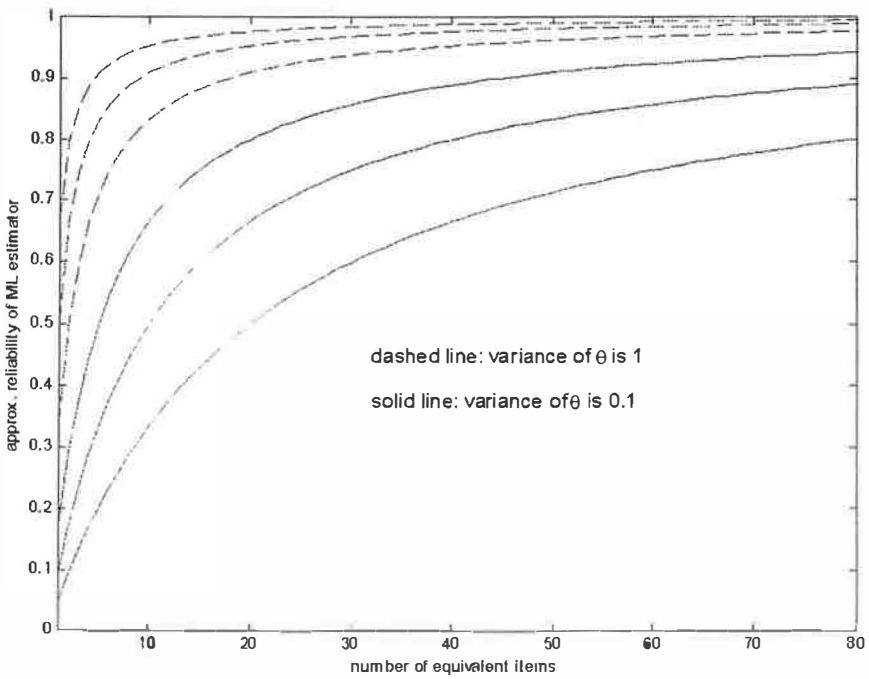

FIGURE 4.

Plot of $\rho^2_{\hat{\theta}}$ against the number of equivalent items in a test. The curves differ in the expected information of each item and the dispersion of θ.

When a ML or Warm estimator (Warm, 1989) is employed, $Var(\hat{\theta}|\theta)$ is equal to

the inverse of the test information function, $TInf^{-1}(\theta)$, when the number of items becomes large. If $E[Var(\hat{\theta}|\theta)] = E[TInf(\theta)]^{-1}$ is substituted in Equation (20), and bias is ignored (i.e., $Var(E[\hat{\theta}|\theta]) = Var(\theta)$), it follows that

$$\rho_{\hat{\theta}}^2 \approx \frac{E[TInf(\theta)]Var(\theta)}{1 + E[TInf(\theta)]Var(\theta)} \tag{22}$$

(see Thissen, 1990; Mellenbergh, 1994, Equation 22; Samejima, 1994, Equation 21). Given $Var(\theta)$, the reliability depends exclusively on $E[TInf(\theta)]$. Thus, if one reports ability estimates, it is desirable to have high expected test information in the population of interest. An alternative approximation to $\rho_{\hat{\theta}}^2$ is discussed by Verhelst, Glas and Verstralen (1995, p. 64), and Rost (1996, pp. 353-354).

3.3. 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. Furthermore, subjects with test scores less than c will fail the test and subjects 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:

$$\Pr(I_p = 1|\theta) = \sum_{y=c}^{\max(Y)} \Pr(Y = y|\theta) \tag{23}$$

$$= \sum_{y=c}^{\max(Y)} \left[\sum_{\mathbf{x}:\sum_i w_i x_i = y} \prod_i \Pr(X_i = x_i|\theta) \right], \tag{24}$$

The marginal probability of passing equals $\Pr(I_p = 1) = E[\Pr(I_p = 1|\theta)]$. The calculation of $\Pr(Y = y|\theta)$ is discussed in the Appendix.

If we apply the definition of reliability, given in Equation (1), to the variable I_p

we obtain the reliability of classification; that is,

$$\rho^2_{Clas} = \frac{Var(E[I_p|\theta])}{Var(E[I_p|\theta]) + E[Var(I_p|\theta)]} = \frac{E[\Pr(I_p = 1|\theta)^2] - E[\Pr(I_p = 1|\theta)]^2}{E[\Pr(I_p = 1|\theta)] - E[\Pr(I_p = 1|\theta)]^2}. \tag{25}$$

It was demonstrated earlier that ρ^2_{Clas} equals the correlation between classifications across two exchangeable administration 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. Cohen 's kappa is equal to

$$\kappa = \frac{P_o - P_c}{1 - P_c}, \tag{26}$$

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 change. Superscript* (r) *denotes that the random variable is registered at the* r*th administration. Let* $I_p^{(r)}$ *denote passing on the* r*th administration. 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 since $\Pr(I_p^{(1)} = 1|\theta) = \Pr(I_p^{(2)} = 1|\theta)$, *by assumption. In*

the same way we find that

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

If we expand P_o and P_c, and substitute the resulting expressions in Equation (26) we find that:

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

This ends the proof. □

We have seen occasions where decision reliability could actually be calculated as Cohen's kappa because there were two independent ratings of the same subjects. 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.

Imagining two exchangeable administrations of the same examination, the probability of consistent classification given θ equals

$$P_o(\theta) = \Pr(I_p^{(1)} = 1, I_p^{(2)} = 1|\theta)] + \Pr(I_p^{(1)} = 0, I_p^{(2)} = 0|\theta) \tag{27}$$

$$= \left[\sum_{y=c}^{\max(Y)} \Pr(Y = y|\theta)\right]^2 + \left[\sum_{y=0}^{c-1} \Pr(Y = y|\theta)\right]^2 . \tag{28}$$

This function is called the *test characteristic decision curve (TCDC).* The probability of inconsistent classification is, of course, $1 - P_o(\theta)$. When the TCDC is integrated over the reference population we obtain the probability of consistent classification when the test is applied to the reference population using $y = c$ as a cutoff. 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.

In general, respondents are classified into C mutually exclusive categories using their test score. Let I_c $(c = 1, ..., C)$ denote a random variable that is 1 if a respondent's test score falls in the cth category, and 0 otherwise. If $C > 2$, The weighted version of Cohen's kappa (Cohen, 1968) may be taken as a *weighted index of reliability of classification.* To be more specific:

$$\rho^2_{Clas,\mathbf{w}} = \frac{P_o - P_c}{1 - P_c}, \tag{29}$$

where

$$\begin{aligned} P_o &= E[P_o(\theta)] \\ &= E\left[\sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij} \Pr(I_i = 1|\theta) \Pr(I_j = 1|\theta)\right], \end{aligned} \tag{30}$$

and

$$\begin{aligned} P_c &= \sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij} \Pr(I_i = 1) \Pr(I_j = 1) \\ &= \sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij} E[\Pr(I_i = 1|\theta)] E[\Pr(I_j = 1|\theta)]. \end{aligned} \tag{31}$$

Expressions similar to (23) may be used to calculate $\Pr(I_i = 1|\theta)$, $i = 1, .., C$. The weights, w_{ij}, are chosen on substantive grounds to express the relative similarities among the categories; $w_{ij} = 1$ if $i = j$ and 0 if $i \neq j$ yields Cohen's kappa. Following Yang en Chen (1978), we assume that $0 \leq w_{ij} \leq 1$, $w_{ii} = 1$, and $w_{ij} = w_{ji}$, for all $i, j = 1, ..., C$. This is not a serious restriction on the weights. If test constructors decide that the numbers d_{ij} $(i, j = 1, ..., C)$ express the difference between the categories, the weights can be calculated as

$$w_{ij} = 1 - \frac{d_{ij}}{\max_{i,j}(d_{ij})}. \tag{32}$$

Alternative ways to quantify and investigate the quality of classifications are discussed by Livingston and Lewis (1995), Verstralen (1997b), 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.

4. Applications

In this section we discuss a number of applications involving the combined use of CTT and IRT. In the first paragraph, we illustrate that reliability can be determined from a single administration of a test using numerical integration. In the second paragraph, we demonstrate how relations between test characteristics, the population of test takers, and test scores may be explored graphically. In the third paragraph, we demonstrate how the correlation between latent traits measured by different tests can be calculated. In the fourth paragraph, we discuss the selection of items from a pilot test when the pilot test could not be administered to the intended population. We describe how we assist item writers in the construction of an examination and also how we determine the reliability of classification of an existing examination. Finally, in the fifth paragraph we briefly describe IRT-based test equating.

4.1. Calculating Reliability with a single Administration of a Test

The easiest application is to use formulae in the first section to calculate reliability using a single test administration.[3] To illustrate this possibility, we use the so-called "KFT data" that are listed on page 99 and 100 in the book by Jürgen Rost (1996).[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

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

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

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 and Rabinowitz, 1984, chapter 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
	$\rho^2_{\hat{\theta}} = 0.74$	$\rho^2_Y = 0.75$	

It turns out that the items are nearly parallel so that Cronbach's alpha is, in this case, only slightly lower than the estimated test reliability. The reliability of the unweighted test score equals that of the estimated abilities. This is to be expected since the unweighted scores carries all the information used to estimate abilities and the relation between unweighted scores and estimated abilities is approximately linear.

4.2. Investigate Relations Between the IRT model, the Population, and Properties of the Items and the Test

Plots are often instrumental to illustrate relations between test characteristics as determined by an IRT model, the population of test takers, and test scores, especially when such relations can not be described analytically. For example, Lord (1953; L&N, fig. 16.14.1 through 16.14.6) uses plots of the relation between ability and 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 formulae presented here may be used to produce such plots.

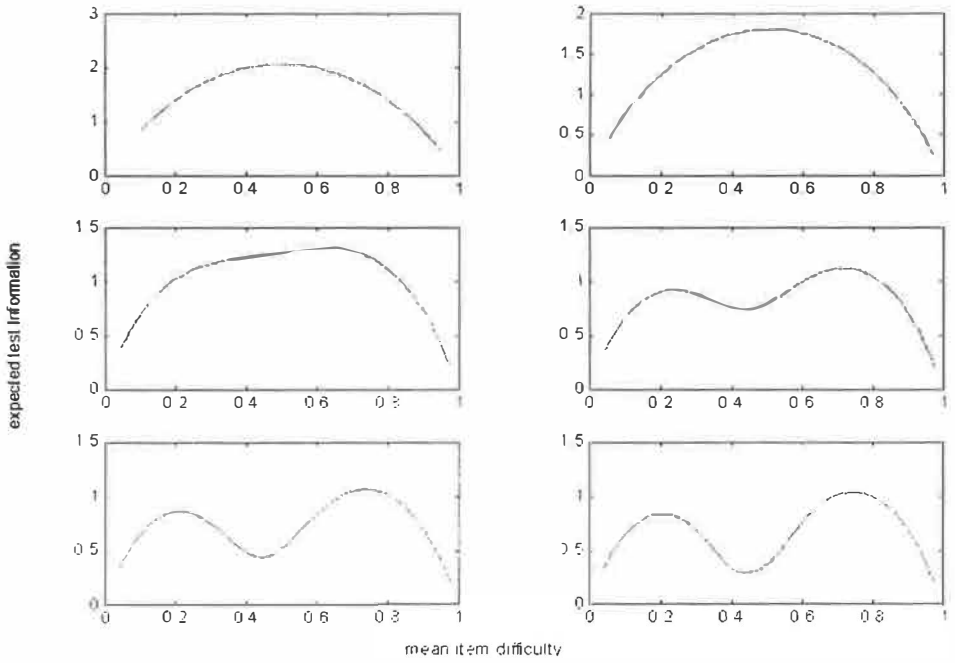

FIGURE 5.

Expected test information plotted against mean item difficulty for varying item category parameters. Mean difficulty increases row wise. The model used was a generalized partial credit model (Muraki, 1992).

There were various examples in the previous sections, such as Figure (2). It is interesting to investigate how a particular test will behave when administered

to different populations. An illustration is provided by Figure (5) which shows the relation between the expected test information and the mean item difficulty in a standard normal population. The mean difficulty was varied by varying the value of the mean of the population. The test consists of 10 binary items. All discrimination parameters were set to 1 but the category parameters were systematically varied over the plots. The plot in the upper left corner was produced with all category parameters equal to zero. In the ensuing plots, items 1 to 4 were systematically made more easy while items 6 to 10 were gradually made more difficult. These plots illustrate that the expected test information (and $\rho^2_{\hat{\theta}}$) is not necessarily high when the mean item difficulty is close to 0.50. The test provides little information about the abilities of the respondents if it consists of items that are either very difficult or very easy for the population of interest (see also Muraki, 1993).

One more illustration is provided by Figure (6), which shows the effect of the discrimination parameter on the TCDC. It is seen that the TCDC becomes more concentrated when the discrimination parameter increases. This illustrate that the quality of a decision increases when items discriminate better.

4.3. Calculating the Correlation Between Two Latent Traits

Let $Corr(\hat{\theta}, \hat{\xi})$ denote the correlation between the estimates of θ and estimates of some other latent trait ξ. If both estimates are unbiased, it can be shown that

$$Corr(\theta, \xi) = Corr(\hat{\theta}, \hat{\xi}) / \sqrt{\rho^2_{\hat{\theta}} \rho^2_{\hat{\xi}}}, \tag{33}$$

where $Corr(\theta, \xi)$ denotes the correlation between θ and ξ. Suppose we have two tests with one test being a measure of a latent trait θ, and the other test a measure of a latent trait ξ. Equation (33) shows that $Corr(\hat{\theta}, \hat{\xi})$ may be much lower than $Corr(\theta, \xi)$ if the estimates are unreliable. This is called "attenuation."

While $Corr(\hat{\theta}, \hat{\xi})$ may be estimated from the data, we need the reliabilities in

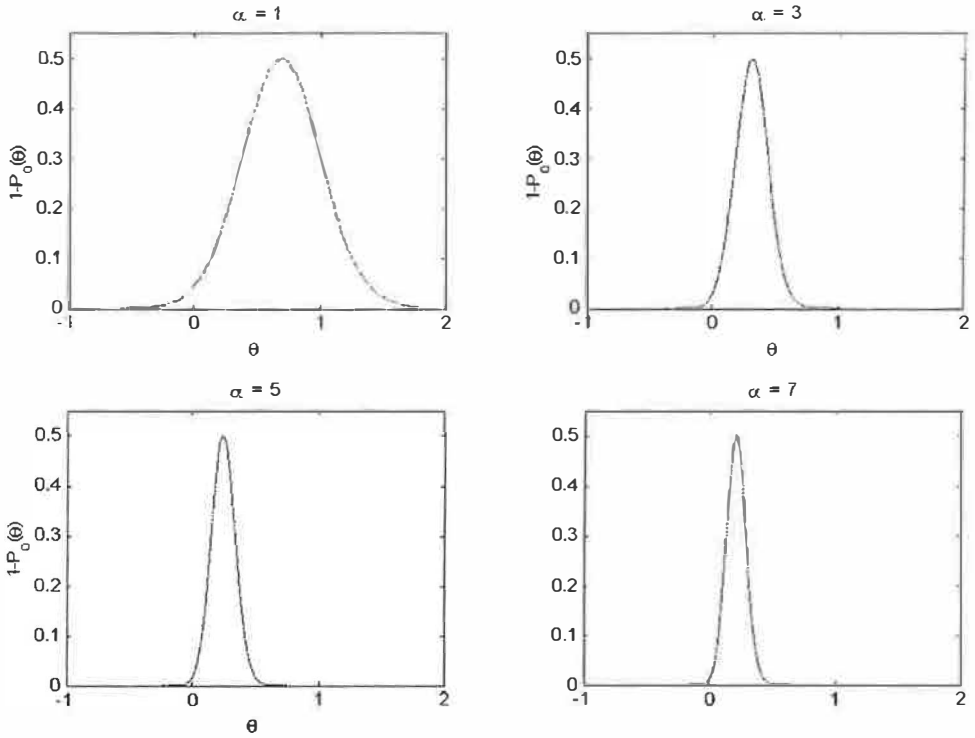

FIGURE 6.

Plots illustrating the effect of the discrimination parameters on the TCDC. The TCDC are based upon the model used to draw Figure (8).

order to correct $Corr(\hat{\theta}, \hat{\xi})$ for attenuation and calculate $Corr(\theta, \xi)$. There are at least three ways to calculate $Corr(\theta, \xi)$. First, when either the ML or the Warm estimator is used, approximate reliabilities can be obtained from Equation (22), using numerical integration, if necessary, to calculate the expected test information. A second procedure becomes feasible when the estimated θ is a one-to-one function of the test score alone; i.e., when Y is minimal sufficient for θ as in the Rasch model. The IRT model gives the distribution of the test score Y given θ; $g(Y = y|\theta)$, where y are the values taken by Y. Each value y results in an estimated ability $\hat{\theta}(y)$ and $g(Y = y|\theta) = g(\hat{\theta} = \hat{\theta}(y)|\theta)$ is the distribution of the estimated abilities given θ.

The variance of $\hat{\theta}$ given θ may now be calculated as

$$Var(\hat{\theta}|\theta) = E[\hat{\theta}^2|\theta] - E[\hat{\theta}|\theta]^2 \tag{34}$$
$$= \sum_y \hat{\theta}^2(y) g(\hat{\theta} = \hat{\theta}(y)|\theta) - \left(\sum_y \hat{\theta}(y) g(\hat{\theta} = \hat{\theta}(y)|\theta) \right)^2,$$

and

$$Var(E[\hat{\theta}|\theta]) = E[E[\hat{\theta}|\theta]^2] - E[E[\hat{\theta}|\theta]]^2 \tag{35}$$

We then calculate the reliabilities via Equation (19) using numerical integration to approximate the expectations, if necessary. Note that the first procedure is based upon the assumption that there is no bias and $E[\hat{\theta}|\theta] = \theta$. The second procedure is expected to be more robust against bias in the estimates. Finally, the correlation may be estimated using the method of maximum likelihood, considering the item and population parameters known. This procedure was described in detail by Verhelst and Veldhuijzen (2002) in an internal report.

4.4. Constructing Examinations

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. A GPCM 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 abilities are transformed to scale scores that serve as examination marks. In this paragraph, we will briefly describe how we assist the item writers with the construction of new examinations for listening and reading. An alternative equating procedure is discussed in the next section.

The construction of a new examination is preceded by a pilot study which entails

the administration of new items to a sample of candidates that participate in a language course. The purpose of the pilot study is to select the items for future examinations. After the data have been collected, they are added to a large data set which contains the data obtained from previous pilot studies and examinations. This data set is called *the data bank*.

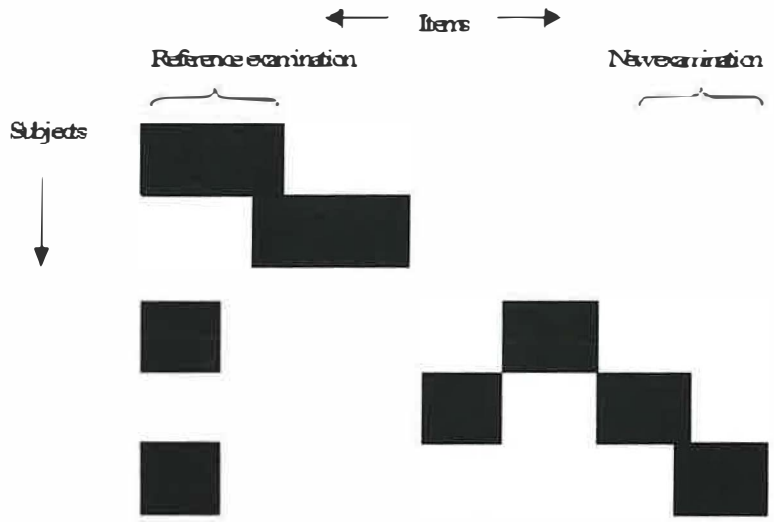

FIGURE 7.
Schematic representation of the data bank.

Schematically, the data bank can be represented by a matrix where the rows are subjects and the columns are items. In Figure (7), the shaded areas represent realized item responses, while the blank areas represent missing responses. The systematic pattern of missing and observed data arises naturally because items are administered in so-called "booklets". While an examination usually consists of a single booklet, the items are spread over various booklets in the pilot to lessen the burden for respondents and allow a large number of items to be tested. The reference test is a subset of the items in the data bank. This reference test was chosen by the examination committee considered a valid and reliable measure of the ability of interest. The reference population is the population of examinees who are generally

more able than the subjects that participate in the pilot study.

The analysis of the pilot data consists of three stages. The IRT model used is the GPCM. We first establish a fitting GPCM using all relevant parts of the data bank and recommend the item writers to discard items that do not conform to the model, and/or items with negative or very low α-parameters. In the second stage, we give the item writers three additional pieces of information. First, we provide the item difficulties in the reference population. The examination committee strives at difficulties between 0.50 and 0.70. Second, we supply expected item information and recommend that those items be included that have the highest values. Thus, we intend to maximize the expected test information of the new examination and the reliability of the estimated abilities. Thirdly, we provide IRC_is 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 item writers then compose a new examination. Once an examination has been constructed, we estimate the reliability of the estimated abilities. This is the third stage of the analysis. The item writers find it convenient to use the common statistics from CTT.

The item difficulties have been reported to the item writers for some years now and it appears that we have been quite successful in predicting the item difficulties in the actual examinations. When, for instance, we look at the last nine examinations of listening, the realized item difficulties ranged between 0.63 and 0.68 as intended. We have not yet gained enough experience with the expected information or the IRCs.

The Reliability of Classifications The examinations discussed in the previous paragraph are high-stakes examinations. In order to gain insight in the quality of the decision made with these tests, we have drawn $1 - P_o(\theta)$ in Figure (8) for

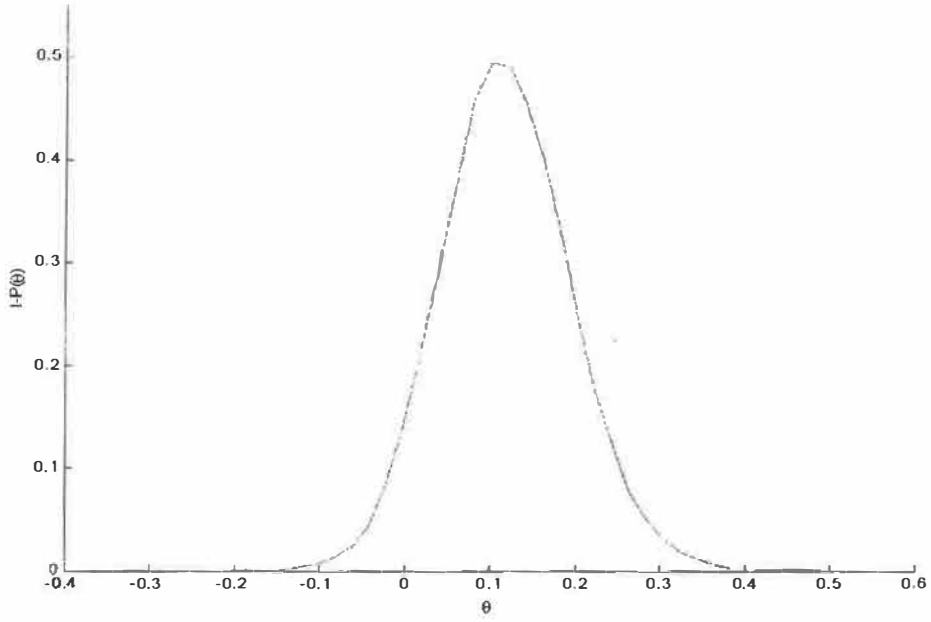

FIGURE 8.

A TCDC for a test with 40 binary items. The GPCM was estimated with 2500 examinees using the method of marginal ML (Muraki, 1992).

one of the examinations. As one might expect, $1 - P_o(\theta)$ increases to 0.5 when θ becomes closer to the ability θ_c corresponding to the cutoff.It is found that $0.25 \leq 1 - P_o(\theta) \leq 0.50$ for about 16% of the examinees. This percentage is dependent upon 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 of immigrants from many different countries. The R_0 test, incorporated in the OPLM software (Glas and Verhelst, 1995), and histograms of estimated abilities confirm this argument. When we consider the distribution of estimated abilities, the mentioned percentage rises from 16 to 35%. This percentage appears much too high for a high-stakes examination but, was to be expected. First, high percentages of inconsistent classifications have been found before using a procedure that assumes that the score distributions are bivariate normal with correlation ρ_Y (Verhelst, 2002a). Second, we looked at data from an examination of the ability to speak. The examinee 's

performance is evaluated twice by independent judges so that we are able to observe the agreement between two independent evaluations of the examinees and obtain an estimate of the reliability of classification. Over examinations, this reliability is found to be remarkably stable at a value of about 0.46.

4.5. Test Score Equating Using an IRT Model

Adjusting scores on different test forms so that they can be compared and used interchangeably is called test equating. Once an examination is constructed and administered to examinees, it is desirable that the scores be equated to the scores on the reference exam to ensure that the achievement of present examinees be comparable to that of previous examinees. This aim could be achieved by means of an *IRT based score equating* procedure (Zeng and Kolen, 1995). This procedure consists of three steps. In the first step, the parameters of an IRT model are estimated. In the second step, the test score distribution of respondents from the reference population on the new examination is determined based on the estimated item and population parameters obtained in the first step. Specifically, for a score point y the expected frequency can be calculated as:

$$n_{ref}\Pr(y) = n_{ref}\sum_{\mathbf{x}:\sum_i w_i x_i = y}\int \Pr(\mathbf{x}|\theta)g_{ref}(\theta)d\theta \tag{36}$$

$$= n_{ref}\int\left[\sum_{\mathbf{x}:\sum_i w_i\, x_i = y}\prod_i \Pr(X_i = x_i|\theta)\right] g_{ref}(\theta)d\theta \tag{37}$$

$$= n_{ref}\int \Pr(Y = y|\theta)g_{ref}(\theta)d\theta, \tag{38}$$

where n_{ref} is the number of respondents in the sample from the reference population, $\mathbf{x}$ a response vector on the present exam, and $g_{ref}(\theta)$ the ability density function in the reference population. The calculation of $\Pr(Y = y|\theta)$ is discussed in the Appendix. In the third step, the test score distribution is used in the determination

of equivalent scores according to any of the available "observed-score" test equating methods (see Kolen and Brennan, 1995 for a survey of such methods). Equipercentile equating, for instance, is based upon the argument that the passing score should be chosen in such a way that the percentage of examinees from the reference population that pass should be equal on the present exam and the reference exam.

5. Generating Exchangeable Test Administrations to Obtain Unbiased Estimates of Statistics of Interest and a Lower Bound for their Sampling Variability

All calculations that were proposed so far are predicated on knowledge of the IRT model and the distribution in the population of interest. In practice, the parameters of the IRT model and the population distribution are estimated and we need to take their sampling error into account. Here we describe a Monte Carlo procedure to obtain a number of exchangeable samples. The generated data can be used to estimate the statistic of interest and the variance over generated samples will provide a consistent estimate of its sampling variance.

Let λ denote all parameters of the IRT model and the population distribution. Here, we take a Bayesian point of view and consider the parameters random variables with prior distribution $\Pr(\theta, \lambda) = \Pr(\theta)\Pr(\lambda)$. First, let $Y^{(1)}$ and $Y^{(2)}$ denote two exchangeable test or item scores; i.e.,

$$\Pr(Y^{(1)}, Y^{(2)}) = \int \Pr(Y^{(2)}|\theta, \lambda)\Pr(Y^{(1)}|\theta, \lambda)\Pr(\theta)\Pr(\lambda)d(\theta, \lambda) \qquad (39)$$

We assume that "nature" has provided us with an identical and independently distributed (i.i.d.) sample from $\Pr(Y^{(1)})$. Now we wish to generate values of $Y^{(2)}$ such that the generated data and the observed data are realizations from $\Pr(Y^{(1)}, Y^{(2)})$.

The observed data are kept constant and we generate $Y^{(2)}$ from $\Pr(Y^{(2)}|Y^{(1)})$, where

$$\Pr(Y^{(2)}|Y^{(1)}) = \frac{\Pr(Y^{(1)}, Y^{(2)})}{\Pr(Y^{(1)})} \tag{40}$$

$$= \int \Pr(Y^{(2)}|\theta, \lambda) \frac{\Pr(Y^{(1)}|\theta, \lambda) \Pr(\theta) \Pr(\lambda)}{\Pr(Y^{(1)})} d(\theta, \lambda) \tag{41}$$

$$= \int \Pr(Y^{(2)}|\theta, \lambda) \Pr(\theta, \lambda|Y^{(1)}) d(\theta, \lambda) \tag{42}$$

To this aim, we employ the method of composition (e.g., Tanner, 1993, section 3.3.2) and generate an i.i.d. sample from $\Pr(Y^{(2)}|\theta, \lambda) \Pr(\theta, \lambda|Y^{(1)})$. First, we must draw θ^* and λ^* from the posterior distribution $\Pr(\theta, \lambda|Y^{(1)})$. Then, we must draw $y_*^{(2)}$ from $\Pr(Y^{(2)}|\theta^*, \lambda^*)$. These steps are repeated N times to yield the desired sample; $(y_1^{(1)}, y_1^{(2)}), ..., (y_N^{(1)}, y_N^{(2)})$. We can do so repeatedly and generate $(y_1^{(1)}, y_1^{(2)}, ..., y_1^{(B)}), ..., (y_N^{(1)}, y_N^{(2)} ..., y_N^{(B)})$. Each of the samples is conditional upon the observed data. The main problem is to construct an algorithm to produce a sample from $\Pr(\theta, \lambda|Y^{(1)})$; the next step is easy; sampling data from $\Pr(Y^{(2)}|\theta, \lambda)$ is simply generating data from the item response model. The following picture schematically depict the procedure.

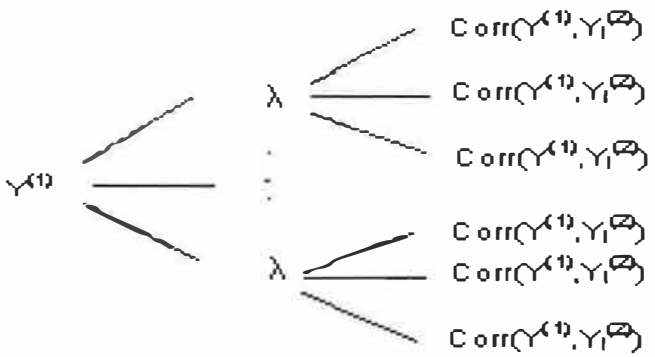

FIGURE 9.

A schematic display of the sampling procedure.

By averaging the correlations we obtain $E[Corr(Y^{(1)}, Y^{(2)})|Y^{(1)}]$. Since

$$E\left\{E[Corr(Y^{(1)}, Y^{(2)})|Y^{(1)}]\right\} = E[Corr(Y^{(1)}, Y^{(2)})] \quad ,$$

the estimator is unbiased. However

$$Var\left\{E[Corr(Y^{(1)}, Y^{(2)})|Y^{(1)}]\right\} < Var(Corr(Y^{(1)}, Y^{(2)}))$$

and we have merely obtained a lower bound of the variance that we want. The reason is that $Y^{(1)}$ is fixed.

Since $\Pr(\theta, \lambda|Y^{(1)}) = \Pr(\theta|Y^{(1)}; \lambda)\Pr(\lambda|Y^{(1)})$, we write

$$\Pr(Y^{(1)}, Y^{(2)}) = \int\int \Pr(Y^{(2)}|\theta, \lambda)\Pr(Y^{(1)}|\theta, \lambda)\Pr(\theta)\Pr(\lambda)d\theta d\lambda$$
$$= \int \left[\int \Pr(Y^{(2)}|\theta; \lambda)\Pr(\theta|Y^{(1)} = y^{(1)}; \lambda)d\theta\right] \Pr(\lambda|Y^{(1)} = y^{(1)})d\lambda$$

If we draw from the probability within brackets, we ignore uncertainty in the item parameters and consider them given.

The main practical problem is to construct an algorithm to produce a sample from $\Pr(\theta, \lambda|Y^{(1)})$. The next step entails generating responses from an IRT model which is quite easy. To produce samples from $\Pr(\theta, \lambda \left| Y^{(1)} \right.)$ a Markov Chain Monte Carlo (MCMC) estimation algorithm for the IRT model can be applied. A number of MCMC estimation algorithms are developed for a wide variety of IRT models including the two parameter logistic model (Patz and Junker, 1999a), the two-parameter normal ogive model (Albert, 1992; Baker, 1998), and the Rasch model (Kim, 2001; Maris and Maris, 2002). These algorithms have been generalized to models with multiple raters, multiple item types and missing data (Patz and Junker, 1999a,b), models with a multi-level structure on the ability parameters (Fox and Glas, 2001), latent class models (Hoijtink and Molenaar, 1997), models with multidimensional latent abilities (Béguin and Glas, 2001), Bock's (1972) nominal response model, mixture item response models (Wollack, Bolt, Cohen and Lee, 2002), the conjunc-

tive Rasch model, the graded response model, the Parella model, and a hierarchical Rasch model (Maris and Maris, 2002). All these estimation procedures have in common that draws from the posterior distribution of the parameters, $\Pr(\theta, \lambda \left| Y^{(1)} \right.)$ are generated during estimation.

6. Discussion

Our aim 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 test in situations where CTT could normally not be applied. We have described a number of applications taken from our daily work 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 construct tests that are parallel in the CTT sense, or Kolen, Zeng, and Hanson (1996) who use IRT to estimate the standard errors of scale scores.

We have considered monotone, unidimensional IRT models but this is not essential. General formulae have been presented here with the aim to facilitate the derivation of reliability, and other CTT concepts using any IRT model. Consider, as an example, a general latent class model where ability is a discrete variable and each value of the ability defines a latent class. With this model, and binary items, test reliability can be shown to be equal to

$$\rho_Y^2 = \frac{\sum_g p_g \left(\sum_i p_{ig}\right)^2 - \left(\sum_g p_g \sum_i p_{ig}\right)^2}{\sum_g p_g \left(\sum_i p_{ig}\right)^2 - \left(\sum_g p_g \sum_i p_{ig}\right)^2 + \sum_g p_g \sum_i p_{ig}(1 - p_{ig})}, \tag{43}$$

where $\sum_g$ sums over latent classes, $\sum_i$ sums over items in the test, p_g denotes the probability of being in class g, and p_{ig} denotes the probability to answer the i-th item correct when one is a member of class g. It is not immediately clear to us what Equation (43) means but it appears to be related to the ability to discriminate

between the classes; that is, $\rho_Y^2 = 1$ if there are one or more items that distinguish perfectly between the latent classes. Here we see an example where IRT offers more specialized indices to judge the quality of the items and the test (e.g., Rost, 1996, pp. 1153-159).

It must be noted that all calculations are predicated on the validity of the IRT model, and the availability of good estimates of the distribution in the population of interest. Further, 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 asses 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, we may get an impression of the sample variance involved in our calculations 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. For example, we calculate the expected test information in a population of interest with $\sigma_\theta = \sigma_\theta^{(\text{low})}$ to get the lower end and then with $\sigma_\theta = \sigma_\theta^{(\text{high})}$ to get the upper end of an approximate 95% interval around the expected information. We choose to vary σ_θ since it is estimated with much less precision than the mean and it is the main determinant of CTT indices. For example, in the analysis discussed in Section 4.1, $SE = 0.149$, which provides the interval: $\rho_Y^2 \in [0.70 - 0.78]$. A more promising method is to use MCMC methods to generate exchangeable replications to obtain estimates of CTT statistics and (a lower bound of) their sampling variance.

7. Appendix: Calculating $\Pr(Y = y|\theta)$

The term within brackets in (24) equals an elementary symmetric function of order y with arguments $\Pr(X_1 = x_1|\theta), ..., \Pr(X_I = x_I|\theta)$ and it can be calculated using any algorithm for the calculation of elementary symmetric functions (e.g., Fischer, 1995, p. 136-137; Verhelst, Glas and Van der Sluis, 1984). Alternatively, $\Pr(Y = y|\theta)$ can be calculated using a slight extension of the recursion formulae that are given by Lord and Wingersky (1984) which we discuss now.

Consider first a number of dichotomous items. Let Y denote the number correct score, and $\Pr_r(Y = y|\theta)$ the probability of number correct score y over the first r items, given ability θ. A moment of thought will allow you to see that $\Pr_2(Y = y|\theta)$ is related to the item scores in the following way:

$$\Pr_2(Y = y|\theta) =$$

$$\begin{cases} \Pr(X_1 = 0|\theta)\Pr(X_2 = 0|\theta) & \text{if } y = 0, \\ \Pr(X_1 = 1|\theta)\Pr(X_2 = 0|\theta) + \Pr(X_1 = 0|\theta)\Pr(X_2 = 1|\theta) & \text{if } y = 1 \\ \Pr(X_1 = 1|\theta)\Pr(X_2 = 1|\theta) & \text{if } y = 2 \end{cases}$$

$$= \begin{cases} \Pr_1(Y = y|\theta)\Pr(X_2 = 0|\theta) & \text{if } y = 0, \\ \Pr_1(Y = y|\theta)\Pr(X_2 = 0|\theta) + \Pr_1(Y = y-1|\theta)\Pr(X_2 = 1|\theta) & \text{if } y = 1 \\ \Pr_1(Y = y-1|\theta)\Pr(X_2 = 1|\theta) & \text{if } y = 2 \end{cases}$$

If we continue to $r = 2$, $r = 3$, etc., we find that in general we may replace $\Pr_2(Y = y|\theta)$ by $\Pr_r(Y = y|\theta)$, $\Pr_1(Y = y|\theta)$ by $\Pr_{r-1}(Y = y|\theta)$, $y = 1$ by $0 < y < r$, and $y = 2$ by $y = r$. Thus we obtain the recursion formulae given by Lord and Wingersky (1984).

We may use to same line of reasoning to derive recursion formulae for polytomous

items. The result is the following: For $1 < r \leq I$,

$$\Pr_r(Y = y|\theta) = \begin{cases} \Pr_{r-1}(y|\theta)(1 - \sum_{h=1}^{J_r} p_{hr}) & \text{if } y = 0 \\ \Pr_{r-1}(y|\theta)(1 - \sum_{h=1}^{J_r} p_{hr}) + \sum_{h=1}^{M} \Pr_{r-1}(y - h|\theta)p_{hr} & \text{if } 0 < y \leq A_{r-1} \\ \sum_{h=\max(1,y-A_{r-1})}^{M} \Pr_{r-1}(y - h|\theta)p_{hr} & \text{if } A_{r-1} < y \leq A_r \end{cases},$$

where $M \equiv \min(J_r, y)$, $A_{r-1} \equiv \sum_{j=1}^{r-1} J_j$, $A_r \equiv \sum_{j=1}^{r} J_j$, $p_{hr} \equiv \Pr(X_r = h|\theta)$, and J_r is the maximum score on item r.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

Baker, F. B. (1998). An investigation of item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement 22,* 153-169.

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

Béguin, A.A. & Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models, *Psychometrika, 66,* 541-562.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37-29-51.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 279-334.

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

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *American Statistician, 49,* 327-335.

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.

Davier, von., 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: Institute for Science Education (IPN).(http://winmira.von-davier.de/)

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

Fischer, G.H. (1995). The linear logistic test model. Chapter 8 in "*Rasch models: Foundations, recent developments, and applications* ", edited by G.H. Fischer and I.W. Molenaar. New-York: Springer.

Fox, J.P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Glas, C.A.W., & Verhelst, N.D. (1995). Tests of fit for polytomous Rasch models. Chapter 18 in "*Rasch models: Foundations, recent developments, and applications* ", edited by G.H. Fischer and I.W. Molenaar. New-York: Springer.

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

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, *62*,171-189.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating.* 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.

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.* New-Jersey: Lawrence Earlbaum Ass.

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

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

Maris, G., & Maris, E. (2002). A MCMC-method for models with continuous latent responses. *Psychometrika, 67*, 335-350.

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: 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.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and*

Behavioral Statistics, 24, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple Item Types, Missing Data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

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

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

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]. Doctoral Dissertation. Arnhem: Cito. (available at http://download/pub/pok/dissertaties/ps-cor.pdf)

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

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.

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

Tanner, M. A. (1993). *Tools for statistical inference.* New-York: Springer-Verlag.

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: Earlbaum.

Verhelst, N.D. (2002a). *Personal Communication.*

Verhelst, N.D. (1998). *Estimating the reliability of a test from a single test administration.* Measurement and Research Reports 98-2, Cito: Arnhem

Verhelst, N. D., & Glas, C.A.W. (1995). *The one parameter logistic Model.* Chapter 12 in "Rasch models: Foundations, recent developments, and applications." Edited by G. H. Fischer and I.W. Molenaar. New-York: Springer.

Verhelst, N. D., Glas, C. A. W., & Van der Sluis, A. (1984). Estimation problems in the Rasch model.: The basis symmetric functions. *Computational Statistics Quarterly, 1*, 245-262.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *One parameter logistic model OPLM.* Computer software manual, Cito: Arnhem. (The program can be obtained for non-commercial purposes via e-mail: norman.verhelst@citogroep.nl)

Verhelst, N.D. & Veldhuijzen, N.H. (2002). *Correlations between latent variables: The programs CORDIM and FMERGE.* R&D Notitie 2002-2, Arnhem: Cito.

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

Verstralen, H.H.F.M. (1997b). *A logistic latent class model for multiple choice items.* Measurement and Research Report, Cito: Arnhem.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, 54, 427-450.

Wollack, J.A., Bold, D.M., Cohen, A.S., & Lee, Y-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*, 339-352.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19*, 231-240.

Yang, G.L., & Chen, M.K. (1978). A note on weighted kappa. *Socio Economic*

Planning Sciences, *12*, 293-294.