Measurement and Research Department Reports 2001–5

The Combined Use of Classical Test Theory and Item Response Theory

H.H.F.M. Verstralen

T.M. Bechger

G.K.J. Maris

Cito groep

Measurement and Research Department Reports 2001–5

The Combined Use of Classical Test Theory and Item Response Theory

H.H.F.M. Verstralen
T.M. Bechger
G.K.J. Maris

Citogroep
Arnhem, September 2001

Abstract

The present paper is about a number of relations between concepts of models from classical test theory (CTT), such as reliability, and item response theory (IRT). It is demonstrated that the use of IRT models allows us to extend the range of applications of CTT, and investigate relations among concepts that are central in CTT such as reliability and item-test correlation.

1 Introduction

The main purpose of this paper is to clarify some aspects of classical (psychometric) test theory (CTT) by combining CTT with item response theory (IRT) modeling. Following, it will be demonstrated that the use of IRT models extends the range of applications of CTT, and allows investigation of concepts that are central in CTT such as reliability and item-test correlation.

In Section 2, a general outline of CTT is presented.[1] Further explanation is given in Section 3 which is about the special case of binary (zero-one) τ-equivalent items. In Sections 4 and 5, some of the relations between CTT and the generalized partial credit model (GPCM) of IRT are discussed. Section 4 is about Fisher's information and its relation to the reliability of estimated abilities. In Section 5, we manipulate characteristics of the GPCM to investigate the effect on reliability or difficulty. We investigate, for instance, whether, under typical circumstances, items with high discrimination parameters are more reliable. In Section 6, we describe two possible applications. In the first application, the use of an IRT model allows the evaluation of classical test reliability of a test that has never been administered to (a sample from) the population of interest. The second application entails the calculation of the correlation between latent variables measured by different tests. Finally, in Section 7 we discuss our findings.

[1] A more extensive discussion of CTT is found in the classical work by Lord and Novick (1968).

2 Classical Test Theory

CTT distinguishes two random experiments: (1) sampling a respondent, and (2) sampling a score within a given respondent. Consider the second experiment first; that is, imagine that we could repeatedly administer an item to a generic subject v in such a way that the answers were independent. One can, for instance, imagine that the subject is "brainwashed" at each administration and forgets having seen the item before.

We assume that our subject's behavior on the test is determined by his value on a vector variable θ that we refer to as "ability." The *true score,* $\tau_i(\theta_v)$, of any person with ability $\theta = \theta_v$ is defined as the expectation $E_v[X_i|\theta = \theta_v]$ of his intra-individual distribution. Let X_{is} denote the score on item i on testing occasion s. We may always write

$$X_{is} \equiv \tau_i(\theta_v) + \epsilon_{is}, \tag{1}$$

where the deviations $\epsilon_{is} \equiv X_{is} - E_v[X_i|\theta = \theta_v]$ represent random measurement error. While the measurement error varies across repeated administrations of an item to a person, the true score is a fixed parameter characterizing the combination of a person and an item. The intra-individual distribution of the measurement errors has zero mean and variance $\sigma^2_{\epsilon_i}(\theta)$.

If we combine the two random experiments, the true score becomes a random variable $E_v[X_i|\theta]$, i.e., the regression of X_i on θ, with values $E_v[X_i|\theta = \theta_v]$. The distribution of the measurement error variables is now a mixture of the individual persons' error distributions.

We assume that items generate discrete measurements, i.e., $X_i \in \{0, ..., M_i\}$, where a higher score is associated with better performance. The *Category*

Response Function (CRF) gives the probability of obtaining a score k, as a function of ability. That is,

$$P_{ik}(\theta) \equiv \Pr(X_i = k|\theta),\ k \in \{0, ..., M_i\}. \tag{2}$$

In practice, we would choose a fitting IRT model (e.g., Equation 26) but at this point we assume no particular functional form for the CRFs. We do assume that all CRFs depend on the same ability, i.e., that different items measure one or more aspects of the same ability.

It follows that $\tau_i(\theta) = \sum_{k=1}^{M_i} kP_{ik}(\theta)$. Hence, the true score is a transformation of the ability. The conditional variance of X_i is equal to

$$\sigma^2_{X_i}(\theta) = \sigma^2_{\epsilon_i}(\theta) = m_i(\theta) - (\tau_i(\theta))^2, \tag{3}$$

where $m_i(\theta) \equiv E_v[X_i^2|\theta] = \sum_{k=1}^{M_i} k^2 P_{ik}(\theta)$, and we assume that $0 < \sigma^2_{X_i}(\theta) < \infty$.

The population of interest will be referred to as *the reference population.* It is assumed that the IRT model holds in the reference population. Integrating $\tau_i(\theta)$ over the distribution of θ in the reference population gives us $\tau_i \equiv E[\tau_i(\theta)] = E[X_i]$, the expected response to item i. The expectation over the reference population is denoted by $E[.]$. The *difficulty* (or expected p-value) of item i is defined as $\pi_i \equiv \frac{\tau_i}{M_i}$. The variance of the observed score in the reference population is

$$\sigma^2_{X_i} = m_i - \tau_i^2,\ (0 < \sigma^2_{X_i} < \infty) \tag{4}$$

where $m_i \equiv E[m_i(\theta)] = E[X_i^2]$. The true-score variance is given by

$$\sigma^2_{\tau_i} = E[(\tau_i(\theta))^2] - \tau_i^2,\ (0 < \sigma^2_{\tau_i} < \infty). \tag{5}$$

The *reliability* of item i in the reference population is defined as the proportion of true variation; that is,

$$\rho^2_{X_i} \equiv \frac{\sigma^2_{\tau_i}}{\sigma^2_{X_i}}. \tag{6}$$

Item reliability indicates the precision with which differences in true score between persons are estimated by differences in observed item scores between individuals. An alternative expression for the item reliability is

$$\rho^2_{X_i} = 1 - \frac{E[\sigma^2_{X_i}(\theta)]}{\sigma^2_{X_i}}, \tag{7}$$

where $E[\sigma^2_{X_i}(\theta)] = m_i - E[(\tau_i(\theta))^2]$. Formula 7 reveals that the reliability of an item depends on the ratio of expected conditional variance and variance in the reference population. A third alternative expression will be given in the next section.

Now, consider a test which consists of $I > 1$ items. We will refer to this test as *the proposed test*. As a test score we consider a linear combination $Y \equiv \sum_{i=1}^{I} \omega_i X_i$ of the item responses, where the w_i are constant weights. The true score on the proposed test is defined as:

$$\tau(\theta) \equiv E_v[Y|\theta = \theta_v] = \sum_{i=1}^{I} \omega_i \tau_i(\theta). \tag{8}$$

Measurement error on the test is defined as $\epsilon \equiv Y - \tau(\theta) = \sum_{i=1}^{I} \omega_i \epsilon_i$. The expected test score $\tau \equiv \sum_{i=1}^{I} \omega_i \tau_i$. Its variance equals

$$\sigma^2_\tau \equiv E[\tau^2(\theta)] - \tau^2. \tag{9}$$

We assume that the measurement errors on different items are independent given θ, so that the error variance of the test score is given by

$$\sigma^2_\epsilon \equiv \sum_{i=1}^{I} \omega_i^2 \sigma^2_{\epsilon_i} = \sum_{i=1}^{I} \omega_i^2 (m_i - E[\tau_i^2(\theta)]). \tag{10}$$

The variance of the score σ_Y^2 is given by $\sigma_\tau^2 + \sigma_\epsilon^2$ and the reliability of the test score in the reference population is given by

$$\rho_Y^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2}. \tag{11}$$

Test reliability is of interest because its square root, ρ_Y, provides an upper bound to the validity of the test with respect to any criterion, i.e., the correlation of the test score with any criterion (see Lord & Novick, 1968, p. 72).

Test reliability can also be shown to equal the square of the correlation between Y and τ (see Lord & Novick, 1968, p. 57). The correlation between Y and τ is not equal to the correlation between Y and θ unless the latter is a linear transformation of τ as, for example, in the *Binomial model* where θ is a scalar random variable, $M = 1$, $P_{i1}(\theta) = \theta$, and $P_{i0}(\theta) = 1 - \theta$ (e.g., 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 it is therefore more appropriate to calculate the reliability of the estimated ability values $\hat{\theta}$. To derive this reliability we first note that

$$\hat{\theta}_v = E[\hat{\theta}|\theta = \theta_v] + e, \tag{12}$$

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. Since reliability is defined as the proportion of true-score variance in the reference population we find that

$$\rho_{\hat{\theta}}^2 = 1 - \frac{E[Var(\hat{\theta}|\theta)]}{\sigma_\theta^2 + E[Var(\hat{\theta}|\theta)]}, \tag{13}$$

where $Var(\hat{\theta}|\theta)$ denotes the variance of the estimated values given the true values of θ. This equals the squared correlation between θ and $\hat{\theta}$ if $E[\hat{\theta}|\theta] =$

$\alpha_1\theta + \alpha_2$, where, $\alpha_1, \alpha_1 \in \mathbb{R}^2$, so that the covariance between $\hat{\theta}$ and θ equals the variance of θ, σ^2_θ.

The reliability of $\hat{\theta}$ shows a similar relation to the validity of the test as ρ^2_Y. To be more specific, if $\rho(\hat{\theta}, \hat{\xi})$ denotes the correlation between the estimates of θ and estimates of some other latent trait ξ, and both estimates are unbiased

$$\rho(\hat{\theta}, \hat{\xi}) = \rho(\theta, \xi)\sqrt{\rho^2_{\hat{\theta}}\rho^2_{\hat{\xi}}}, \tag{14}$$

where $\rho(\theta, \xi)$ denotes *the disattenuated validity*, i.e., the correlation between θ and ξ. If θ and ξ are abilities measured by different subscales of a test, $\rho(\theta, \xi)$ will help to decide whether both subtests should be combined. In Section 5.3, we discuss the estimation of $\rho(\theta, \xi)$.

Another important item property in CTT is the *Item-Total Correlation (ITC)*; the correlation of the score on item i with the score on the proposed test, including the item.

$$\begin{aligned} ITC_i &= \frac{Cov(\tau, \tau_i) + Cov(\epsilon, \epsilon_i)}{\sqrt{\sigma^2_Y \sigma^2_{X_i}}} \\ &= \frac{E[\tau(\theta)\tau_i(\theta)] - \tau\tau_i + \sigma^2_{\epsilon_i}}{\sqrt{\sigma^2_Y \sigma^2_{X_i}}}. \end{aligned} \tag{15}$$

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. 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, we calculate the *item rest correlation (IRC)*; the correlation between the score on an item and the total

score on the proposed test, excluding the item. First, the term $\sigma^2_{\epsilon_i}$ can be deleted from (15). Let $ITC_i^{(-i)}$ denote the ITC_i with $\sigma^2_{\epsilon_i} = 0$. If we apply the correction from Guilford and Fruchter (1978, p. 449, Equation 18.7) for decreasing the length of the test by one, we find that the IRC of item i is estimated by

$$IRC_i = \frac{ITC_i^{(-i)}}{\sqrt{\frac{1-\rho_Y^2}{\rho_Y^2/I} + \rho_Y^2}}, \tag{16}$$

where ρ_Y^2 denotes the reliability of the proposed test.

Consider an arbitrary test with I_R items, excluding the item of interest. To distinguish this test from the proposed test we call it *the reference test.* The correlation between item i and the score on the reference test is given by $ITC^{(-i)}$, which may be interpreted as a measure of the "fit " of item i to the reference test. An application motivating our distinction between a proposed test and a reference test is discussed in Section 6.1.

3 The Case of Binary, τ-Equivalent Items

In this section, we make four simplifying assumptions. First, we assume that all items are binary with $X_i = 1$ if the answer is correct, and $X_i = 0$ otherwise. Second, we assume that all CRFs are equal, that is, $P_{i1}(\theta) = P(\theta)$, and $P_{i0}(\theta) = 1 - P(\theta)$ for all items. Third, the CRFs are appropriately modelled by a *Generalized Partial Credit model (GPCM)*

$$P(\theta; \alpha, \delta) = \frac{\exp(\alpha(\theta - \delta))}{1 + \exp(\alpha(\theta - \delta))}, \tag{17}$$

where the parameters $\alpha, \delta \in \mathbb{R}^2$ are considered known. The ability θ is a real scalar random variable and the items are said to be *unidimensional.*

For the purpose of illustration we have drawn the CRFs and the true score for a GPCM in Figure 1. Note that the category parameter δ is the value of θ where $P(\theta;\alpha,\delta) = 1 - P(\theta;\alpha,\delta) = 0.5$. Fourth, we assume that the

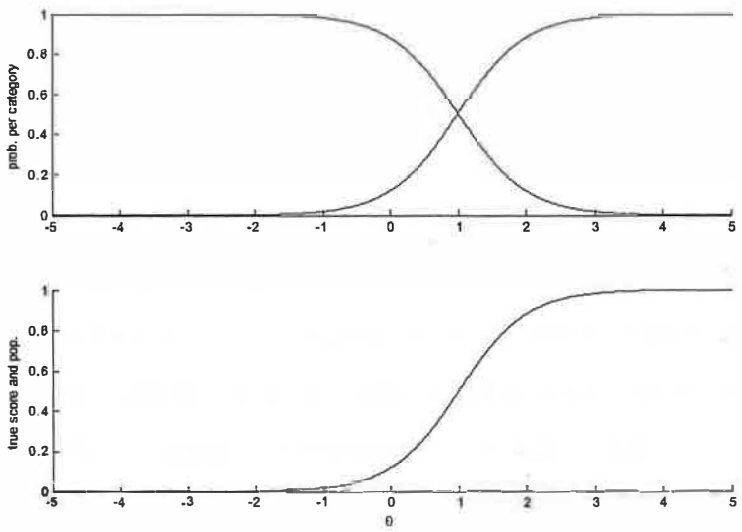

Figure 1: The first figure shows CRFs for a GPCM item with $\delta = 1$, and $\alpha = 2$. The second figure shows the true score as a function of theta.

distribution of θ in the reference population is normal and that we know its mean and variance. This would be a reasonable assumption when, for instance, we have obtained estimates of these quantities using a large sample from the reference population. More information on the GPCM is presented in the next section.

The true score equals the probability of a correct response, i.e., $\tau_i(\theta) = P(\theta;\alpha,\delta)$ (see Figure 1). Since the true-scores of the items are all equal, the items are said to be *τ-equivalent.* The conditional variance of the score for

each item is equal to

$$\sigma_X^2(\theta) = P(\theta;\alpha,\delta)(1 - P(\theta;\alpha,\delta)). \tag{18}$$

Under the GPCM, the derivative of the true score with respect to θ equals $\alpha\sigma_X^2(\theta)$. The α parameter may therefore be interpreted as a discrimination parameter. It is usually required that $\alpha > 0$, so that the true-score increases with ability.

Using the formulae in the previous section we find that

$$\pi = E[P(\theta;\alpha,\delta)], \sigma_X^2 = \pi(1-\pi), \tag{19}$$

and the true-score variance for any item is given by

$$E[(P(\theta;\alpha,\delta))^2] - \pi^2 = \sigma_P^2, \tag{20}$$

where σ_P^2 denotes the variance of the proportions correct (or p-values) in the reference population. We conclude that binary, τ-equivalent items have equal observed variances and equal true score variances. It follows that such items have the same measurement error variance, $\sigma_\epsilon^2 = \pi(1-\pi) - \sigma_P^2$, and the same reliability: $\rho_X^2 = \sigma_P^2/\sigma_X^2$.

Reliability depends on the population and on the CRFs. When $\alpha \to 0$, given δ, $\sigma_P^2 \to 0$, and reliability becomes zero. If $\alpha \to \infty$, $P(\theta;\alpha,\delta) = X_i$, so that $\sigma_P^2 = \sigma_X^2$ and $\rho_X^2 = 1$. To illustrate the effect of increasing α we have drawn the CRF and the true score of an item with a very large α in Figure 2. Figure 2 may be compared to Figure 1 to see the effect of increasing the value of α on the CRFs.

Figure 2 illustrates that, if α becomes very large, the item reliability is perfect although we cannot distinguish between subjects with ability values

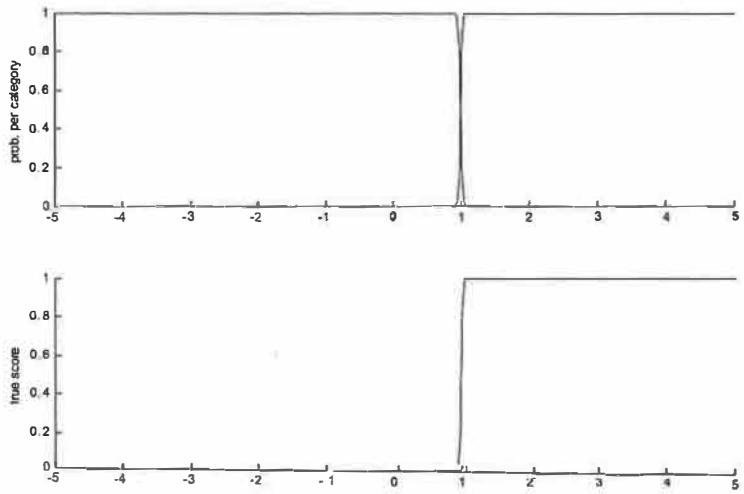

Figure 2: Plots of CRFs and true score against θ for a GPCM item with parameters $\delta = 1$ and $\alpha = 80$.

on the same side of $\theta = \delta$ (see Loevinger, 1954 or Lord & Novick, 1968, p. 465). It appears that high item reliability is not always desirable when the purpose of the study is to estimate abilities. An item such as the one drawn in Figure 2 would, however, be most useful if we wish to classify subjects into two groups, i.e., those with abilities over δ and those with abilities below δ, but worthless if groups were defined otherwise. It should be noted, however, that items where α_i is so high that $X_i \approx P_{i1}(\theta)$ are extremely rare in practice.

In CTT, τ-equivalent items are said to be *parallel* if their error terms are uncorrelated. The covariance between the scores on two parallel items

$(i \neq j)$ is given by

$$\begin{aligned} \sigma_{X_i X_j} &= E[X_i(\theta)X_j(\theta)] - E[X_i(\theta)]E[X_j(\theta)] \\ &= E[(\tau_i(\theta) + \epsilon_i(\theta))(\tau_i(\theta) + \epsilon_j(\theta))] - E[X_i(\theta)]E[X_j(\theta)] \\ &= E[\tau_i^2(\theta)] - E[X_i(\theta)]E[X_j(\theta)] \\ &= \sigma_{\tau_i}^2 = \sigma_P^2. \end{aligned} \tag{21}$$

Furthermore,

$$\frac{\sigma_{X_i X_j}}{\sqrt{\sigma_{X_i}^2 \sigma_{X_i}^2}} = \frac{\sigma_P^2}{\sigma_X^2} = \rho_X^2. \tag{22}$$

Hence, in any population, the covariances of scores on parallel items are equal and positive, and the item reliability equals the correlation between any two parallel items.

The expected unweighted score on a proposed test with I τ-equivalent items, given θ, equals $\tau(\theta) = IP(\theta; \alpha, \delta)$. The expected score on the test equals $E[Y] = \tau = I\pi$, with variance $\sigma_\tau^2 = I^2\sigma_P^2$. If the items are parallel, the error variance on the test is given by $\sum_{i=1}^{I} \sigma_\epsilon^2 = I(\sigma_X^2 - \sigma_P^2)$. It follows that

$$\sigma_Y^2 = I^2\sigma_P^2 + I(\sigma_X^2 - \sigma_P^2). \tag{23}$$

The reliability of the test as a function of the reliability of the items is

$$\begin{aligned} \rho_Y^2 &= \frac{I^2\sigma_P^2}{I^2\sigma_P^2 + I(\sigma_X^2 - \sigma_P^2)} \\ &= \frac{I^2\frac{\sigma_P^2}{\sigma_X^2}}{I^2\frac{\sigma_P^2}{\sigma_X^2} + I - I\frac{\sigma_P^2}{\sigma_X^2}} \\ &= \frac{I\rho_X^2}{(I-1)\rho_X^2 + 1}. \end{aligned} \tag{24}$$

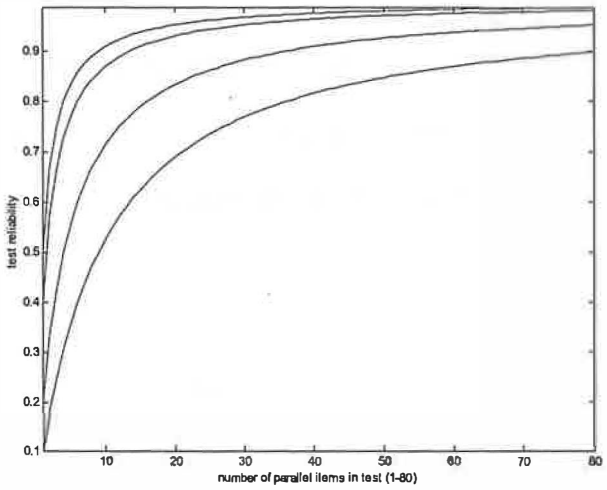

Figure 3: Test reliability plotted against the number of parallel items in the test.

This equation is known as the *Spearman-Brown (SB) formula.* Using the SB formula, the reliability of an aggregated measurement consisting of the sum or average of I parallel measurements of the same persons can be computed. If $I = 1$, for instance, $\rho_Y^2 = \rho_X^2$ the reliability of a single item. For $I = I^* + Z$, we obtain the reliability of a test with I^* items when it is lengthened by adding Z parallel items. The SB formula shows that the reliability of the test-score goes to 1 if I becomes large. As can be seen in Figure 3, however, there is usually a point beyond which adding further parallel items to the test makes little difference for the reliability of the test score.

A bit of algebra shows that in the present circumstances Equation 15

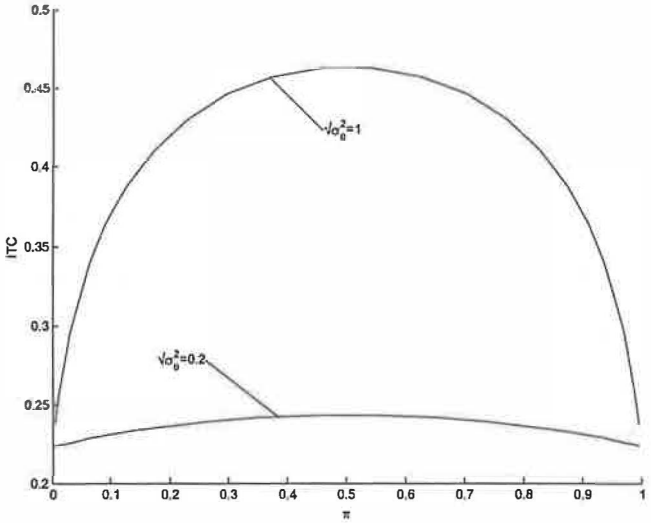

Figure 4: Relation between item difficulty π and the item test correlation assuming 20 parallel GPCM items ($\alpha = 1$).

simplifies to

$$ITC_i = \frac{\sqrt{(I-1)\sigma_p^2 + \pi(1-\pi)}}{\sqrt{I}\sqrt{\pi(1-\pi)}}. \tag{25}$$

Using formula (19) we see that $\lim_{I\to\infty} ITC_i = \sqrt{\rho_X^2}$, which provides an interpretation for ρ_X^2. A plot of the ITC_i against π (for two values of σ_θ^2) in Figure 4 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 σ_θ^2, and on the number of items in the test. One should therefore be careful to give rules-of-thumb for the selection of items based on the ITC (see e.g. Ebel & Frisbie, 1986).

To obtain an idea of the discrimination power of a particular item it would

be more appropriate to look at a plot of the percentage correct responses to this item against the score on the proposed test. To this aim, the score must be divided into intervals (score groups) to make sure that enough observations are made on the item in each score group (Verstralen, 1989). We may take any other variable besides the test score (e.g., income) to see if the item discriminates between levels of this variable or use these plots to investigate differential item functioning (DIF). We frequently use such figures in our daily work and options to produce them are incorporated in our software for CTT (Heuvelmans, 2001).

4 Fisher Information and Reliability

In the previous section we introduced the GPCM for binary items. For polytomous items, the GPCM states that

$$P_{ik}(\theta;\alpha_i,\boldsymbol{\delta}_i) = \frac{1}{D_i}\exp(\alpha_i\sum_{p=1}^{k}(\theta-\delta_{ip})), \tag{26}$$

where $\sum_{p=1}^{0}(\theta-\delta_{ip}) \equiv 0$, $\boldsymbol{\delta}_i = (\delta_{i1},..,\delta_{iM_i})$, and D_i is a constant that is added to make sure that $\sum_{k=0}^{M_i} P_{ik}(\theta) = 1$. For illustration purposes, we have drawn the CRFs of an GPCM item with three categories in Figure 5. Figure 5 also shows a plot of the true score as a function of θ, and the Fisher information function of the item which will be defined below.

The *category parameters*, δ_{ip}, are the values of θ where the CRFs of adjacent categories are equal. When the category parameters are ordered (i.e., $\delta_{i0} < ... < \delta_{iM_i}$) there are segments, $S_k = (\delta_k, \delta_{k+1})$, on the range of θ where category k is more likely than any other category, and these segments

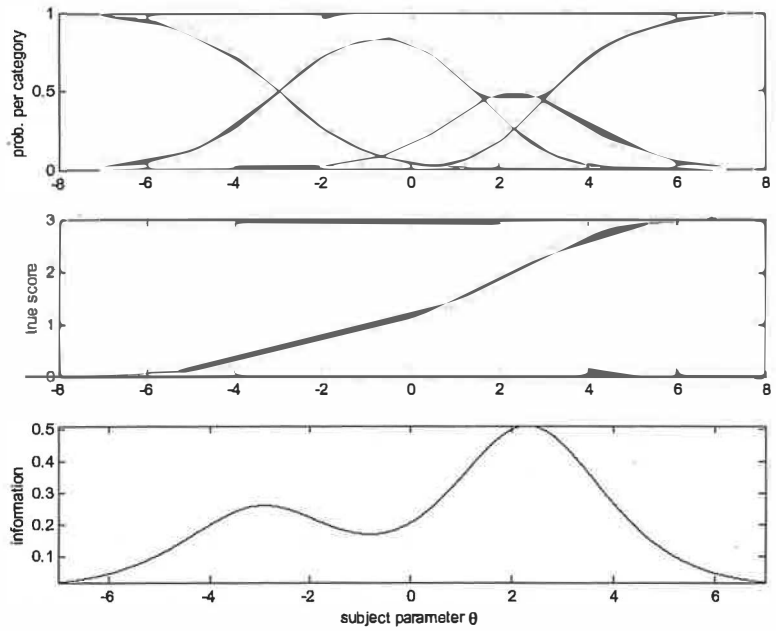

Figure 5: CRFs, true score and information for an GPCM item with three categories; $\boldsymbol{\delta}_i = (-3, 1.67, 3)$, and $\alpha_i = 1$.

are arranged conform the ordering of the category parameters. In this case, $\delta_{k+1} - \delta_k$ is called the *length of category* k.

In general, when some CRFs are rising at a certain θ, others must be falling since $\sum_{k=0}^{M_i} \frac{\partial}{\partial\theta} P_{ik}(\theta) = \frac{\partial}{\partial\theta} \sum_{k=0}^{M_i} P_{ik}(\theta) = 0$. Figure 5 shows two additional aspects that are typical of the GPCM (see Appendix): (1) $P_{i0}(\theta)$ is non-increasing, and $P_{iM_i}(\theta)$ is non-decreasing in θ, and (2) the CRFs for categories $1, ..., (M_i - 1)$ are always bell-shaped, but not necessarily symmetric. It can be shown that the CRFs are not exclusively related to the corresponding categories, which impedes substantive interpretation of their

location and shape.

Let $L(\theta|X_i = k) = P_{ik}(\theta)$ denote the likelihood function of θ given the observed response $X_i = k$. The *item (Fisher) information function* is defined as

$$\begin{aligned} I_i(\theta) &\equiv E_v\left[\left(\frac{\partial}{\partial\theta}\ln L(\theta|X_i)\right)^2|\theta\right] \\ &= P_{i0}(\theta)\left(\frac{\partial}{\partial\theta}\ln P_{i0}(\theta)\right)^2 + \cdots + P_{iM_i}(\theta)\left(\frac{\partial}{\partial\theta}\ln P_{iM_i}(\theta)\right)^2 \\ &= \sum_{k=0}^{M_i}\frac{\left[\frac{\partial}{\partial\theta}P_{ik}(\theta)\right]^2}{P_{ik}(\theta)}. \end{aligned} \tag{27}$$

This shows that, in general, the item information depends on the combined rate of change in the CRFs. Using Equation 33 in the Appendix, the item information function of the GPCM is found to be equal to

$$\begin{aligned} I_i(\theta) &= \alpha_i^2\sum_{k=0}^{M_i}P_{ik}(\theta;\alpha_i,\boldsymbol{\delta}_i)\left(k - E_v[X_i|\theta]\right)^2 \\ &= \alpha_i^2\left(E_v[X_i^2|\theta] - (E_v[X_i|\theta])^2\right) \\ &= \alpha_i^2\sigma_{X_i}^2(\theta) = \alpha_i^2\sigma_{\epsilon_i}^2(\theta). \end{aligned}$$

Thus, the item information is proportional to the conditional (error) variance.

When the item responses are independent given θ, which is a standard assumption in IRT, *the test information function* is the sum of the item information functions, i.e.,

$$I(\theta) \equiv \sum_{i=1}^{I} I_i(\theta), \tag{28}$$

and $E[I(\theta)] = \sum_{i=1}^{I} E[I_i(\theta)]$. Thus, items with high (expected) information will yield a test with high (expected) test information.

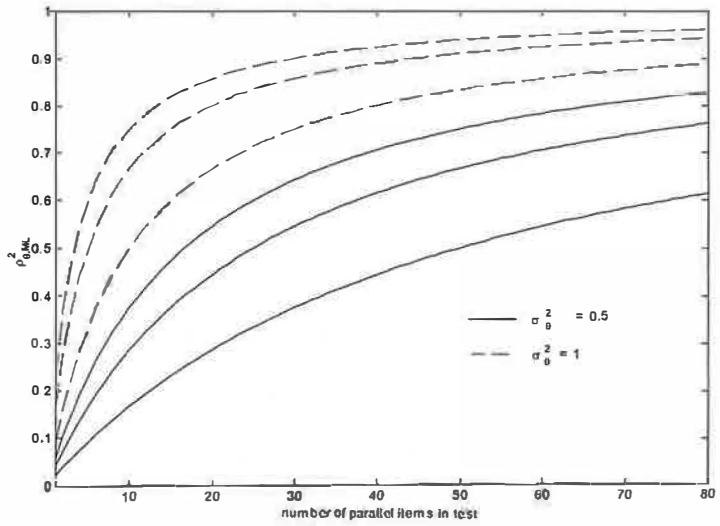

Figure 6: Plot of $\rho^2_{\hat{\theta},ML}$ against number of parallel items in a test. The curves are drawn for situations that differ by the expected information of the items and the homogeneity of the population. As indicated in the legenda, σ^2_θ largely determines how many items are required.

Information is an important concept in IRT because it is related to the precision with which we can estimate abilities in the reference population. When abilities are estimated with the method of Maximum Likelihood (ML), the asymptotic (in the sense of many items) variance of $\hat{\theta}$, given θ, equals $I^{-1}(\theta)$. It follows from Equation 13 that

$$\rho^2_{\hat{\theta},ML} \approx \frac{\sigma^2_\theta}{\sigma^2_\theta + 1/E[I(\theta)]}. \tag{29}$$

Given σ^2_θ, the reliability depends exclusively on $E[I(\theta)]$, which is additive in the expected information per item. Thus, if one reports ML estimates of the abilities, it is desirable to have high expected test information in the reference

population. This is also true for some of the alternative ability estimators, although $Var(\hat{\theta}|\theta)$ may differ from that of the ML-estimator. Note that the ML-estimator is not unbiased so that $\rho^2_{\hat{\theta},ML}$ can only approximate the square of the correlation between $\hat{\theta}$ and θ.

Figure 6 plots $\rho^2_{\hat{\theta},ML}$ against the number of parallel items in a test. We see that $\rho^2_{\hat{\theta},ML}$ shows a pattern that is similar to ρ^2_Y (see Figure 3). Figure 6 also shows that the number of items needed to achieve a certain level of reliability is dependent upon the homogeneity of the population, i.e., more items are needed when the population is more homogeneous.

As was noted before, there are situations where the expected information will be very low while the reliability is high. The relation between reliability, expected information and the GPCM is the subject of the next section.

5 Exploring the Dependency of Reliability and Expected (Fisher) Information on the Parameters of the GPCM

5.1 Introduction

It is quite difficult to describe precisely how properties such as reliability, expected information, etc. relate to the parameters of the postulated IRT model, or to the distribution of θ in the reference population. The situation is aggravated when the expectations do not exist in closed form, which is the case, e.g. when θ is distributed normally. It is not surprising, therefore, that little is known about such relations.

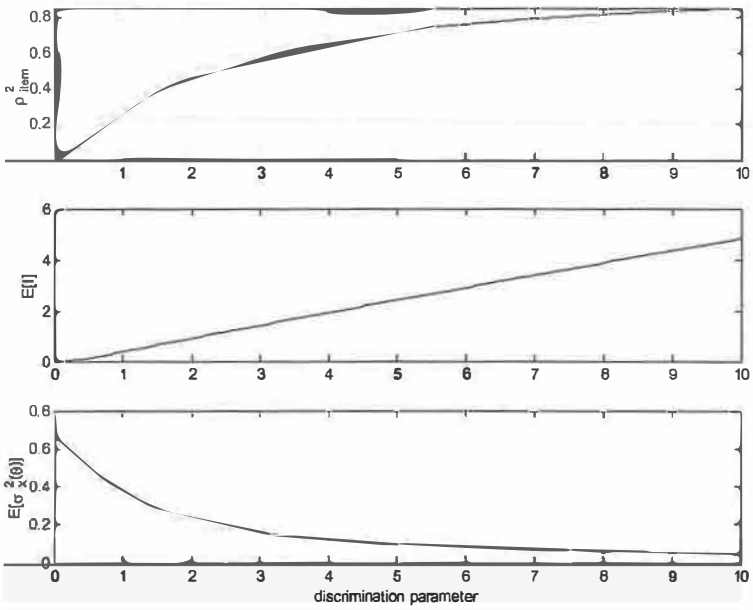

Figure 7: Plot of $\rho^2_{X_i}$, $E[I_i(\theta)]$, and $E[\sigma^2_{X_i}(\theta)]$ against the discrimination parameter α_i. $\boldsymbol{\delta} = (-1, 1)$. The population is normal with $E[\theta] = 0$, and $\sigma^2_\theta = 1$.

In this section, we demonstrate how the formulae in Section 2 enable graphical exploration of the dependencies between different properties of items using standard numerical integration techniques to calculate expectations. Graphical exploration is not nearly as good as solid mathematical proof but it may nevertheless be instrumental in suggesting directions for research.

To be more specific, we will investigate the validity of two widely held convictions:

1. Both item-reliability and expected item information increase when α_i increases.

2. It is best to select items with π values near 0.50.

We are especially interested to know what happens in "the typical situation." The typical situation has the following characteristics: (i) $0 < \alpha_i < 10$. (ii) The category parameters are ordered from small to large. (iii) The maximum absolute difference between the category parameters lies between 0.5 and 6. (iv) The items are neither very difficult nor very easy for the reference population. (v) The reference population is neither very homogeneous nor very heterogeneous compared to the maximum absolute difference between the category parameters.

5.2 Are Items with Large Discrimination Parameters Better?

In the typical situation, we find that items with higher α values (all other parameters constant) are indeed more reliable, and more informative. Figure 7 portrays a typical item. Although $E[\sigma^2_{X_i}(\theta)]$ decreases with α_i, α_i^2 increases more rapidly so that $E[I_i(\theta)] = \alpha_i^2 E[\sigma^2_{X_i}(\theta)]$ increases with α_i.

The situation may change radically when the item is either very difficult or very easy for the reference population. This is illustrated with Figures 8, and 9. While reliability continues to increase with α, the expected item information decreases and beyond a certain value of α, the item contributes virtually nothing to the expected test information.

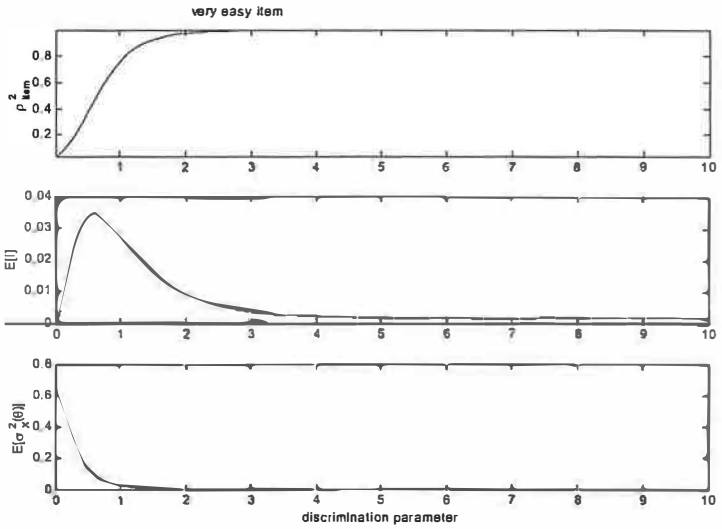

Figure 8: Plots of ρ^2_{item}, $E[I_i(\theta)]$, and $E[\sigma^2_{X_i}(\theta)]$ against the discrimination parameter α. $\boldsymbol{\delta} = (0, 2)$. The population is normal with mean 6 and variance 1.

The idea that higher α is associated with better items is also found wrong in some cases where the population is very homogeneous (i.e., σ^2_θ is small). Figures 13 to 15 in the Appendix illustrate some of the possibilities. Summarizing, we conclude that items with large discrimination parameters are usually better.

5.3 Is 0.50 the Ideal Item Difficulty ?

Under typical circumstances both the expected information and the reliability are maximized when π-values are within the interval $[0.30 - 0.70]$, everything else constant (see Figure 10). Thus, items with difficulties around 0.50

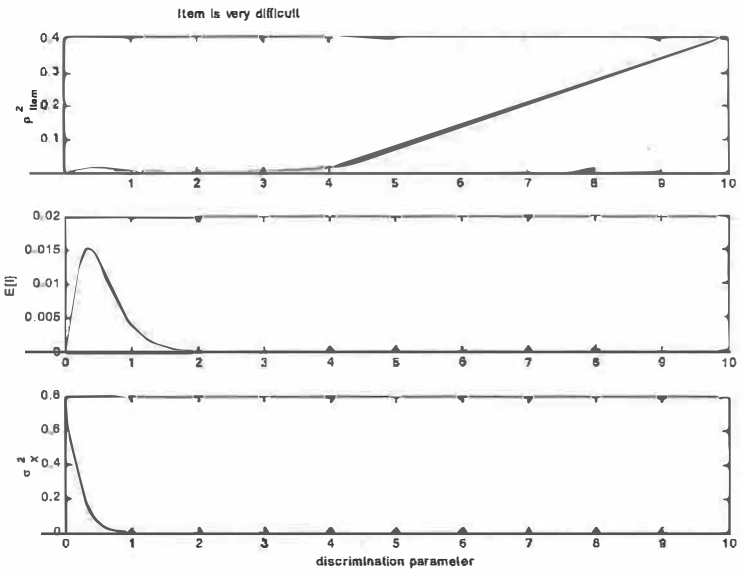

Figure 9: Plots of ρ^2_{item}, $E[I_i(\theta)]$, and $\sigma^2_{X_i}$ against the α-parameter. $\boldsymbol{\delta} = (0, 2)$. The population is normal with mean -6 and variance 1.

are usually preferably.

Divergent patterns are illustrated in Figure 11, and Figure 16 in the Appendix. Figure 16 shows a pattern which we found to occur in a situation where the category parameters are not ordered from small to large. Yet another interesting pattern is shown in Figure 11. Here, both the reliability and the expected information of the item can be zero while neither π nor the α-parameter are unusual. What we see in Figure 11 can be explained by the fact that it requires little ability to score in the second category and much ability to score in the third category (i.e., category 1 is quite long). When π is near 0.5, the bulk of the population lies within the segment S_1 (see Figure 17 in the Appendix). This example is a bit contrived but serves

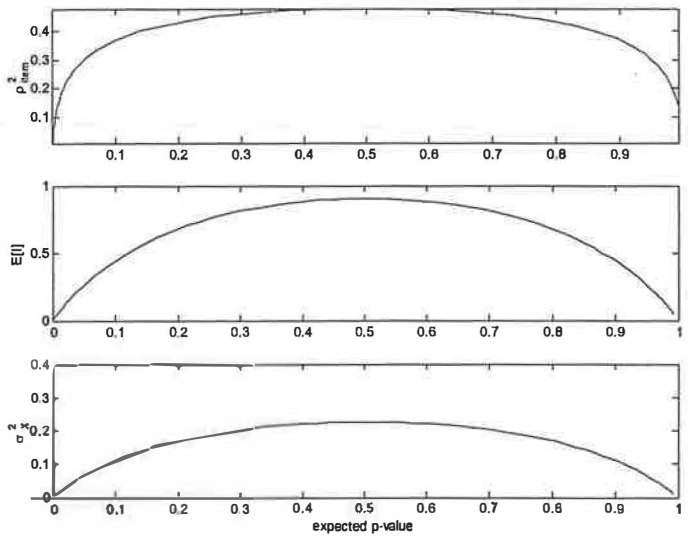

Figure 10: Plots of ρ^2_{item}, $E[I_i(\theta)]$, and $\sigma^2_{X_i}$ against π. $\boldsymbol{\delta} = (0, 1)$, and $\alpha = 2$. The population is normal with variance 1.

to demonstrate that it is not always true that items with difficulties around 0.50 are to be preferred.

Lord (1952) used simulation to investigate the relation between π and test reliability. His result suggests that the optimal value of π lies between 0.70 and 0.85. It is clear that his results depend on the IRT model and the distribution of θ used to simulate the data. His recommendations regarding the "optimal value of π" are neither more valid nor less valid than recommendations by other authors (e.g., Crocker & Algina, 1986; Feldt, 1993), except when the IRT model he used is more appropriate.

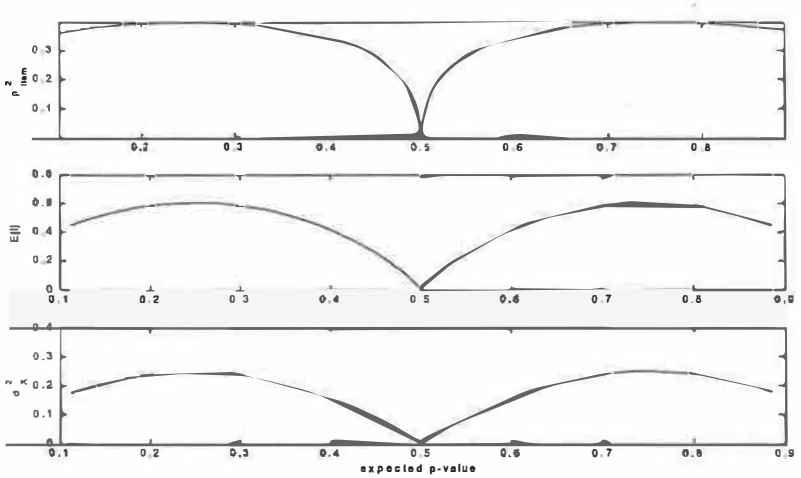

Figure 11: $\rho^2_{X_i}$, $E[I_i(\theta)]$, and $\sigma^2_{X_i}$ as a function of π. $\alpha_i = 2$, and $\boldsymbol{\delta} = (-6, 6)$.

6 Applications

6.1 Selecting Items from a Pilot Test

The state examination of Dutch as a second language (hereafter abbreviated to "the Stex ") is a large-scale examination of the ability to use the Dutch language in practical situations; sometimes called "functional literacy". Four aspects of functional literacy; listening, speaking, writing, and reading, are examined separately. 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. We will briefly discuss how the fomulae in Section 2 are used in the Stex to guide the construction of a new examination.

In the Stex, the construction of a new examination is preceded by a pilot

study which entails the administration of new items to people that follow a Dutch language course. The purpose of this pilot study is to select those items that are best to figure in the coming examination.

After the data from the pilot study 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* for later reference.

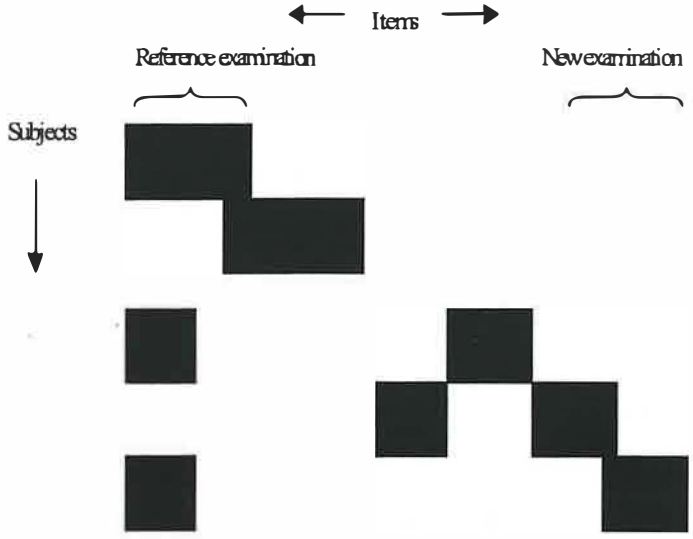

Figure 12: Schematic representation of the data bank.

Schematically, the data bank can be represented (as in Figure 12) by a matrix where the rows are subjects and the columns are items. In Figure 12, 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 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

for more items to be tested.

The reference examination is a subset of the items in the data bank. This reference examination was chosen by the examination committee and is believed to measure the ability of interest. The design of the pilot study is chosen to ensure that all items in the pilot study can be linked to one another, and to the reference examination via items that are common to one or more booklets. This is a technical condition which allows us to fit a GPCM to the item responses, and place the items on the same scale as the items in the reference examination. We also obtain an estimate of the distribution of the latent ability in the population of examinees. This population is the *reference population.*

The first and most important step in the analysis of the pilot data is to establish a fitting GPCM using all relevant parts of the data bank. All subjects that are included in the analysis are assumed to be drawn from the reference population. Although the samples that participate in the pilot testing are generally less able, they are assumed to differ only in their ability distribution, which ensures that the parameters of the GPCM apply to them (e.g., Steyer & Eid, 1993, par. 18.4).

In the Stex, we first use a special purpose program to estimate the parameters of the GPCM (see Verstralen, 1996a,1996b). We then scale the α's and round them to the nearest integer, consider them as fixed, and use the OPLM program (Verhelst, Glas & Verstralen, 1995) to reestimate the category parameters, and establish the fit of the model. To this aim, OPLM produces several "goodness-of-fit " statistics (see Glas & Verhelst, 1995). Items that do not conform to the model, and/or items with negative α-parameters are

discarded because the true score is expected to be increasing in θ.

Following the calibration of the data from the pilot examination, we provide the examination authorities with three 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, which will normally achieve optimal expected information, as we have seen in the previous paragraph. Second, we report expected item information and we recommend to discard those items that have the lowest values. This seems appropriate since estimated abilities will be reported as examination marks and expected item information is positively related to the reliability of the estimated abilities. Thirdly, we provide IRC_is since the items should fit the reference examination. With this information, the item writers then compose a proposed examination. The item difficulties are used to place more difficult items at the end of the examination booklets in order not to discourage candidates. When a proposed examination has been constructed, we provide an estimate of the reliability of the estimated abilities. Since the reliability provides a lower bound to the validity it should be sufficiently high.

It could be argued that the Stex involves a classification problem with two classes; candidates below or above a cutting point θ_0. The cutting point is determined by the examination committee who decides which score is required to pass the reference examination. Spray and Reckase (1994) have argued that one should select items with highest information at θ_0. The information of the items at θ_0 need not to correlate perfectly with their expected information and information at θ_0 is an additional selection criterion. Eggen (1999) describes an elaborate item selection method which gives similar results.

In the Stex, the item difficulties have been reported to the item writers for some years now. It appears that we have been quite successful in predicting the realized item difficulties. When we look at the last nine examinations of the ability to listen (program 1), the correlation between the expected and the realized difficulties, averaged over items, is about 0.82. Both sets of p-values ranged between 0.63 and 0.68 as intended by the examination committee. We have not yet gained any experience with the expected information or the IRCs.

6.2 Estimating the Correlation Between two Latent Traits

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 14 shows that the correlation $\rho(\hat{\theta}, \hat{\xi})$ may be much lower than $\rho(\theta, \xi)$ if the estimates are not very reliable. While $\rho(\hat{\theta}, \hat{\xi})$ may be estimated from a data set we need to estimate the reliabilities in order to calculate $\rho(\theta, \xi) = \rho(\hat{\theta}, \hat{\xi}) / \sqrt{\rho^2_{\hat{\theta}} \rho^2_{\hat{\xi}}}$.

To do so, we could use the estimated asymptotic variance of $\hat{\theta}$ (or $\hat{\xi}$) given θ (or ξ) (e.g., Verhelst, Glas & Verstralen, 1995, pp. 63-64). This works well as long as there are sufficient items to justify the use of the asymptotic variance, and as long as the estimates are unbiased. When the number of items is small and/or when there are many extreme test scores the estimated reliabilities may be wrong.

An alternative procedure is as follows. 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 gives us an estimated ability $\hat{\theta}(y)$ and $g(Y = y|\theta) = g(\hat{\theta} =$

$\hat{\theta}(y)|\theta)$; the distribution of the estimated abilities given θ. The variance of $\hat{\theta}$ given θ may now be calculated as

$$Var(\hat{\theta}|\theta) = E[\hat{\theta}^2|\theta] - (E[\hat{\theta}|\theta])^2 \quad (30)$$

$$= \sum_y \hat{\theta}^2(y) g(\hat{\theta} = \hat{\theta}(y)|\theta) - \left(\sum_y \hat{\theta}^2(y) g(\hat{\theta} = \hat{\theta}(y)|\theta) \right)^2 \quad (31)$$

We may then calculate the reliabilities via Equation (13) using numerical integration to calculate the expectation. These estimates are expected to be more robust against bias in the estimates.

7 Discussion

The purpose of this paper has been to clarify CTT. Our presentation of CTT was non-standard for two reasons. First, we have assumed that measurements are discrete. Second, we have assumed that the CRFs are defined by a suitable IRT model. The main advantage of the use of an IRT model is that we can test hypotheses on the relation between ability and the measurements. Computer programs such as OPLM (Verhelst, Glas & Verstralen, 1995) or CONQUEST (Wu, Adams, & Wilson 1997) provide a myriad of information on the appropriateness of the IRT model. Through the analysis we learn whether it is reasonable to assume that the measures are unidimensional, and whether the true score increases with ability or not, etc. When an appropriate IRT model is used this allows us to calculate classical indices for properties of items and test in situations where CTT could normally not be applied. For example, when the items of interest were not administered to the population of interest.

Throughout, the IRT model that we have used was the GPCM which is often used in practice. We could have used any other model. We could, for example, have considered 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

$$\rho_Y^2 = \frac{\sum_g p_g \left(\sum_i p_{ig}\right)^2 - \left(\sum_g p_g \sum_i p_{ig}\right)^2}{\sum_g p_g \left(\sum_i p_{ig}\right)^2 - \left(\sum_g p_g \sum_i p_{ig}\right)^2 + \sum_g p_g \sum_i p_{ig}(1 - p_{ig})}, \qquad (32)$$

where $\sum_g$ sums over latent classes, $\sum_i$ sums over items, 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 clearly perceptible what Equation 32 means but it appears to be related to the ability to discriminate between the classes. For example, $\rho_Y^2 = 1$ if there are one or more items that distinguish perfectly between the classes. More specialized indices have been developed to judge the quality of the items (e.g., Rost, 1996, pp. 1153-159).

It must be noted that all calculations depend on the validity of the IRT model and the availability of good estimates of the distribution in the population of interest. If we assume that the distribution of ability was normal with mean μ and variance σ_θ^2, an approximate 95% interval of uncertainty may be constructed by varying σ_θ between $\sigma_\theta^{(\text{low})} = \sigma_\theta^{(g)} - 1.64SE$ and $\sigma_\theta^{(\text{high})} = \sigma_\theta + 1.64SE$, respectively, where SE denotes the standard error of the standard deviation of the reference population. For example, we calculate the expected information in a population of interest with $\sigma_\theta^{2(g)} = (\sigma_\theta^{(\text{low})})^2$ to get the lower end of an approximate 95% interval, and then with $\sigma_\theta^{2(g)} = (\sigma_\theta^{(\text{high})})^2$ 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 $\rho^2_{X_i}$, ρ^2_Y or $E[I(\theta)]$. Another possibility is to take expectations over the posterior distribution of the parameters (e.g., Lewis, 2001).

As a topic for future research the relation between item properties of CTT and IRT item parameters to the quality of decision making should be considered. Many tests are used to make decisions on students while they are not composed in a way that is known to minimize the probability of erroneous decisions.

8 References

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New-York: Holt, Rinehart and Winston.

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Pscyhological Measurement, 23,* 249-261.

Feldt, L.S. (1993). The relationship between the distribution of item difficulties and test reliabilility. *Applied Psychological Measurement, 6,* 37-49.

Guilford, J.P., & Fruchter, B. (1978). *Fundamental statistics in Psychology and Education.* (6th Ed.) Tokyo: McGraw-Hill Kogakusha.

Heuvelmans, A. (2001). *TiaPlus user's manual.* Cito: Arnhem. Available at http://www.citogroep.nl/pok/poc/eind_fr.htm.

Lewis, C. (2001). Expected response functions. Chapter 9 in "*Essays on item response theory.*" Edited by A.Boomsma, M.A.J. van Duijn and T. Snijders. New-York: Springer

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51,* 493-504.

Lord, F.M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17,* 181-194.

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

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion.* [Textbook for

test theory and test construction] Hans Huber : Bern.

Steyer, R., & Eid, M. (1993). *Messen und testen.* 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.

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., & Verstralen, H.H.F.M. (1995). *One parameter logistic model OPLM.* Computer software manual, Cito: Arnhem.

Verstralen, H.H.F.M. (1989). *Het groeperen van frequentieverdelingen.* [Dividing frequency distribution into groups]. Measurement and Research Report 98-3. Cito: Arnhem.

Verstralen, H.H.F.M. (1996a). *Estimating integer parameters in IRT models for polytomous items.* Measurement and Research Report, Cito: Arnhem.

Verstralen, H.H.F.M. (1996b). OPCAT: *Estimating integer category weights in OPLM.* User Manual. Arnhem: Cito

Wu, M.L., Adams, R.J., & Wilson, M.R. (1997). *ConQuest: Generalized item response modelling software.* Melbourne: ACER.

9 Appendix

Under the GPCM, the derivatives are given by

$$\begin{aligned} \frac{\partial}{\partial\theta}P_{ik}(\theta) &= \alpha_i P_{ik}(\theta)[k - \sum_{r=1}^{M_i} r P_{ir}(\theta)] \\ &= \alpha_i P_{ik}(\theta)\left(k - E_v[X_i|\theta]\right). \end{aligned} \tag{33}$$

It is easy to see that the derivatives are uniformly bounded, that is,

$$\alpha_i(k - M_i) \leq \frac{\partial}{\partial\theta}P_{ik}(\theta) \leq \alpha_i k, \text{ for all } \theta. \tag{34}$$

It follows that $\frac{\partial}{\partial\theta}P_{i0}(\theta) \leq 0$, and $\frac{\partial}{\partial\theta}P_{iM_i}(\theta) \geq 0$; $P_{i0}(\theta)$ is non-increasing, and $P_{iM_i}(\theta)$ is non-decreasing in θ. If $\alpha_i > 0$, the true score is a continuous, strictly increasing function that varies smoothly between 0 and M_i. This implies that, for all $0 < k < M_i$, there is a real value θ_k such that $E_v[X_i|\theta = \theta_k] = k$. From 33 It follows that $\frac{\partial}{\partial\theta}P_{ik}(\theta_k) = 0$. If $\theta > \theta_k$, $E_v[X_i|\theta] > E_v[X_i|\theta_k]$ and $\frac{\partial}{\partial\theta}P_{ik}(\theta)$ is negative. If $\theta < \theta_k$, $E_v[X_i|\theta] < E_v[X_i|\theta_k]$ and $\frac{\partial}{\partial\theta}P_{ik}(\theta)$ is positive.

Note that if $\alpha_i \to \infty$, the true score function $E_v[X_i|\theta]$ becomes a step function (see Figure 2). The information becomes zero almost everywhere, except at those values of θ where $E_v[X_i|\theta] = k$. Thus, when α_i increases beyond limits, the information becomes concentrated around values of θ where $E_v[X_i|\theta] = k$, and the CRFs of categories $0 < k < M_i$ peak.

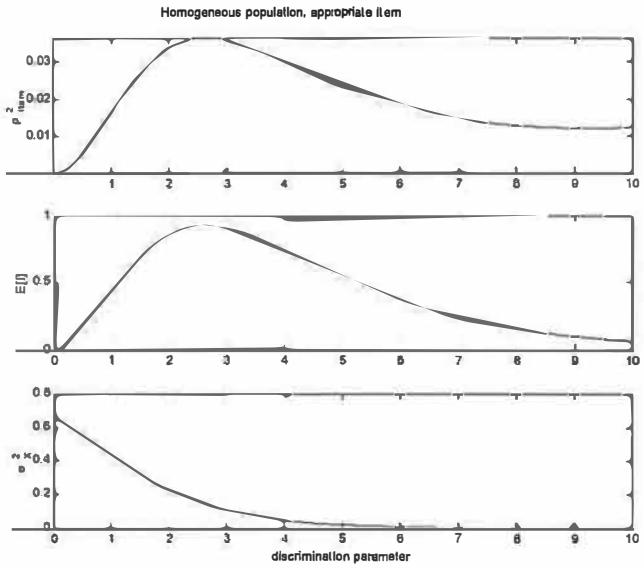

Figure 13: Plots of ρ^2_{item}, $E[I_i(\theta)]$, and $\sigma^2_{X_i}$ against the α-parameter. $\boldsymbol{\delta} = (0, 2)$ and the population is normal with mean 1 and standard deviation 0.2.

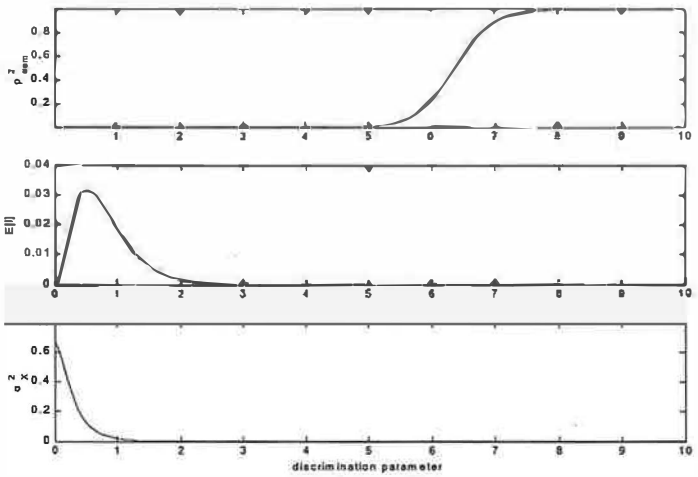

Figure 14: Plots of ρ^2_{item}, $E[I_i(\theta)]$, and $\sigma^2_{X_i}$ against the α-parameter. $\boldsymbol{\delta} = (0, 2)$ and the population is normal with mean 6 and variance 0.09. This population is way out of the effective range of the item.

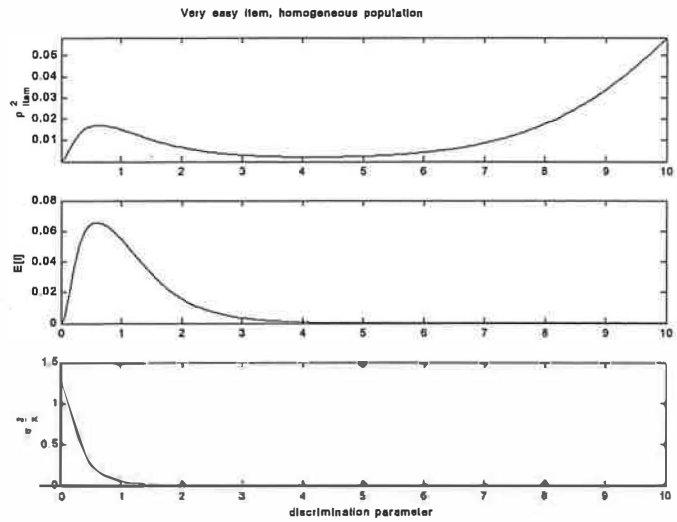

Figure 15: Plots of ρ^2_{item}, $E[I_i(\theta)]$, and $\sigma^2_{\mathbf{X}_i}$ against the α-parameter. $\boldsymbol{\delta} = (-1, 0, -0.1)$. The population is normal with mean -4 and variance (0.25).

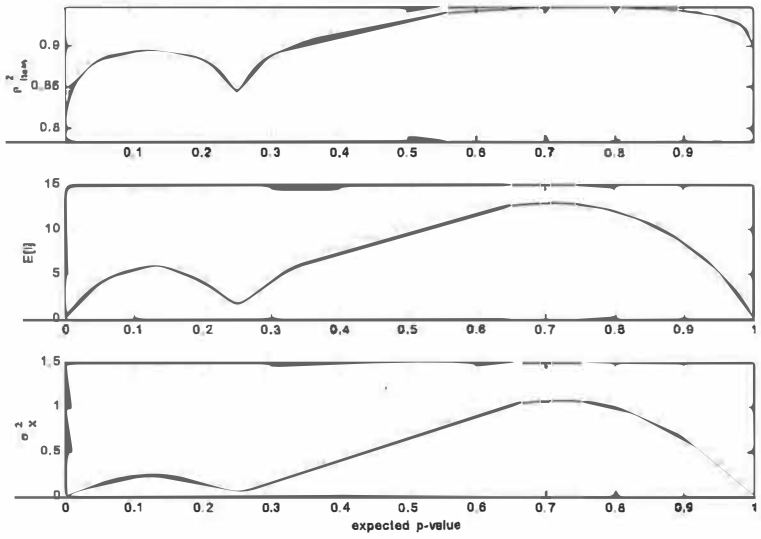

Figure 16: $\rho^2_{\bar{X}_i}$, $E[I_i(\theta)]$, and $\sigma^2_{\bar{X}_i}$ as a function of π. $\alpha_i = 15$, and $\boldsymbol{\delta} = (-3, 1, 7, -1)$.

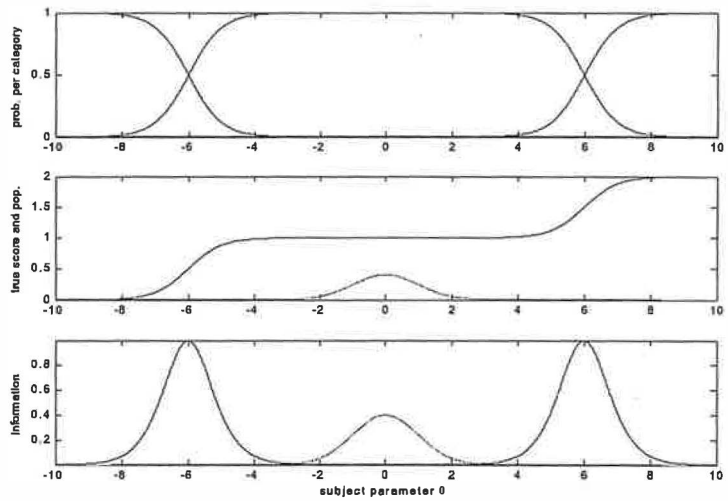

Figure 17: CRFs, information and true score. $\alpha_i = 2$, and $\boldsymbol{\delta} = (-6, 6)$. The bell-shaped curve shows the location of the population which is normal with mean 0 and standard deviation 1 so that $\pi = 0.5$.