Reliability Measures In Item Response Theory: Manifest Versus Latent Correlation Functions

MILANZI, Elasma; MOLENBERGHS, Geert; ALONSO ABAD, Ariel; VERBEKE, Geert & De Boeck, Paul (2015) Reliability Measures In Item Response Theory: Manifest Versus Latent Correlation Functions. In: British Journal of Mathematical & Statistical Psychology, 68 (1), p. 43-64.

DOI: 10.1111/bmsp.12033
Handle: http://hdl.handle.net/1942/16178

Reliability Measures In Item Response Theory: Manifest Versus Latent Correlation Functions

Elasma Milanzi[1] **Geert Molenberghs**[1,2]

Ariel Alonso[1] **Geert Verbeke**[2,1]

Paul De Boeck[3]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*

[3] *Department of Psychology, Higher Cognition and Individual Differences, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium & Universiteit van Amsterdam, the Netherlands*

Abstract

For item response theory (IRT) models, which belong to the class of generalized linear or non-linear mixed models, reliability at the scale of observed scores, i.e., manifest correlation, is usually of greater scientific interest, while its calculation is more difficult than latent correlation based reliability. This is not in the least because it cannot be calculated explicitly when the logit link is used in conjunction with normal random effects. As such, approximations like Fisher's information coefficient, Cronbach's α, or the latent correlation are calculated, allegedly because it is easy to do so. Cronbach's α has well-known and serious drawbacks, Fisher's information is not meaningful under certain circumstances and there is an important but often overlooked difference between latent and manifest correlations. Here, manifest correlation refers to correlation between observed scores, while latent correlation refers to correlation between scores at the latent scale, e.g., logit or probit scale. Thus, using one in place of the other can lead to erroneous conclusions. Taylor series based reliability measures, which are based on manifest correlation functions, are derived and a careful comparison of reliability measures based on latent correlations, Fisher's information, and exact reliability is effectuated. The latent correlations are virtually always considerably higher than their manifest counterparts, Fisher's information measure shows no coherent behavior (it is even negative in some cases), while the newly introduced Taylor series based approximations reflect the exact reliability very closely. Comparisons among the various types of correlations, for various IRT models, are made using algebraic expressions, Monte Carlo simulations, and data analysis. Considering the light computational burden and the performance of Taylor series based reliability measures, their use is recommended.

Some Keywords: 1PL model; 2PL model; Logit link; Probit link; Rasch model.

1 Introduction

Measurement studies play a vital role in exploring various attributes, like social and intellectual behavior. The relevance of such studies depends on several, equally important factors, which include reliability of the tool used for measurement. Culligan (2008) defines reliability as a measure of

the consistency of the application of an instrument to a particular population at a particular time. Classical test theory (CTT) remains one of the most used paradigms for the analysis of measurement studies, within which the concept of reliability is well developed. In CTT, the reliability measure is simply the proportion of true to observed variance. Limitations in classical theory (Schaeffer *et al.*, 1986; O'Brien, 1995) have led to the development of several alternatives, in some of which deriving a reliability measure is straightforward, such as in CTT, while in others it is not.

Examples include generalizability theory (GT), which makes use of linear mixed models to estimate various variance components used in estimating reliability (Van Leeuwen et al., 1998). With linear mixed models, the definition of reliability as a proportion of true to observed variance is easily carried forward due to the nice properties of the normal distribution, which is usually assumed for the observed responses, the most notable being the separation of mean and variance parameters.

For binary response items, Item response theory (IRT) has indisputably commanded wide application among measurement studies, mostly for its advantages over classical theory (Rasch, 1960; De Ayala, 2009). Much as IRT models are commonly used for measurement of variables like attributes and attitudes (Van der Linden & Hambleton, 1997), the question of reliability of measurements (Spearman, 1904), which is crucial for such studies, cannot be ignored.

Unfortunately, peculiarities emerge when dealing with binary responses. For normally distributed outcomes, reliability of measurement reduces to, $\sigma_\theta^2/(\sigma_\theta^2 + \sigma_\varepsilon^2)$, where σ_θ^2 is the variance of the person trait θ and σ_ε^2 is the variance of the distribution assumed for the errors. This is commonly referred to as intraclass correlation (Molenberghs & Verbeke, 2005). Directly using intraclass correlation for dichotomous responses produces what is known as *latent correlation* because it gives the correlation between responses at a logit or probit scale. It follows that reliability measured using latent correlation will be at a latent scale. More often than not, scientific interest is in the reliability of the observed scores rather than the latent ones such that meaningful reliability measures have to be based on *manifest correlation*, i.e., correlation between observed scores. While for normally distributed outcomes latent and manifest correlations coincide, this is no longer the case for binary or other non-normal outcomes. Hence, to obtain meaningful reliability measures, appropriate quantities for the intraclass correlation formula have to be derived.

Reliability measures based on manifest correlations are not as well developed for dichotomous responses as is the case for continous responses. Briggs & Wilson (2007) and Rodríguez & Elo (2003) note that these are usually difficult to derive because they involve the evaluation of integrals that lack closed forms and thus are not widely used. To bypass such difficulties, approximate reliability measures are preferred and these include: Cronbach's α, the intra-class correlation, and Fisher's Information measure. Drawbacks for using Cronbach's α in IRT have been well documented (Cronbach & Shavelson, 2004), and Fisher's Information measure has limited application, given that under some conditions it can be negative (Mesbah *et al.*, 2002). Therefore, it is not meaningful in some cases. Furthermore, its extension to models with multi-dimension traits is not clear. Some drawbacks regarding the use of the intraclass correlation in IRT, i.e., using reliability measures based on latent correlation, will be highlighted in the sections to follow.

Most of the IRT models fall into the family of generalized linear mixed models (GLMM), an extension of linear mixed models to a special family of non-linear mixed models (Molenberghs & Verbeke, 2005; Rijmen *et al.*, 2003), where the outcome is of a non-Gaussian type, but the effects of predictor variables still enter a so-called linear predictor function. Vangeneugden *et al.* (2010) derived approximate manifest variance-covariance functions and correlation functions for the GLMM family. The approximation totally evades the need to evaluate integrals and requires the input of estimates that

are easily obtained during models estimation.

Taking into account that One- and Two-parameter Logistic (1PL and 2PL) models belong to the GLMM family, we explore the usefulness of such approximations in these two IRT models towards obtaining reliability measures. Whereas the goal of Vangeneugden *et al.* (2010) was to estimate manifest correlations between two binary outcomes within a subject (which are the correlations of interest for most model members of the GLMM), by studying these approximations in the context of a combination of Classical Test Theory (CTT) and IRT, we derive reliability measures, both at the expected item score and expected sum score levels, that directly correspond to the definition of reliability in CTT, i.e., proportion of true to observed variance, which are based on manifest correlations. The performance of these measures will be assesed through a simulation study which will compare the newly derived approximations, latent and Fisher Information based reliability measures to the exact reliability. Applicability will be shown through an empirical data analysis of Verbal Aggression and Law School Admission Test datasets.

The organization of this paper is as follows. A brief description of the case studies is presented in Section 2. Section 3 is dedicated to the measures of reliability in IRT, and Section 4 reviews the general correlation functions as derived by Vangeneugden *et al.* (2010), including the derivation of the correlation functions for the specific 1PL and 2PL models. The design and results of the simulation study are described in Section 5. The analysis of the case studies in Section 6 is followed by concluding remarks in Section 7.

2 Data Description

2.1 Verbal Aggression Data

The data consist of subjects' responses to questions about verbal aggression. The instrument is a behavioral questionnaire. All items refer to verbally aggressive reactions in a frustrating situation. The data can also be considered as from a psychological experiment which has three design factors. (1) Behavior mode: a differentiation is made between two levels, i.e., wanting to do and the actual doing. (2) Situation type: This factor has two levels, namely other-to-blame and self-to-blame situation type, and each of these levels has two situations. Self to blame situations were: *'The grocery closes just as I am about to enter'* and *'The operator disconnects me when I had used up my last 10 cents'*. Other to-blame situations were: *'A bus fails to stop for me'* and *'I miss the train because a clerk gave me faulty information'*. So, the situations can also be viewed as nested in the situation type. (3) Behavior type: this had three kinds of behaviors, namely shout, scold, and curse. An example of an item in this instrument was: *'A bus fails to stop for me. I would want to curse'*. Possible answers were, no (0), perhaps (1), and yes (2). In our application, we will use the dichotomized version of the response, in which 'no' and 'perhaps' are recoded as 0 and 'yes' as 1. A detailed description of the data and its items can be found in Vansteelandt (2000) and De Boeck & Wilson (2004).

2.2 Law School Admission Test (LSAT6) Data

LSAT is a standardized test administered to prospective law students and designed to assess reading comprehension, logical and verbal reasoning proficiencies. The data comprise scores on five items of Section 6 of of the LSAT, for 1000 examinees. The data are publicly available in the R package

mirt; they are well described in Bock & Lieberman (1970).

3 Reliability Measures in One Parameter Logistic (1PL) and Two Parameter Logistic (2PL) Models

Reliability measures of common interest for 1PL and 2PL models, which are also the focus of this work, include reliability of expected item score and reliability of the expected sum score. This section reviews both exact and approximate methods for estimating such measures.

3.1 Exact Reliability Measures

Customarily, expected item and sum score reliability measures are computed based on observed scores, which are binary in nature for 1PL and 2PL models. While exact methods for estimating such measures are widely known, they are rarely used in practice because they are computationally intensive; instead, approximations are preferred. The exact measures are reviewed to facilitate comparison with the approximate one.

Define Y_{ji} as a realized score on item $i = 1, \ldots, I$ by person $j = 1, \ldots, N$. It is common to express the observed score as:

$$Y_{ji} = \mu_{ji} + \varepsilon_{ji},$$

where μ_{ji} is the true score and ε_{ji} is the error score. Equivalently, the 2PL model formulates the observed score as

$$Y_{ji} = \mu_{ji} + \varepsilon_{ji} = \frac{\exp\left[\alpha_i(\theta_j - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_j - \beta_i)\right]} + \varepsilon_{ji}, \tag{1}$$

where θ_j is the person trait score, with θ_j having variance σ_θ^2, α_i and β_i the discrimination and difficulty values for item i, and ε_{ji} the error term, which in this case is a function of μ_{ji}.

Recall that our focus is on the expected item and expected sum scores, which are given by:

$$\begin{aligned} Y_i &= \int \left\{ \frac{\exp\left[\alpha_i(\theta_j - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_j - \beta_i)\right]} + \varepsilon_{ji} \right\} \phi(\theta_j|0, \sigma_\theta^2)\, d\theta \;\; = \mu_i + \varepsilon_i \\ S_T &= \int \sum_{i=1}^{I} \left\{ \frac{\exp\left[\alpha_i(\theta_j - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_j - \beta_i)\right]} + \varepsilon_{ji} \right\} \phi(\theta_j|0, \sigma_\theta^2)\, d\theta \;\; = \mu + \varepsilon, \end{aligned} \tag{2}$$

respectively. Dimitrov (2003) defines their corresponding variances as:

$$\begin{aligned} \mathsf{Var}(\varepsilon_i) &= \int \mu_{ji}(1 - \mu_{ji})\, \phi(\theta_j|0, \sigma_\theta^2)\, d\theta, \\ \mathsf{Var}(\mu_i) &= \mu_i(1 - \mu_i) - \mathsf{Var}(\varepsilon_i), \\ \mathsf{Var}(\varepsilon) &= \sum_{i=1}^{I} \mathsf{Var}(\varepsilon_i), \\ \mathsf{Var}(\mu) &= \int \left\{ \sum_{i=1}^{I} \frac{\exp\left[\alpha_i(\theta_j - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_j - \beta_i)\right]} \right\}^2 \phi(\theta_j|0, \sigma_\theta^2)\, d\theta \\ &\quad + \left\{ \int \sum_{i=1}^{I} \frac{\exp\left[\alpha_i(\theta_j - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_j - \beta_i)\right]} \phi(\theta_j|0, \sigma_\theta^2)\, d\theta \right\}^2, \end{aligned}$$

with reliability defined as the proportion of true variance to observed variance. Further,

$$\begin{aligned}\rho_i &= \frac{\mathsf{Var}(\mu_i)}{\mathsf{Var}(\mu_i)+\mathsf{Var}(\varepsilon_i)},\\ \rho_s &= \frac{\mathsf{Var}(\mu)}{\mathsf{Var}(\mu)+\mathsf{Var}(\varepsilon)}\end{aligned}$$

measure the expected item score and expected sum score reliability, respectively. Each of the components for obtaining these measures involves integration of normal random effects over binary data distributions, which are known to lack a closed form. The computational burden associated with these measures arguably results into their infrequent use in practice.

3.2 Intra-class Correlation (Latent)

The definition of the intra-class correlation stems directly from the definition of reliability in CTT, which is the ratio of the true over the observed variance. Let Y_{ji} be a continuous observed score for person $j = 1, \ldots, N$ on item $i = 1, \ldots, I$, and further $\boldsymbol{Y}_j^c = (Y_{j1}, \ldots, Y_{jI})$. Using a linear mixed model, the observed score can be expressed in terms of the true score $\boldsymbol{\mu}_j$ and error as follows:

$$\boldsymbol{Y}_j^c = \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_j = \boldsymbol{\theta}_j - \boldsymbol{\beta} + \boldsymbol{\varepsilon}_j, \tag{3}$$

where $\boldsymbol{\beta}$ is the vector of item difficulties, $\theta_j \sim N(0, \sigma_\theta^2)$, and $\boldsymbol{\varepsilon}_j \sim N(0, \sigma_\varepsilon^2 \mathbf{I}_I)$, are the measures of person trait and random errors, respectively. Further, $\mathsf{Cov}(\theta_j, \boldsymbol{\varepsilon}_j) = 0$. It has been shown that

$$\mathsf{Var}(\boldsymbol{Y}_j^c) = \mathsf{Var}(\boldsymbol{\mu}_j) + \mathsf{Var}(\boldsymbol{\varepsilon}_j) = \mathbf{1}_I \sigma_\theta^2 \mathbf{1}'_I + \sigma_\varepsilon^2 \mathbf{I}_I$$

(Verbeke & Molenberghs, 2000). For illustrative purposes, consider the case of $I = 2$. Then,

$$\mathsf{Var}(\boldsymbol{\mu}_j) = \begin{bmatrix}\sigma_\theta^2 & \sigma_\theta^2\\ \sigma_\theta^2 & \sigma_\theta^2\end{bmatrix}, \qquad \mathsf{Var}(\boldsymbol{\varepsilon}_j) = \begin{bmatrix}\sigma_\varepsilon^2 & 0\\ 0 & \sigma_\varepsilon^2\end{bmatrix}. \tag{4}$$

Item and sum score reliability measures follow as:

$$\rho_{i_1} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}, \quad \rho_{s_2} = \frac{2\sigma_\theta^2}{2\sigma_\theta^2 + \sigma_\varepsilon^2}, \tag{5}$$

which correspond to ICC(1) and ICC(k) measures in McGraw & Wong (1996).

The same arguments can be followed when $\boldsymbol{\mu}_j = \alpha(\boldsymbol{\theta}_j - \boldsymbol{\beta})$ and the equivalent quantities are:

$$\begin{aligned}\mathsf{Var}(\boldsymbol{Y}_j^c) &= \mathsf{Var}(\boldsymbol{\mu}_j) + \mathsf{Var}(\boldsymbol{\varepsilon}_j) = \boldsymbol{\alpha}_I \sigma_\theta^2 \boldsymbol{\alpha}'_I + \sigma_\varepsilon^2 \mathbf{I}_I,\\ \mathsf{Var}(\boldsymbol{\mu}_j) &= \begin{bmatrix}\alpha_i^2\sigma_\theta^2 & \alpha_i\alpha_{i'}\sigma_\theta^2\\ \alpha_i\alpha_{i'}\sigma_\theta^2 & \alpha_{i'}^2\sigma_\theta^2\end{bmatrix},\\ \mathsf{Var}(\boldsymbol{\varepsilon}_j) &= \begin{bmatrix}\sigma_\varepsilon^2 & 0\\ 0 & \sigma_\varepsilon^2\end{bmatrix},\\ \rho_{i_2} &= \frac{\alpha_i^2\sigma_\theta^2}{\alpha_i^2\sigma_\theta^2 + \sigma_\varepsilon^2}, \qquad (6)\\ \rho_{s_2} &= \frac{\sigma_\theta^2(\alpha_i + \alpha_{i'})^2}{\sigma_\theta^2(\alpha_i + \alpha_{i'})^2 + 2\sigma_\varepsilon^2}. \qquad (7)\end{aligned}$$

Now, consider a binary observed score modeled through the 1PL model as:

$$\text{logit}\left[P(\boldsymbol{Y}_j = 1)\right] = \boldsymbol{\theta}_j - \boldsymbol{\beta}. \tag{8}$$

It has been noted (Agresti, 2002) that (8) can be expressed in the form of (3), as follows:

$$\boldsymbol{Y}_j^* = \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_j = \boldsymbol{\theta}_j - \boldsymbol{\beta} + \boldsymbol{\varepsilon}_j^*, \tag{9}$$

where Y_{ji}^* is assumed to be a latent continuous score underlying the dichotomization of Y_{ji}, such that $Y_{ji} = 1$ if $Y_{ji}^* \geq C$ and 0 otherwise, with C a pre-specified threshold and $\varepsilon^* \sim$ Logistic$(0, 1)$. Model (9) is a typical linear mixed model, hence the theory behind the derivation of reliability measures (5), (6), and (7), can be applied directly and the corresponding reliability measures are:

$$\rho_{i_{l1}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \frac{\pi^2}{3}} \quad \text{and} \quad \rho_{s_{l1}} = \frac{2\sigma_\theta^2}{2\sigma_\theta^2 + \frac{\pi^2}{3}}, \tag{10}$$

for item and test score, respectively, for the 1PL model. Equivalent expressions for the 2PL model are:

$$\rho_{i_{l2}} = \frac{\alpha_i^2\sigma_\theta^2}{\alpha_i^2\sigma_\theta^2 + \frac{\pi^2}{3}}, \quad \text{and} \quad \rho_{s_{l2}} = \frac{\sigma_\theta^2(\alpha_i + \alpha_{i'})^2}{\sigma_\theta^2(\alpha_i + \alpha_{i'})^2 + \frac{2\pi^2}{3}}, \tag{11}$$

for the item and test reliability, respectively, $\pi^2/3$ is the variance of the underlying error distribution, the standard logistic distribution. While the mathematical motivation is appealing, it is evident that (10) and (11) only depend on σ_θ^2. Hence, any change in this value, for example, due to change in identification restrictions, will result in varying reliability. These are examples of reliability measures based on latent correlation and, as noted before, for the binary case such measures do not coincide with their manifest correlation based counterparts. In a way this is unfortunate, especially in cases where there are no closed forms for the marginal model stemming from the hierarchical formulation. Note that the difference between latent and manifest correlation is independent of the existence of a closed form. For example, a probit link with normal random effects allows for a closed-form formulation (Molenberghs *et al.*, 2010), but also there the two correlations have a different expression. In general, latent correlation should not be used when manifest correlation is of interest.

3.3 Fisher Information

A common approach for dichotomous IRT models, is to use a Fisher information based test score reliability measure, which for the 2PL model is obtained as follows. Define total information as:

$$I(\theta_j; I) = \sum_{i=1}^{I} \alpha_i^2 \frac{\exp(\eta_{ji})}{[1 + \exp(\eta_{ji})]^2},$$

with $\eta_{ji} = \alpha_i(\theta_j - \beta_i)$, where β_i is the difficulty value for item i, θ_j is the value of the latent trait for person j, and I is the total number of items. The expression is intuitively appealing, given that it is a sum over standard logistic variances. An approximation of the reliability coefficient follows as:

$$\rho_f = 1 - \frac{1}{N}\sum_{j=1}^{N} \frac{I(\theta_j; I)^{-1}}{\widehat{\sigma}_\theta^2} \tag{12}$$

(Lord, 1983), where $\widehat{\sigma}_\theta^2$ is the estimated variance of the latent trait, e.g., the observed variance of person parameters and N the total number of persons. This approximation is only valid when the number of items is large; it also requires the knowledge of true values of difficulty parameters, information which usually is beyond reach. A 1PL equivalent follows by setting $\alpha = 1$.

4 Taylor-series-based Derivation of the Correlation Function

4.1 Manifest Correlation Functions For GLMM

We present here a brief review of the explicit but approximate correlation functions, as based on Taylor-series approximations as derived for GLMM family and explained in detail in Vangeneugden *et al.* (2010). Let Y_{ji} be the i^{th} outcome measured on person j , $j = 1, \ldots N$ and $i = 1, \ldots, I$; further let $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn_j})'$.

Write the general model as $\boldsymbol{Y}_j = \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_j$, where the conditional mean, given the random effects are written as $\boldsymbol{\mu}_j = h(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{\theta}_j)$, and where $\boldsymbol{\beta}$ is a vector of fixed-effects parameters, $\boldsymbol{\theta}_j$ are random effects, $\boldsymbol{X}_j$ and $\boldsymbol{Z}_j$ are known design matrices, and h is a known link function. Finally, $\boldsymbol{\varepsilon}_j = (\varepsilon_{j1}, \ldots, \varepsilon_{jn_j})'$ is the residual error component.

The general formula for the variance-covariance matrix of $\boldsymbol{Y}_j$ was derived without any restriction on the distribution of the outcome variable, nor on the complexity of the model, e.g., allowing for serial correlation or not. This maximizes the similarity with the case of continuous, normally distributed outcomes. However, a key distinction is that in the Gaussian case the mean and variance are functionally independent parameters, whereas here the residual variance will follow from the mean. The variance-covariance matrix can be written as:

$$\boldsymbol{V}_j = \mathsf{Var}(\boldsymbol{Y}_j) = \mathsf{Var}(\boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_i) = \mathsf{Var}(\boldsymbol{\mu}_j) + \mathsf{Var}(\boldsymbol{\varepsilon}_j) + 2\mathsf{Cov}(\boldsymbol{\mu}_j, \boldsymbol{\varepsilon}_j). \tag{13}$$

Because $\boldsymbol{\mu}_j$ depends on $\boldsymbol{\theta}_j$ only, which is independent of $\boldsymbol{\varepsilon}_j$, it follows that $\mathsf{Cov}(\boldsymbol{\mu}_j, \boldsymbol{\varepsilon}_j) = 0$, and the first term in (13), using a first-order Taylor series expansion around $\boldsymbol{\theta}_j = 0$ reduces to:

$$\begin{aligned} \mathsf{Var}(\boldsymbol{\mu}_j) &= \mathsf{Var}[\boldsymbol{\mu}_j(\boldsymbol{\eta}_j)] = \mathsf{Var}[\boldsymbol{\mu}_j(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{\theta}_j)] \\ &\cong \left(\frac{\partial\boldsymbol{\mu}_j}{\partial\boldsymbol{\eta}_j}\frac{\partial\boldsymbol{\eta}_j}{\partial\boldsymbol{\theta}_j}\mid_{\boldsymbol{\theta}_j=0}\right)\boldsymbol{D}\left(\frac{\partial\boldsymbol{\mu}_j}{\partial\boldsymbol{\eta}_j}\frac{\partial\boldsymbol{\eta}_j}{\partial\boldsymbol{\theta}_j}\mid_{\boldsymbol{\theta}_j=0}\right)' \cong \boldsymbol{\Delta}_j\boldsymbol{Z}_j\boldsymbol{D}\boldsymbol{Z}_j'\boldsymbol{\Delta}_j', \end{aligned} \tag{14}$$

where $\boldsymbol{\Delta}_j = \frac{\partial\boldsymbol{\mu}_j}{\partial\boldsymbol{\eta}_j}\mid_{\boldsymbol{\theta}_j=0}$. The second term in (13), leads to:

$$\mathsf{Var}(\boldsymbol{\varepsilon}_j) = \mathsf{Var}[E(\boldsymbol{\varepsilon}_j \mid \boldsymbol{\theta}_j)] + E[\mathsf{Var}(\boldsymbol{\varepsilon}_j|\boldsymbol{\theta}_j)] = E[\mathsf{Var}(\boldsymbol{\varepsilon}_j \mid \boldsymbol{\theta}_j)] = \boldsymbol{\Xi}_j^{\frac{1}{2}}\boldsymbol{\Sigma}_j\boldsymbol{\Xi}_j^{\frac{1}{2}}, \tag{15}$$

where $\boldsymbol{\Xi}$ is a diagonal matrix with the overdispersion parameters along the diagonal. If there are no overdispersion parameters, $\boldsymbol{\Xi}_j$ is set equal to the identity matrix. Expand the variance function $\boldsymbol{\Sigma}_j$ so that

$$\mathsf{Var}(\boldsymbol{\varepsilon}_j) = \boldsymbol{\Xi}_j^{\frac{1}{2}}\boldsymbol{A}_j^{\frac{1}{2}}\boldsymbol{R}_j\boldsymbol{\Xi}_j^{\frac{1}{2}}\boldsymbol{A}_j^{\frac{1}{2}}, \tag{16}$$

where $\boldsymbol{R}_j$ is the correlation matrix and $\boldsymbol{A}_j$ is a diagonal matrix containing the variances following from the generalized linear model specification of Y_{ji} given the random effects $\boldsymbol{\theta}_j = 0$, i.e., with diagonal elements $v(\mu_{ji})|_{\boldsymbol{\theta}_j=0}$. Using (14) and (16), we have the following expression for the variance-covariance matrix (13):

$$\boldsymbol{V}_j = \boldsymbol{\Delta}_j\boldsymbol{Z}_j\boldsymbol{D}\boldsymbol{Z}_j'\boldsymbol{\Delta}_j' + \boldsymbol{\Xi}_j^{\frac{1}{2}}\boldsymbol{A}_j^{\frac{1}{2}}R_j\boldsymbol{\Xi}_j^{\frac{1}{2}}\boldsymbol{A}_j^{\frac{1}{2}}. \tag{17}$$

Evidently, from this variance-covariance matrix, we can easily obtain the correlations. While the above derivation is referred to as a first-order Taylor series expansion, the exact same expression follows if a second-order expansion is considered, owing to terms vanishing. Therefore, we are authorized to refer to it as a second-order Taylor series expansion which according to Vangeneugden *et al.* (2011), who explored the quality of approximation by considering higher-order series, gives a good approximation.

4.2 Taylor Series Based Reliability Measures For 1PL and 2PL Models.

Loosely, reliability is an indicator of strength of agreement of particular scores (depending on the form of reliability); usually it takes the form of a correlation function. To obtain reliability measures on the scale of observed scores, manifest correlation functions have to be used. While such functions are easily obtained for continuous data, usually for binary data, approximations are employed. This section discusses the relevance of the approximate variance-covariance matrix in a GLMM for assessing reliability of the expected item score and the expected sum score.

Without loss of generality, consider a measurement tool with $i = 1, \ldots, I$ items, responded to by $j = 1, \ldots, N$ persons and further let Y_{ji} be the binary score on item i by person j. Then, parameterize the 2PL model as

$$Y_{ji} = \underbrace{\frac{\exp\left[\alpha_i(\theta_j - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_j - \beta_i)\right]}}_{\mu_{ji}} + \varepsilon_{ji}, \tag{18}$$

where $\theta_j \sim N(0, \sigma_\theta^2)$, β_i and α_i are the difficulty value and discrimination parameters, respectively, for item i. The model formulation in Section 4.1, $\boldsymbol{Y}_j = \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_j$, is basically the matrix representation of (18), where $\boldsymbol{Y}_j$ is the vector of Y_{ji}'s, for person j on all items; we proceed similarly for $\boldsymbol{\mu}_j$ and $\boldsymbol{\varepsilon}_j$.

Vangeneugden *et al.* (2010) approximate $\mathsf{Var}(\boldsymbol{Y}_j)$ by using a first-order Taylor series expansion of the variance function around $\theta_j = 0$. Implicitly, this assumes that $P(y_{ji} = 1|\theta_j) = P(y_i = 1)$, which defines the expected item score (Section 3.1). Further, $\boldsymbol{Y}_j = \boldsymbol{Y}_{j'}$, $j \neq j'$ and $\boldsymbol{Y}_j = \boldsymbol{Y}_i$, where $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, is the vector of expected item scores. Consequently,

$$\mathsf{Var}(\boldsymbol{Y}_i) = \mathsf{Var}(\boldsymbol{Y}_j) \cong \mathsf{Var}(\boldsymbol{\mu}_j) + \mathsf{Var}(\boldsymbol{\varepsilon}_j), \quad \text{where} \tag{19}$$

$$\mathsf{Var}(\boldsymbol{\mu}_j) \cong \left(\frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\eta}_j}\frac{\partial \boldsymbol{\eta}_j}{\partial \boldsymbol{\theta}_j}\,|_{\boldsymbol{\theta}_j=0}\right)\boldsymbol{D}\left(\frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\eta}_j}\frac{\partial \boldsymbol{\eta}_j}{\partial \boldsymbol{\theta}_j}\,|_{\boldsymbol{\theta}_j=0}\right)', \tag{20}$$

$$\mathsf{Var}(\boldsymbol{\varepsilon}_j) \cong \boldsymbol{A}_j^{\frac{1}{2}}\boldsymbol{A}_j^{\frac{1}{2}}. \tag{21}$$

The fact that $\boldsymbol{R}_j$ and $\boldsymbol{\Xi}_j$ disappear is a result of conditional independence and the assumption of no overdispersion in the 2PL model. From (2), it is easy to deduce that the expected sum score merely is the sum over all expected item scores. Hence, the variance of S_T is just the sum of all components in $\mathsf{Var}(\boldsymbol{Y}_j)$, i.e.,

$$\mathsf{Var}(S_T) = \sum \mathsf{Var}(\boldsymbol{Y}_j) = \sum \mathsf{Var}(\boldsymbol{\mu}_j) + \sum \mathsf{Var}(\boldsymbol{\varepsilon}_j).$$

Using the classical definition of reliability, i.e., the proportion of true to observed variance, we obtain:

$$\rho_{iA} = \frac{\mathsf{Var}(\mu_i)}{\mathsf{Var}(\mu_i) + \mathsf{Var}(\varepsilon_i)} \quad \text{and} \quad \rho_{sA} = \frac{\mathsf{Var}(\sum_i \boldsymbol{\mu}_{ji})}{\mathsf{Var}(\sum_i \boldsymbol{\mu}_{ji}) + \mathsf{Var}(\sum_i \boldsymbol{\varepsilon}_{ji})}, \tag{22}$$

as reliability measures for the expected item and test scores, respectively. $\mathsf{Var}(\mu_i)$ and $\mathsf{Var}(\varepsilon_i)$ correspond to the i^{th} diagonal elements of $\mathsf{Var}(\boldsymbol{\mu}_j)$ and $\mathsf{Var}(\boldsymbol{\varepsilon}_j)$, respectively. Equivalent expressions for the 1PL model follow when $\alpha_i = 1$.

It is assumed that variance estimation comes after estimation of other model parameters. As such, the assumptions made in this section apply to variance estimation only.

We acknowledge alternative approximations for reliability measures based on manifest correlations (Dimitrov, 2003). Nevertheless, the simplicity and easy-to-follow nature of our approximations make them a valuable addition to the existing methods.

4.2.1 Illustration For 1PL Model

This section illustrates computation of (19) and (22) for 1PL, a step that also allows the exploration of differences in latent and manifest correlations.

Define the 1PL model as (18), where $\alpha_i = 1$. It follows that:

$$\eta_{ji} = \theta_j - \beta_i, \qquad \frac{\partial \eta_{ji}}{\partial \theta_j}|_{\theta_j=0} = 1, \qquad \frac{\partial \mu_{ji}}{\partial \eta_{ji}}|_{\theta_j=0} = \frac{\exp(\beta_i)}{[1+\exp(\beta_i)]^2} = v(\mu_{ji})|_{\boldsymbol{\theta}_j=0}. \tag{23}$$

Assume a test with only two items, then,

$$\frac{\partial \boldsymbol{\mu}_{ji}}{\partial \boldsymbol{\eta}_{ji}}|_{\theta_j=0} = \boldsymbol{A}_j = \begin{bmatrix} v_{ji}(0) & 0 \\ 0 & v_{ji'}(0) \end{bmatrix}, \qquad \boldsymbol{D} = \begin{bmatrix} \sigma_\theta^2 \end{bmatrix}, \qquad \frac{\partial \boldsymbol{\eta}_{ji}}{\partial \theta_j}|_{\boldsymbol{\theta}_j=0} = 1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{24}$$

$$\mathsf{Var}(\boldsymbol{\mu}_j) = \begin{bmatrix} v_{ji}^2(0)\sigma_\theta^2 & v_{ji}(0)v_{ji'}(0)\sigma_\theta^2 \\ v_{ji'}(0)v_{ji}(0)\sigma_\theta^2 & v_{ji'}^2(0)\sigma_\theta^2 \end{bmatrix}, \qquad \mathsf{Var}(\boldsymbol{\varepsilon}_j) = \begin{bmatrix} v_{ji}(0) & 0 \\ 0 & v_{ji'}(0) \end{bmatrix},$$

$$\rho_{i_{A1}} = \frac{v_{ji}(0)\sigma_\theta^2}{1 + v_{ji}(0)\sigma_\theta^2}, \qquad \rho_{s_{A1}} = \frac{\sigma_\theta^2\left[v_{ji}(0) + v_{ji'}(0)\right]}{1 + \sigma_\theta^2\left[v_{ji}(0) + v_{ji'}(0)\right]},$$

where $v_{ji}(0) = v(\mu_{ji})|_{\boldsymbol{\theta}_j=0}$.

Manifest versus Latent Correlation

The latent correlation measure presented in Section 3.3 is obviously appealing and easy to obtain, and one can be tempted to use it as a reliability measure. We study the relationship between the reliability measures based on latent and manifest correlation.

Consider $\rho_{i_{A1}}$ and $\rho_{s_{A1}}$ at their maximum possible values, which are easily obtained by realizing that the maximum value for $v_{ji}(0) = 0.25$. A comparison of these to the latent reliability measures in (10), reveals the following relationship;

$$\rho_{i_{A1}} = \frac{\sigma_\theta^2}{4 + \sigma_\theta^2} < \rho_{i_{l1}} = \frac{\sigma_\theta^2}{\frac{\pi^2}{3} + \sigma_\theta^2}, \tag{25}$$

$$\rho_{s_{A1}} = \frac{\sigma_\theta^2}{2 + \sigma_\theta^2} < \rho_{s_{l1}} = \frac{\sigma_\theta^2}{\frac{\pi^2}{6} + \sigma_\theta^2}. \tag{26}$$

Latent correlation based score reliability is always greater than its manifest correlation based counterpart, hence if scientific interest is on reliability of observed scores, great caution has to be exercised in using latent correlation based reliability measures.

Note that the above statement does not imply that manifest correlation is always better than latent correlation or, for that matter, vice versa. Rather, when interest genuinely is in latent correlation,

then it is fair to say that manifest correlation, when viewed as an approximation to latent correlation, is attenuated because of measurement error. In the reverse case, when interest is placed in manifest correlation, latent correlation can be considered disattentuated because it fails to accommodate this very measurement error.

4.2.2 Illustration For 2PL Model

Similar to Section (4.2.1) we illustrate computation of (19) and (22) for the 2PL model, and further explore the relationship between latent and manifest correlations.

Define the 2PL model as (18), it follows that,

$$\eta_{ji} = \alpha_i(\theta_j - \beta_i), \quad \frac{\partial \eta_{ji}}{\partial \theta_j}|_{\theta_j=0} = \alpha_i \quad \frac{\partial \mu_{ji}}{\partial \eta_{ji}}|_{\theta_j=0} = \frac{\exp(\alpha_i\beta_i)}{[1+\exp(\alpha_i\beta_i)]^2} = v(\mu_{ji})|_{\boldsymbol{\theta}_j=0}. \tag{27}$$

Again consider a test with only two items, then,

$$\frac{\partial \boldsymbol{\mu}_{ji}}{\partial \boldsymbol{\eta}_{ji}}|_{\theta_j=0} = A_i = \begin{bmatrix} v_{ji}(0) & 0 \\ 0 & v_{j'i}(0) \end{bmatrix}, \quad D = \begin{bmatrix} \sigma_\theta^2 \end{bmatrix}, \qquad \frac{\partial \boldsymbol{\eta}_{ji}}{\partial \theta_j}|_{\boldsymbol{\theta}_j=0} = \begin{bmatrix} \alpha_i \\ \alpha_{i'} \end{bmatrix}, \tag{28}$$

$$\mathsf{Var}(\boldsymbol{\mu}_j) = \begin{bmatrix} \alpha_i^2 v_{ji}^2(0)\sigma_\theta^2 & \alpha_{i'}\alpha_i v_{ji}(0)v_{ji'}(0)\sigma_\theta^2 \\ \alpha_{i'}\alpha_i v_{ji'}(0)v_{ji}(0)\sigma_\theta^2 & \alpha_{i'}^2 v_{ji'}^2(0)\sigma_\theta^2 \end{bmatrix}, \qquad \mathsf{Var}(\boldsymbol{\varepsilon}_j) = \begin{bmatrix} v_{ji}(0) & 0 \\ 0 & v_{ji'}(0) \end{bmatrix},$$

$$\rho_{i_{A2}} = \frac{v_{ji}(0)\alpha_i^2\sigma_\theta^2}{1 + v_{ji}(0)\alpha_i^2\sigma_\theta^2}, \qquad \rho_{s_{A2}} = \frac{\sigma_\theta^2\left[v_{ji}(0)\alpha_i + v_{ji'}(0)\alpha_{i'}\right]^2}{\left[v_{ji}(0) + v_{ji'}(0)\right] + \sigma_\theta^2\left[v_{ji}(0)\alpha_i + v_{ji'}(0)\alpha_{i'}\right]^2},$$

where $v_{ji}(0) = v(\mu_{ji})|_{\boldsymbol{\theta}_j=0}$.

Manifest versus Latent Correlation.

The following relationship exists between latent and manifest correlation based reliability measures for the 2PL model:

$$\rho_{i_{A2}} = \frac{\alpha_i^2\sigma_\theta^2}{4 + \alpha_i^2\sigma_\theta^2} \quad < \quad \rho_{i_{l2}} = \frac{\alpha_i^2\sigma_\theta^2}{\frac{\pi^2}{3} + \alpha_i^2\sigma_\theta^2}, \tag{29}$$

$$\rho_{s_{A2}} = \frac{\sigma_\theta^2(\alpha_i + \alpha_{i'})^2}{8 + \sigma_\theta^2(\alpha_i + \alpha_{i'})^2} \quad < \quad \rho_{s_{l2}} = \frac{\sigma_\theta^2(\alpha_i + \alpha_{i'})^2}{\frac{2\pi^2}{3} + \sigma_\theta^2(\alpha_i + \alpha_{i'})^2} \tag{30}$$

The relationship between latent and manifest correlations based reliability measures observed in the 1PL model is the same for the 2PL model.

5 Simulation Study

Due to lack of closed-form quantities, the performance of reliability measures based on Taylor series approximations of the variance-covariance matrix in Section 4.2 will be assessed through a simulation study. We will compare these with the exact measures described in Section 3.1. Additionally, the relationship observed between manifest and latent correlation based measures in Sections 4.2.1 and 4.2.2, will be studied for more than two items.

5.1 Design

5.1.1 Based on Theoretical Values

Measuring tools calibrated under the 1PL and 2PL models are considered. Each has 24 items, whose difficulty values (β_i) are generated from a uniform distribution within the range $[-4, 4]$, the discrimination parameter values (α_i) for the 2PL model are sampled from $N(2, 0.64)$, and the number of respondents is set to 400. To assess the quality of our approximations at different levels of reliability, three values for the variance of person trait scores (θ_j), which influences reliability, were considered, i.e., $\sigma_\theta^2 = 0.25$, 1, or 4, implying that each model will produce three measuring tools that will be responded to by three different sets of individuals. With these values, the score Y_{ji} for item i by person j is generated from a Bernoulli(π_{ji}), where

$$\pi_{ji} = \begin{cases} \frac{\exp(\theta_j - \beta_i)}{1+\exp(\theta_j - \beta_i)} & \text{for the 1PL model,} \\ \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1+\exp[\alpha_i(\theta_j - \beta_i)]} & \text{for the 2PL model.} \end{cases}$$

For the exact, Taylor series approximation and latent correlation based approximations, both expected item reliability and expected sum score reliability will be computed, while for the Fisher information coefficient, only the expected sum score will be obtained. To concentrate on the performance of the reliability measures, the models are not fit. Rather, we assume that the generated samples represent the true population. As such, simulated values are plugged into all the formulas and the integrals in Section 3.1 are obtained by simply averaging over the relevant quantities. This helps to eliminate behavior that may be observed due other issues, like lack of model convergence resulting in poor estimates.

5.1.2 Based on Model Estimates

While the generated values were used to compute all reliability measures in Section 5.2.1, in this section the models under which the data was generated were fitted. Model estimates were thus plugged in to compute all reliability measures except for the exact reliability where the generated values were used. Further, both the number of items and persons were varied, with 6, 12, and 24, and 30, 100, and 200 considered for the items and number of persons, respectively. A total of 50 datasets were generated under each setting and only $\sigma_\theta^2 = 1$ was used. All other settings were similar to those in Section 5.2.1.

5.2 Simulation Results

5.2.1 Based on Theoretical Values

Table 1 indicates that the reliability of the expected sum, estimated using the newly introduced,Taylor series approximation is very similar to the exact reliability. Practically, using one in place of the other should lead to virtually the same substantive conclusions. In addition, Taylor series approximations are not computationally intensive, because they do not involve the evaluation of integrals with no closed forms, and quantities for approximating the variance-covariance matrix follow directly from model estimation. Clearly, these strengths make the Taylor series approximation a valuable addition to the theory of reliability measurement.

Table 1: *Expected sum score reliability computed based on theoretical values for 1PL and 2PL models, using various approximation methods. ρ_{S_A} gives the reliability estimated using the Taylor series approximation, ρ_f, uses the Fisher information measure, ρ_{S_l} is the latent variance of logistic regression, i.e., $\pi^2/3$, is used as variance of error (σ^2_ε) and ρ_S is the exact sum score reliability. σ^2_θ is the variance of person trait.*

		Measure of Reliability			
σ^2_θ	Model	ρ_{S_A}	ρ_f	ρ_{S_l}	ρ_S
0.25	1PL	0.461	-0.120	0.646	0.480
	2PL	0.687	0.615	0.889	0.704
1	1PL	0.774	0.676	0.879	0.778
	2PL	0.898	0.881	0.970	0.895
4	1PL	0.932	0.887	0.967	0.928
	2PL	0.972	0.891	0.992	0.963

Also revealed in Table 1 is a possible drawback for the commonly used Fisher information based measure: when $\sigma^2_\theta = 0.25$, reliability is negative and does not have a meaningful interpretation. Further, in other cases, like for the 2PL model with $\sigma^2_\theta = 1$, it overestimates the exact reliability while in the 1PL model with $\sigma^2_\theta = 0.25$, it underestimates the exact reliability. In general, Fisher information based reliability will not always yield a truthful picture of the reliability of the expected sum.

In line with theory, latent correlation based reliability is consistently greater than the manifest correlation based measures (Taylor approximation and exact). If reliability at the logit scale is of interest, latent correlations are meaningful; otherwise, they should be avoided.

Expected item score reliability is useful when creating item banks, as it can help in choosing items that are reliable. Results in Table 2 emphasize why latent correlation based reliability may not be best suited to be used for such a process, especially for the 1PL model, because not only is it always greater than the manifest correlation based version, but it is also constant for all items. The Taylor series based versions approximate the exact expected item score reliability closely for the 1PL model, such that decision making based on the former is likely to reflect the decisions that would result from using the latter. For the 2PL model, Taylor series approximation is not good for some items like 3, 4, 8, 9, 11, 13, 15, and 21, which greatly underestimate the exact reliability, suggesting the need to improve the approximations if they are to be used for item reliability in the 2PL model. Possible considerations include expansion of Taylor series around $\theta_j = \widehat{\theta}_j$, the maximum likelihood estimate of the trait score, instead of $\theta_j = 0$, but this is beyond the scope of the current work. Latent reliability in 2PL increases with discriminative power, i.e., items with high discriminative power have high reliability, regardless of the difficulty level. Results for cases with $\sigma^2_\theta = 0.25, 4$ are presented in the Appendix.

Table 2: *Item reliability for 1PL and 2PL, where person trait variance $\sigma_\theta^2 = 1$. ρ_{S_A} is the item reliability obtained using the Taylor series approximation, ρ_{S_l}, is the latent correlation based item reliability and ρ_S is the exact item reliability and finally, β and α are the simulated item difficulty and discrimination parameter values.*

	1PL				2PL				
Item	β	ρ_{i_l}	ρ_{i_A}	ρ_i	β	α	ρ_{i_l}	ρ_{i_A}	ρ_i
1	1.221	0.233	0.150	0.168	0.444	3.441	0.782	0.634	0.608
2	-0.670	0.233	0.183	0.173	3.785	1.210	0.308	0.015	0.054
3	1.205	0.233	0.151	0.169	0.974	3.731	0.809	0.258	0.608
4	0.162	0.233	0.199	0.185	-1.389	2.233	0.602	0.170	0.352
5	0.780	0.233	0.177	0.179	-0.490	2.043	0.559	0.450	0.398
6	-3.850	0.233	0.020	0.047	3.008	1.789	0.493	0.014	0.135
7	3.330	0.233	0.032	0.071	-0.623	2.680	0.686	0.489	0.487
8	-0.197	0.233	0.198	0.182	2.971	1.960	0.538	0.011	0.150
9	2.373	0.233	0.072	0.119	1.244	2.602	0.673	0.197	0.465
10	2.037	0.233	0.093	0.135	-0.711	2.517	0.658	0.437	0.458
11	0.423	0.233	0.193	0.184	-1.353	2.350	0.627	0.175	0.372
12	-1.702	0.233	0.115	0.136	-2.184	0.590	0.096	0.056	0.061
13	0.948	0.233	0.168	0.176	-2.332	2.241	0.604	0.026	0.267
14	2.979	0.233	0.044	0.088	-0.061	2.757	0.698	0.654	0.521
15	3.873	0.233	0.020	0.049	-2.552	2.274	0.611	0.015	0.251
16	0.288	0.233	0.197	0.185	-0.519	2.020	0.554	0.440	0.393
17	0.042	0.233	0.200	0.185	-1.281	2.046	0.560	0.209	0.336
18	-1.461	0.233	0.133	0.146	0.118	0.395	0.045	0.038	0.040
19	0.480	0.233	0.191	0.184	-0.156	2.106	0.574	0.519	0.423
20	-3.227	0.233	0.035	0.069	-1.769	1.286	0.334	0.123	0.177
21	3.741	0.233	0.022	0.054	3.688	2.413	0.639	0.001	0.053
22	0.096	0.233	0.200	0.185	-1.501	2.011	0.551	0.152	0.309
23	-2.821	0.233	0.050	0.086	1.581	1.981	0.544	0.136	0.342
24	3.780	0.233	0.021	0.052	0.630	1.619	0.443	0.338	0.330

5.2.2 Based on Model Estimates

The pattern observed in the estimation of expected sum reliability when theoretical values are used, is carried forward to the setting where model-based estimates are used. Latent reliability is always higher than the exact and Taylor approximated reliability. The disadvantage of Fisher information is also prominent in settings with only few items, e.g., 6, where reliability is either negative or very small compared to the true reliability. In general, Taylor approximated sum reliability is the closest to the exact reliability, although the difference is noticeable for 2PL models with 6 items. The number of items may affect the quality of approximation. This can be seen in Table 3.

On the other hand, item reliability seems to heavily depend on the quality of estimates used. Results from the 2PL model for Taylor approximation and latent reliability are presented in Figures 1 and 2, respectively. Where the approximated reliability matches the exact, points are expected to lie along the diagonal line. This is hardly the case for latent reliability and Taylor approximated reliability for

Table 3: *Expected sum score reliability estimated for 1PL and 2PL models, using various approximation methods. ρ_{S_A} gives the reliability estimated using the Taylor series approximation, ρ_f, uses the Fisher information measure, ρ_{S_l} is the latent reliability and ρ_S is the exact sum score reliability. Inum and Pnum represent the number of items and persons used for each model.*

				Measure of Reliability			
Model	Inum	Pnum	$\widehat{\sigma}^2_\theta$	ρ_S	ρ_{S_A}	ρ_l	ρ_{S_f}
1PL	30	6	1.25	0.49	0.52	0.66	-0.03
		12	1.16	0.63	0.65	0.80	0.44
		24	1.01	0.76	0.75	0.88	0.66
	100	6	1.06	0.49	0.52	0.65	-0.01
		12	0.96	0.63	0.64	0.78	0.40
		24	0.99	0.76	0.77	0.88	0.67
	200	6	1.04	0.49	0.52	0.65	0.00
		12	0.98	0.63	0.65	0.78	0.42
		24	1.01	0.76	0.77	0.88	0.68
2PL	30	6	1.00	0.59	0.66	1.00	0.10
		12	1.00	0.80	0.85	0.96	0.79
		24	1.00	0.89	0.89	0.97	0.91
	100	6	1.00	0.58	0.47	0.90	0.40
		12	1.00	0.80	0.83	0.91	0.73
		24	1.00	0.89	0.91	0.97	0.90
	200	6	1.00	0.58	0.49	0.92	0.25
		12	1.00	0.80	0.82	0.92	0.74
		24	1.00	0.89	0.89	0.97	0.90

models with 6 items. This observation is in line with the noticeable differences between exact and Taylor approximation for the expected sum reliability. Even when the estimates are biased (relative bias $> 10\%$), Taylor approximated reliability is still relatively closer to the exact reliability. Results for the 1PL model are presented in Figures 4 and 3 and they convey a similar story.

6 Analysis of Case Study

The application of the reliability measures introduced in Section 4.2 is demonstrated through the analysis of the two datasets described in Section 2. Using various datasets that are measuring different attributes brings out just how broadly these measures can be used to assess reliability of measuring tools across different fields of research. Both 1PL and 2PL models were fitted using the NLMIXED procedure in SAS, which employs adaptive Gauss-Hermite quadrature to compute the integrals and ultimately the maximum likelihood based parameter estimates.

Results for the LSAT6 data are presented in Table 4. These generally indicate low reliability both at

Table 4: *Results from the analysis of the LSAT6 data. $\widehat{\beta}$ is the estimate of item difficulty; $\widehat{\alpha}$, is the discrimination parameter estimate; ρ_{i_A} indicates the Taylor series approximated expected item score reliability; ρ_{i_l} corresponds to the latent correlation based counterpart; ρ_s gives the expected sum score reliability of the corresponding item reliability; ρ_f is the Fisher information based reliability measure.*

	1PL			2PL			
item	$\widehat{\beta}$	ρ_{i_A}	ρ_{i_l}	$\widehat{\beta}$	$\widehat{\alpha}$	ρ_{i_A}	ρ_{i_l}
1	-2.730	0.032	0.148	-3.359	0.826	0.036	0.172
2	-0.999	0.101	0.148	-1.370	0.723	0.094	0.137
3	-0.240	0.123	0.148	-0.280	0.891	0.163	0.194
4	-1.306	0.087	0.148	-1.866	0.688	0.074	0.126
5	-2.099	0.053	0.148	-3.126	0.657	0.042	0.116
ρ_S		0.304	0.464			0.312	0.465
ρ_f		-1.343			-1.240		

the expected item score level and at the expected sum score level. With a five-item measuring tool, this is not surprising as these may not be enough to capture all relevant information. In addition, the negative estimates for the difficulty parameters for both the 1PL and 2PL models indicate a low difficulty level for all items and the small person trait variance of 0.570 suggests that examinees exhibit similar levels of ability, a scenario that is well known to be less informative. Latent correlation based reliability is larger than the Taylor series approximation and in this case this would lead to different conclusions regarding reliability of the expected sum score. The Fisher information based measure is not useful in this case given that it is negative, which can also be attributed to the relatively small number of items.

The questionnaire for Verbal Aggression data has 24 items, which can be considered of average length and according to Table 5, the expected sum score reliability is high: above 0.85 for all measures. Further, the item difficulty estimates for 1PL and 2PL models have maxima, 2.976 and 2.439, and minima, -1.748 and -1.212, respectively, suggesting a mixture of high and low difficulty level items. The person trait variance of 1.919 suggests a cross-section of persons with varying abilities, forming a desirable scenario to achieve high reliability. Again, the latent correlation based reliability is higher than the manifest correlation based counterpart, although in this case, similar conclusions regarding reliability of both the expected item and the sum score would be reached, regardless of the measure of reliability used.

7 Concluding Remarks

Beyond doubt, reliability measures based on manifest correlations are of considerable importance in IRT. The reason for their relatively rare use can be largely attributed to lack of efficient means of estimation given that marginalizing the joint distribution of normal random effects, combined with binary data distributions is computationally challenging.

This paper has outlined a procedure for approximating reliability measures based on manifest correlations, and illustrated their application for 1PL and 2PL models for both expected item score and expected sum score. We have further explored the relationship between latent and manifest

correlations based reliability measures, where it was shown that latent correlation based reliability measures are always greater than their manifest correlation based counterparts. Hence, using one in place of the other should be avoided. A simulation study to assess the performance of the newly introduced Taylor series based reliability measures indicated that they give a true reflection of the exact reliability, especially at the expected sum score level. In comparison to Fisher information based sum score reliability, Taylor series based approximations, perform consistently better, including in the cases where the Fisher information based measure gives negative values, which are not meaningful.

Taylor series based reliability measures do not involve evaluation of integrals with no closed forms. Rather, they use quantities that are easily obtained during model estimation; computation can be handled by most standard statistical software tools. Thus, they represent a less computationally intensive, readily available solution to obtaining reliability of either item score or sum score, whichever truly reflects the required reliability. However, the quality of reliability estimates heavily relies on the quality of model estimated parameters. For example, results from poorly converged models may not reflect the true reliability.

Generally, our findings are useful and relevant for practice and expand on the available tools for measuring reliability. When studying reliability or generalizability, manifest correlation is a more intuitive measure, as it captures the correlation between what is actually observed, and not what happens at the level of a latent construct.

Acknowledgments

The authors gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy). For the computations, simulations and data processing, we used the infrastructure of the VSC—Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government - department EWI.

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: John Wiley & Sons.

Bock, R.D., & Lieberman M. (1970). Fitting a response curve model for dichotomously scored items. *Psychometrika*, **35**, 179-198.

Briggs, D.C., and Wilson, M. (2007). Generalizability in item response modelling. *Journal of Education Measurement*, **44**, 131–155.

Cronbach, L.J., & Shavelson, R.J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, **64**, 391-418.

Culligan, B. (2008). Estimating word difficulty using Yes/No tests in an IRT framework and its application for pedagogic objectives. Unpublished dissertation. Tokyo: Temple University Japan.

De Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guilford Press.

De Boeck, P. (2009). Random Item Item Response Theory Models. *Technical Report*.

De Boeck, P. & Wilson, M. (2004).(Eds.) *Explanatory Item Response Models: A Generalized Linear and non-linear Approach*. New York: Springer.

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

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 package. *Journal of Statistical Software*, **20**, 1–18.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Press.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions, Vol. 2.* Boston: Houghton-Mifflin.

Laenen, A. (2008). *Psychometric Validation of Continuous Rating Scales from Complex Data*. Unpublished PhD Thesis. Hasselt University, Belgium.

Lord, F.M. (1980). Unbiased estimators of ability parameters of their variance and of their parallel-forms reliability. *Psychometrika*,**48**, 233–245.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, **1**, 30–46.

Mesbah, M., Cole, B.F., & Ting Lee, M.-L. (2002). *Statistical Methods for Quality of Life Studies: Design, Measurement, and Analysis*; Questionnaire reliability under the Rasch model, pp. 155–166. New York: Kluwer Academic Publishers.

Molenberghs G. and Verbeke G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., & Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

O'Brien, R.M. (1995). Generalizability coefficients are reliability coefficients. *Quality and Quantity*, **29**, 421–428.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens P. (2003). A non-linear mixed model framework for item response theory. *Pyschological Methods*, **8**, 185–205.

Rodríguez, G. & Elo, I. (2003). Intra-class correlation in random effects models for binary data. *The Stata Journal*, **3**, 32–46.

Schaeffer, G.A., Carlson, R.E., and Matas, R.L. (1986). Assessing the reliability of criterion-referenced measures used to evaluate health-education programs. *Evaluation Review*, **10**, 115–125.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**, 72–101.

Van der Linden, W.J. & Hambleton, R.K. (Eds.) (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, **25**, 13–30.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, **61**, 295–304.

Vangeneugden, T., Molenberghs, G., Laenen, A., Geys, H., Beunckens, C., and Sotto, C.(2010) Marginal correlationin longitudinal binary data based on generalized linear mixed models. *Communications in Statistics - Theory and Methods*, **39**, 3540–3557.

Vangeneugden, T., Molenberghs, G., Verbeke, G., & Demétrio, C. (2011). Marginal correlation from an extended random-effects model for repeated and overdispersed counts. *Journal of Applied Statistics*, **38**, 215–232.

Van Leeuwen, D.M., Barnes, M.D., & Pase, M. (1998). Generalizability theory: A unified approach to assessing the dependability (reliability) of measurements in the health sciences. *Journal of Outcome Measurement*, **2**, 302–325.

Vansteelandt, K. (2000). *Formal models for Contextualized Personality Psychology. Unpublished doctoral dissertation*, KU Leuven, Belgium.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York:Springer.

Zeger, S.L., Liang, K.-Y., & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

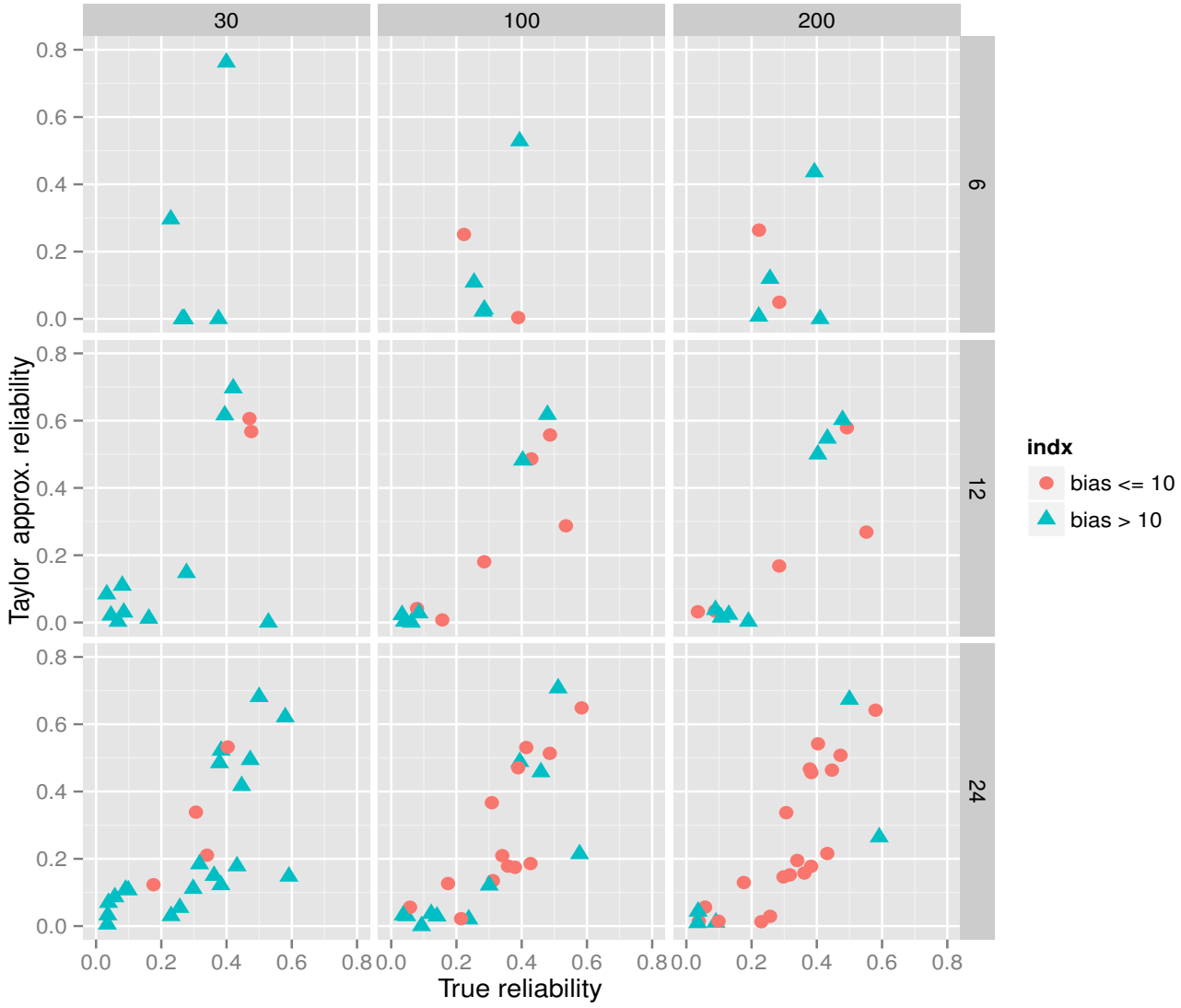

Figure 1: *Scatter plot for exact vs Taylor approximated item reliability from 2PL model. On the upper and right margins are the number of persons (30, 100, 200) and items (6, 12, 24), respectively. The variable* 'indx' *indicates whether both the item difficulty and discrimination parameter estimates used in computing reliability had relative bias of greater than 10% (bias>10) or below (bias >= 10).*

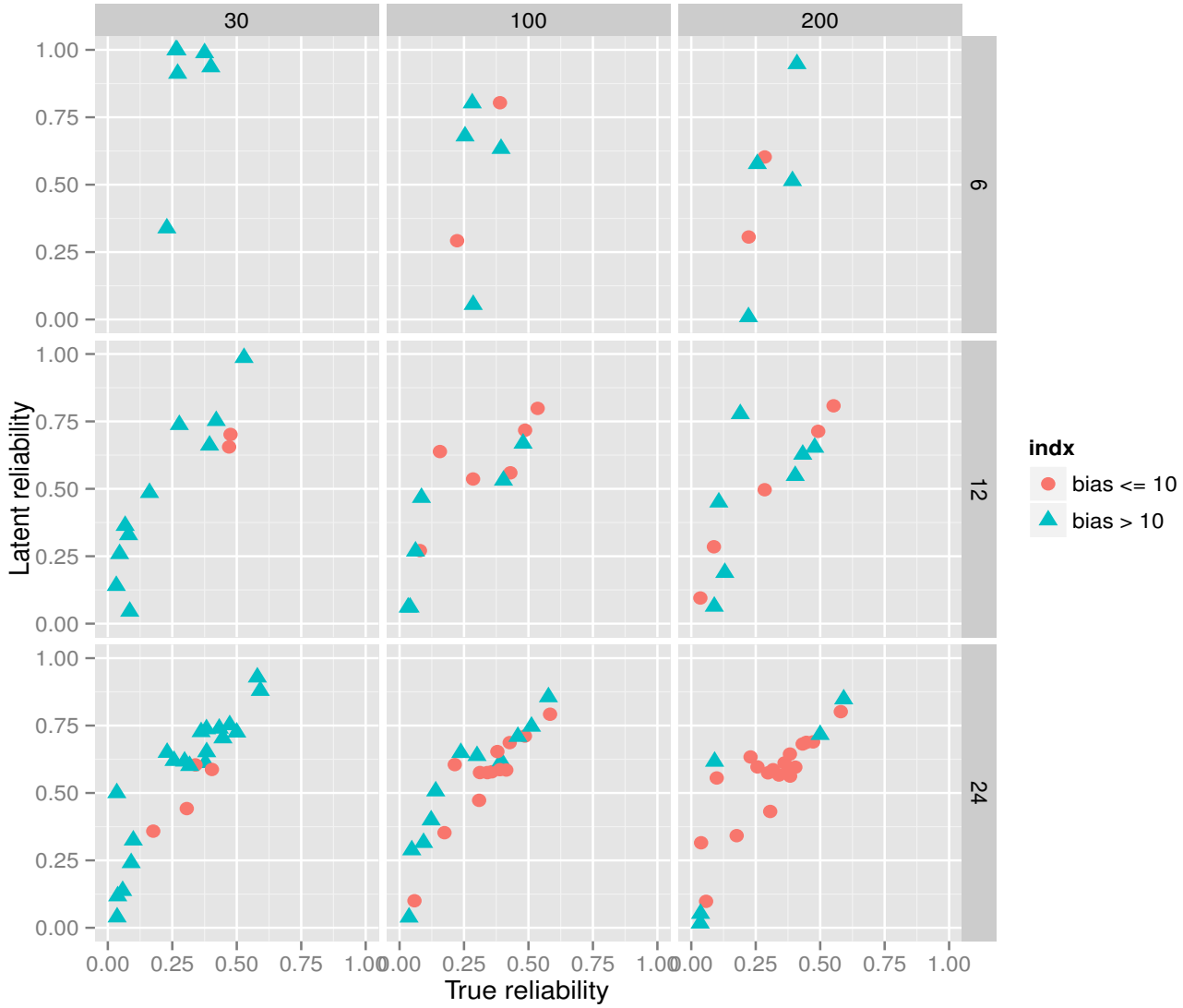

Figure 2: *Scatter plot for exact vs latent item reliability from 2PL model. On the upper and right margins are the number of persons (30, 100, 200) and items (6, 12, 24), respectively. The variable* 'indx' *indicates whether both the item difficulty and discrimination parameter estimates used in computing reliability had relative bias of greater than 10% (bias>10) or below (bias >= 10).*

Table 5: *Results from the analysis of the Verbal Aggression Data. $\widehat{\beta}$ is the estimate of item difficulty; $\widehat{\alpha}$, is the discrimination parameter estimate; ρ_{i_A} indicates the Taylor series approximated expected item score reliability; ρ_{i_l} corresponds to the latent correlation based counterpart; ρ_s gives the expected sum score reliability of the corresponding item reliability; ρ_f is the Fisher information based reliability measure.*

	1PL			2PL			
item	$\widehat{\beta}$	ρ_{i_A}	ρ_{i_l}	$\widehat{\beta}$	$\widehat{\alpha}$	ρ_{i_A}	ρ_{i_l}
1	-1.221	0.252	0.368	-0.886	1.372	0.249	0.364
2	-0.565	0.307	0.368	-0.387	1.551	0.355	0.422
3	-0.080	0.324	0.368	-0.062	1.373	0.320	0.364
4	-1.748	0.195	0.368	-1.212	1.483	0.211	0.400
5	-0.707	0.298	0.368	-0.476	1.601	0.357	0.438
6	-0.012	0.324	0.368	-0.012	1.285	0.292	0.334
7	-0.529	0.309	0.368	-0.510	0.891	0.159	0.194
8	0.686	0.299	0.368	0.479	1.436	0.315	0.385
9	1.527	0.220	0.368	1.438	0.933	0.125	0.209
10	-1.082	0.266	0.368	-0.877	1.148	0.205	0.286
11	0.349	0.318	0.368	0.223	1.628	0.391	0.446
12	1.044	0.270	0.368	0.935	0.996	0.167	0.232
13	-1.221	0.252	0.368	-0.786	1.720	0.326	0.473
14	-0.389	0.316	0.368	-0.230	2.351	0.563	0.627
15	0.871	0.285	0.368	0.606	1.451	0.304	0.390
16	-0.872	0.285	0.368	-0.602	1.512	0.319	0.410
17	0.057	0.324	0.368	0.023	2.030	0.507	0.556
18	1.482	0.225	0.368	0.963	1.656	0.278	0.454
19	0.211	0.322	0.368	0.173	1.116	0.236	0.274
20	1.504	0.222	0.368	1.094	1.361	0.218	0.360
21	2.976	0.081	0.368	2.439	1.140	0.067	0.283
22	-0.707	0.298	0.368	-0.510	1.401	0.302	0.373
23	0.384	0.316	0.368	0.261	1.471	0.343	0.397
24	2.000	0.168	0.368	1.571	1.209	0.142	0.307
ρ_S		0.900	0.933			0.908	0.936
ρ_f		0.859			0.863		

Manifest Versus Latent Correlation Functions in Item Response Theory

Elasma Milanzi[1] **Geert Molenberghs**[1,2]

Ariel Alonso[1] **Geert Verbeke**[2,1]

Paul De Boeck[3]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*

[3] *Department of Psychology, Higher Cognition and Individual Differences, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium & Universiteit van Amsterdam, the Netherlands*

Supplementary Materials

Additional Results From The Simulation Study

Table 6: *Item reliability for 1PL and 2PL, where person trait variance, $\sigma^2_\theta = 4$. ρ_{S_A} is the item reliability obtained using the Taylor series approximation; ρ_{S_l} is the latent correlation based item reliability; ρ_S is the exact item reliability; β and α are the simulated item difficulty and discrimination parameter values.*

	1PL				2PL				
Item	β	ρ_{i_l}	ρ_{i_A}	ρ_i	β	α	ρ_{i_l}	ρ_{i_A}	ρ_i
1	1.221	0.549	0.413	0.407	0.444	3.441	0.935	0.874	0.782
2	-0.670	0.549	0.472	0.398	3.785	1.210	0.640	0.056	0.380
3	1.205	0.549	0.415	0.407	0.974	3.731	0.944	0.582	0.817
4	0.162	0.549	0.498	0.410	-1.389	2.233	0.858	0.451	0.655
5	0.780	0.549	0.463	0.411	-0.490	2.043	0.835	0.766	0.642
6	-3.850	0.549	0.075	0.268	3.008	1.789	0.795	0.055	0.555
7	3.330	0.549	0.118	0.337	-0.623	2.680	0.897	0.793	0.715
8	-0.197	0.549	0.498	0.406	2.971	1.960	0.824	0.043	0.586
9	2.373	0.549	0.238	0.377	1.244	2.602	0.892	0.496	0.736
10	2.037	0.549	0.290	0.389	-0.711	2.517	0.885	0.756	0.700
11	0.423	0.549	0.489	0.411	-1.353	2.350	0.870	0.459	0.672
12	-1.702	0.549	0.343	0.367	-2.184	0.590	0.297	0.191	0.197
13	0.948	0.549	0.446	0.410	-2.332	2.241	0.859	0.097	0.598
14	2.979	0.549	0.155	0.353	-0.061	2.757	0.902	0.883	0.723
15	3.873	0.549	0.074	0.309	-2.552	2.274	0.863	0.058	0.589
16	0.288	0.549	0.495	0.411	-0.519	2.020	0.832	0.758	0.638
17	0.042	0.549	0.500	0.409	-1.281	2.046	0.836	0.514	0.631
18	-1.461	0.549	0.379	0.376	0.118	0.395	0.160	0.135	0.130
19	0.480	0.549	0.486	0.411	-0.156	2.106	0.844	0.812	0.653
20	-3.227	0.549	0.128	0.297	-1.769	1.286	0.668	0.359	0.450
21	3.741	0.549	0.083	0.316	3.688	2.413	0.876	0.003	0.652
22	0.096	0.549	0.499	0.409	-1.501	2.011	0.831	0.418	0.617
23	-2.821	0.549	0.175	0.316	1.581	1.981	0.827	0.386	0.644
24	3.780	0.549	0.080	0.314	0.630	1.619	0.761	0.671	0.585

Table 7: *Item reliability for 1PL and 2PL, where person trait variance, $\sigma_\theta^2 = 0.25$. ρ_{S_A} is the item reliability obtained using the Taylor series approximation; ρ_{S_l} is the latent correlation based item reliability; ρ_S is the exact item reliability; β and α are the simulated item difficulty and discrimination parameter values.*

	1PL				2PL				
Item	β	ρ_{i_l}	ρ_{i_A}	ρ_i	β	α	ρ_{i_l}	ρ_{i_A}	ρ_i
1	1.221	0.071	0.042	0.048	0.444	3.441	0.473	0.303	0.342
2	-0.670	0.071	0.053	0.055	3.785	1.210	0.100	0.004	0.006
3	1.205	0.071	0.042	0.048	0.974	3.731	0.514	0.080	0.286
4	0.162	0.071	0.058	0.060	-1.389	2.233	0.275	0.049	0.097
5	0.780	0.071	0.051	0.055	-0.490	2.043	0.241	0.170	0.169
6	-3.850	0.071	0.005	0.007	3.008	1.789	0.196	0.004	0.009
7	3.330	0.071	0.008	0.012	-0.623	2.680	0.353	0.193	0.220
8	-0.197	0.071	0.058	0.060	2.971	1.960	0.226	0.003	0.008
9	2.373	0.071	0.019	0.025	1.244	2.602	0.340	0.058	0.149
10	2.037	0.071	0.025	0.031	-0.711	2.517	0.325	0.163	0.194
11	0.423	0.071	0.056	0.059	-1.353	2.350	0.296	0.050	0.106
12	-1.702	0.071	0.032	0.036	-2.184	0.590	0.026	0.015	0.016
13	0.948	0.071	0.048	0.053	-2.332	2.241	0.276	0.007	0.025
14	2.979	0.071	0.011	0.016	-0.061	2.757	0.366	0.320	0.277
15	3.873	0.071	0.005	0.007	-2.552	2.274	0.282	0.004	0.017
16	0.288	0.071	0.058	0.060	-0.519	2.020	0.237	0.164	0.164
17	0.042	0.071	0.059	0.061	-1.281	2.046	0.241	0.062	0.099
18	-1.461	0.071	0.037	0.041	0.118	0.395	0.012	0.010	0.010
19	0.480	0.071	0.056	0.059	-0.156	2.106	0.252	0.213	0.194
20	-3.227	0.071	0.009	0.012	-1.769	1.286	0.111	0.034	0.043
21	3.741	0.071	0.006	0.008	3.688	2.413	0.307	0.000	0.001
22	0.096	0.071	0.059	0.061	-1.501	2.011	0.235	0.043	0.078
23	-2.821	0.071	0.013	0.017	1.581	1.981	0.230	0.038	0.078
24	3.780	0.071	0.005	0.008	0.630	1.619	0.166	0.113	0.124

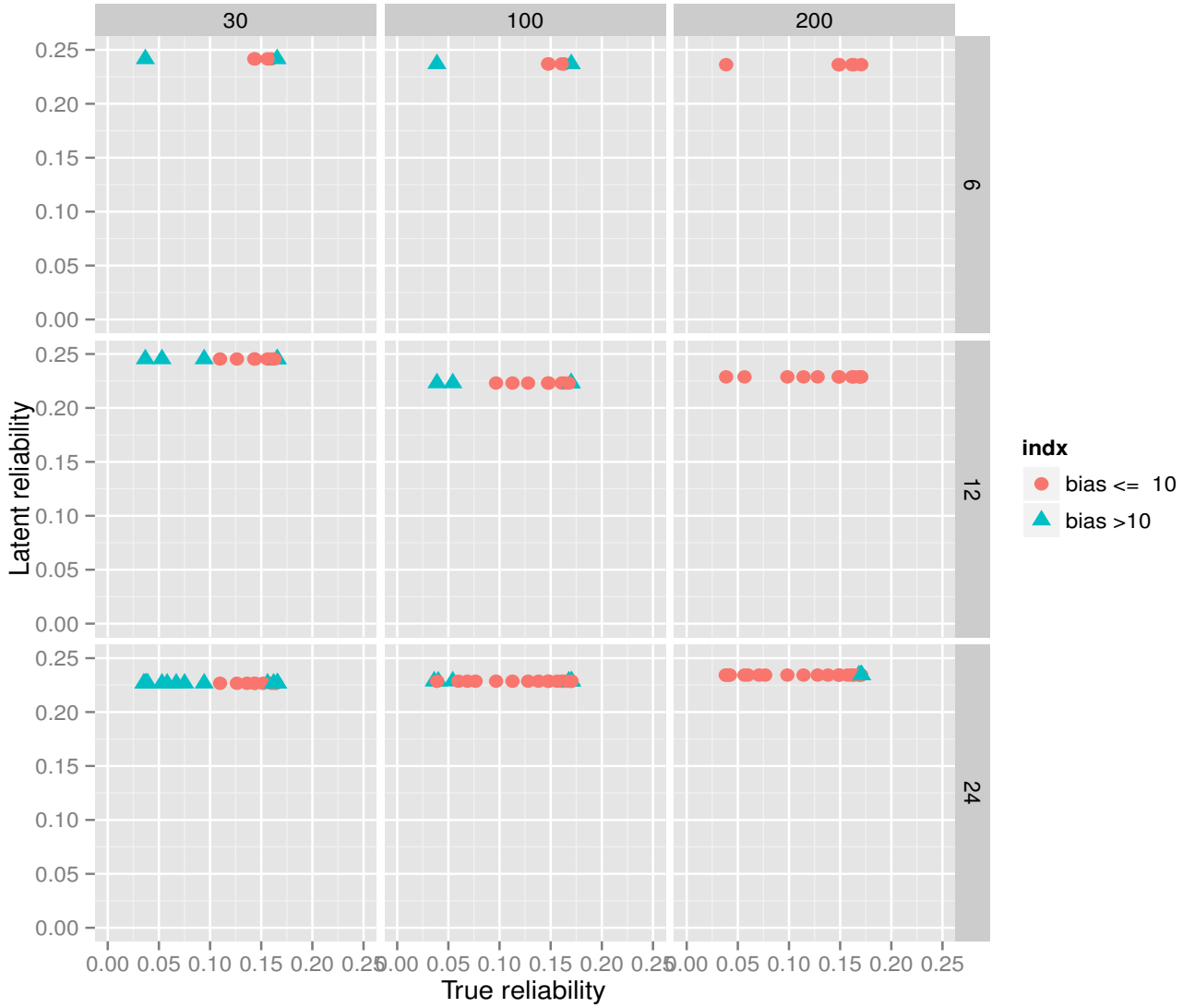

Figure 3: *Scatter plot for exact vs latent item reliability from 1PL model. On the upper and right margins are the number of persons (30, 100, 200) and items (6, 12, 24), respectively. The variable* 'indx' *indicates whether the item difficulty estimate used in computing reliability had relative bias of greater than 10% (bias>10) or below (bias >= 10).*

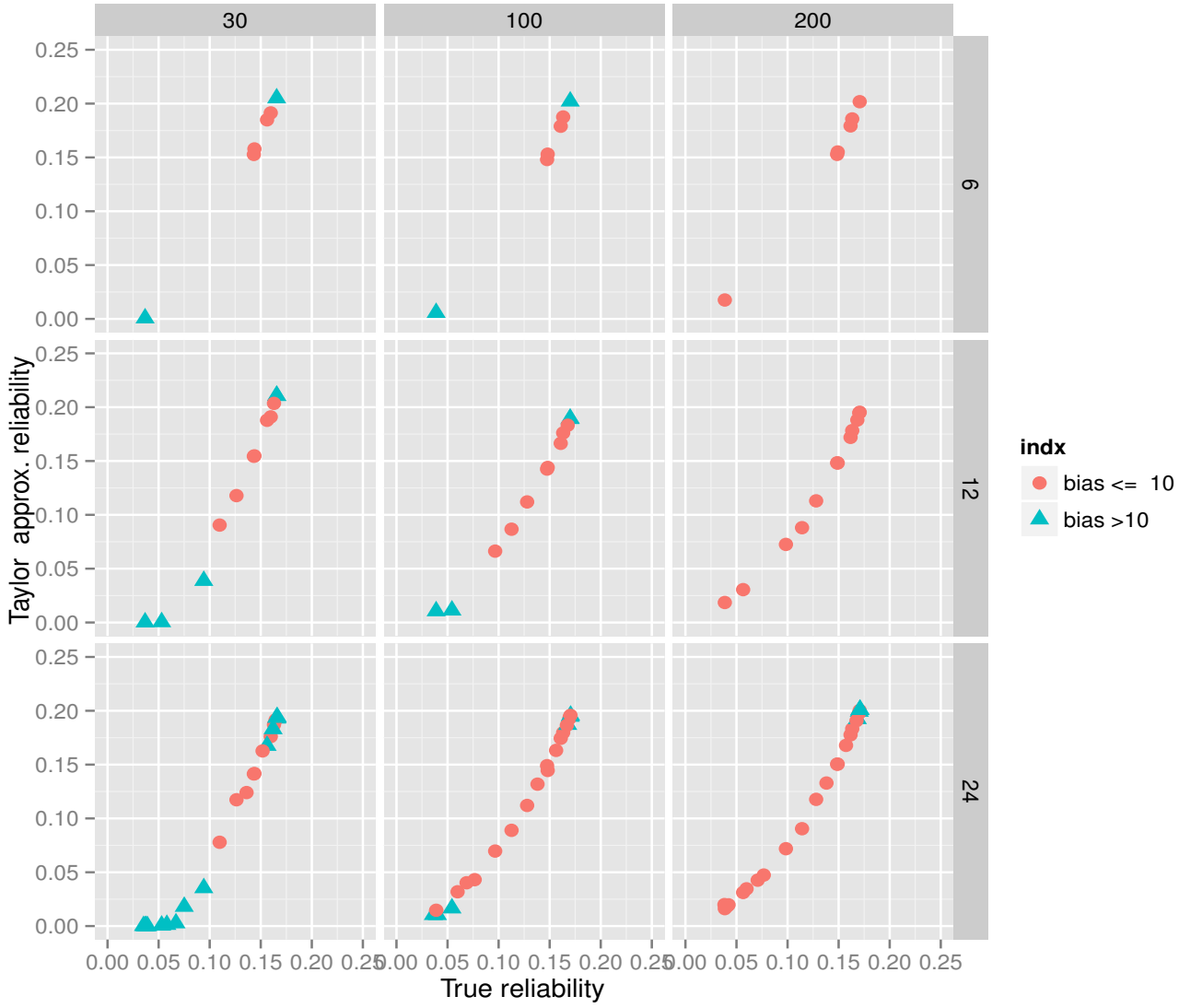

Figure 4: *Scatter plot for exact vs Taylor approximated item reliability from 2PL model. On the upper and right margins are the number of persons (30, 100, 200) and items (6, 12, 24), respectively. The variable* 'indx' *indicates whether the item difficulty estimate used in computing reliability had relative bias of greater than 10% (bias>10) or below (bias >= 10).*