Large Sample Confidence Intervals for Item Response Theory Reliability Coefficients

Educational and Psychological
Measurement

2018, Vol. 78(1) 32–45

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164417713570

journals.sagepub.com/home/epm

Björn Andersson¹ and Tao Xin¹

Abstract

In applications of item response theory (IRT), an estimate of the reliability of the ability estimates or sum scores is often reported. However, analytical expressions for the standard errors of the estimators of the reliability coefficients are not available in the literature and therefore the variability associated with the estimated reliability is typically not reported. In this study, the asymptotic variances of the IRT marginal and test reliability coefficient estimators are derived for dichotomous and polytomous IRT models assuming an underlying asymptotically normally distributed item parameter estimator. The results are used to construct confidence intervals for the reliability coefficients. Simulations are presented which show that the confidence intervals for the test reliability coefficient have good coverage properties in finite samples under a variety of settings with the generalized partial credit model and the three-parameter logistic model. Meanwhile, it is shown that the estimator of the marginal reliability coefficient has finite sample bias resulting in confidence intervals that do not attain the nominal level for small sample sizes but that the bias tends to zero as the sample size increases.

Keywords

reliability, item response theory, confidence intervals, asymptotic variance

In classical test theory, the reliability of a test plays a central role. The reliability is a measure of the consistency of the scores from a test. Reliability, together with the

¹Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China

Corresponding Author:

Björn Andersson, Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, No. 19 Xijiekou Wai Street, Haidian District, Beijing 100875, China.

Email: bjoern.andersson@bnu.edu.cn

concept of item and test information, also plays a role in item response theory (IRT). Two different measures of reliability in IRT are marginal reliability (Cheng, Yuan, & Liu, 2012) and test reliability (Kim & Feldt, 2010; Lord, 1977, 1980). The marginal reliability denotes the ratio of the true score variance to the total variance, expressed with respect to the estimated latent abilities. Test reliability, on the other hand, denotes the ratio of true score variance to total variance expressed with respect to the sum scores. Note that, in the terminology used in this article, both the marginal reliability and the test reliability refer to reliability with regard to a population as a whole and hence are in some sense both “marginal” measures of reliability. Having a measure of the reliability of the scores from a test in the context of IRT is useful since it provides an indication of the overall consistency of the test scores in measuring the underlying trait or the observed score generated from the IRT model. Since the object of a test in IRT is to measure the latent concept for everyone taking the test, an overall measure of the reliability of the test scores across the spectrum of the latent distribution is important to consider.

When using IRT in empirical research, a reliability coefficient is often reported. The reported reliability coefficients are estimates of the true reliability coefficient and these estimates have a certain amount of variance associated with them due to the item parameter estimation. The large sample variance for several reliability coefficient estimators in classical test theory have been derived (Yuan & Bentler, 2002). However, analytical estimates of the variance of the IRT reliability coefficients are not available in the literature and an estimated standard error or confidence interval is therefore usually not reported. In this article, the large sample variance of the estimators of the IRT marginal reliability coefficient and the IRT test reliability coefficient are derived using standard asymptotic theory. The results may be used by empirical researchers to estimate confidence intervals for the reliability coefficients.

The article is structured as follows. First, IRT is introduced and common models for either dichotomous or polytomous data are briefly described. Then, IRT marginal and test reliability coefficients are defined and the asymptotic variance of estimators of these are derived. It is then investigated how well the large sample results work with finite samples through a simulation study. Finally, concluding remarks are given.

Item Response Theory

Consider a test consisting of J number of dichotomous or polytomous items. Denote the IRT item parameter vector by α and let $\alpha_j \in \alpha$ be the item parameter vector for item j . Let $P_{jk}(\theta; \alpha_j)$ be the probability of obtaining category $k \in \{1, \dots, m_j\}$ on item j conditional on the latent variable θ , where m_j denotes the number of possible categories for item j . For dichotomous items, the three-parameter logistic (3-PL) model (Lord, 1980) defines the probability as

$$P_{jk}(\theta; \alpha_j) = u_{jk}P_j(\theta; \alpha_j) + (1 - u_{jk})(1 - P_j(\theta; \alpha_j)), \quad (1)$$

where $u_{j1} = 0$ and $u_{j2} = 1$ and

$$P_j(\theta; \alpha_j) = c_j + \frac{1 - c_j}{1 + \exp[-a_j(\theta - b_j)]}, \quad (2)$$

where a_j is called the discrimination parameter, b_j the location parameter and c_j the pseudo-guessing parameter for item j . For polytomous items, the generalized partial credit model (GPCM) (Muraki, 1992) can be used, with probabilities defined by

$$P_{jk}(\theta; \alpha_j) = \frac{\exp\left[\sum_{v=1}^k a_j(\theta - b_{j,v})\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^c a_j(\theta - b_{j,v})\right]}, \quad (3)$$

where a_j is again the discrimination parameter and each $b_{j,v}$ are called the item category parameters, with $b_{j,1} \equiv 0$. In IRT, the concept of item and test information has an important role. The test information is inversely related to the variance to the maximum likelihood estimator of the ability. Define the expected item information for a dichotomous IRT model as

$$I_j(\theta; \alpha_j) = \sum_{k=1}^{m_j} \frac{\left(\frac{\partial P_{jk}(\theta; \alpha_j)}{\partial \theta}\right)^2}{P_{jk}(\theta; \alpha_j)}, \quad (4)$$

where, for the 3-PL model,

$$\frac{\partial P_{jk}(\theta; \alpha_j)}{\partial \theta} = u_{jk} \frac{\partial P_j(\theta; \alpha_j)}{\partial \theta} - (1 - u_{jk}) \frac{\partial P_j(\theta; \alpha_j)}{\partial \theta} \quad (5)$$

where (Hambleton & Swaminathan, 1985)

$$\frac{\partial P_j(\theta; \alpha_j)}{\partial \theta} = \frac{a_j[1 - P_j(\theta; \alpha_j)][P_j(\theta; \alpha_j) - c_j]}{1 - c_j}. \quad (6)$$

For a polytomous IRT model, define the expected item information as (Magis, 2015)

$$I_j(\theta; \alpha_j) = \sum_{k=1}^{m_j} \left[\frac{\left(\frac{\partial P_{jk}(\theta; \alpha_j)}{\partial \theta}\right)^2}{P_{jk}(\theta; \alpha_j)} - \frac{\partial^2 P_{jk}(\theta; \alpha_j)}{\partial \theta^2} \right], \quad (7)$$

where, for the GPCM (Muraki, 1992),

$$\frac{\partial P_{jk}(\theta; \alpha_j)}{\partial \theta} = P_{jk}(\theta; \alpha_j) a_j \left(k - \sum_{c=1}^{m_j} c P_{jc}(\theta; \alpha_j) \right) \quad (8)$$

and

$$\frac{\partial^2 P_{jk}(\theta; \alpha_j)}{\partial \theta^2} = \frac{\partial P_{jk}(\theta; \alpha_j)}{\partial \theta} a_j \left(k - \sum_{c=1}^{m_j} c P_{jc}(\theta; \alpha_j) \right) - P_{jk}(\theta; \alpha_j) a_j \sum_{c=1}^{m_j} c \frac{\partial P_{jc}(\theta; \alpha_j)}{\partial \theta}, \quad (9)$$

resulting in the expected information for the GPCM having the expression

$$I_j(\theta; \alpha_j) = \sum_{k=1}^{m_j} P_{jk}(\theta; \alpha_j) a_j \sum_{c=1}^{m_j} c \frac{\partial P_{jc}(\theta; \alpha_j)}{\partial \theta}. \quad (10)$$

The test information is the sum of the information for each item, i.e.,

$$I(\theta; \alpha) = \sum_{j=1}^J I_j(\theta; \alpha_j). \quad (11)$$

Let $\hat{\theta}_{MLE}$ be the maximum likelihood estimator of the ability. As $J \rightarrow \infty$, the estimator $\hat{\theta}_{MLE}$ has variance (Hambleton & Swaminathan, 1985)

$$\text{Var}(\hat{\theta}_{MLE} | \theta) = \frac{1}{I(\theta; \alpha)}. \quad (12)$$

Item Response Theory Marginal Reliability

One measure of reliability for scores on the latent variable metric which has been proposed in the literature is the marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984), sometimes referred to as parallel forms reliability (Kim, 2012). With the assumption of an ability distribution with density g and variance 1, this reliability coefficient is defined as (Cheng et al., 2012)

$$\rho_{\Theta}(\alpha) = \int_{-\infty}^{\infty} \frac{I(\theta; \alpha)}{I(\theta; \alpha) + 1} g(\theta) d\theta. \quad (13)$$

The marginal reliability coefficient can be interpreted as the reliability with regard to the maximum likelihood ability estimates. The density function g is often assumed to be the standard normal distribution during estimation and this is the assumption that will be used hereafter.

Item Response Theory Test Reliability

The IRT test reliability coefficient $\rho_{XX'}(\alpha)$ is defined as the ratio of the true score variance $\sigma_T^2(\alpha)$ and the observed score variance $\sigma_X^2(\alpha)$. It can be interpreted as the parallel forms reliability with regard to the sum score. Define W_{jk} as the score for obtaining category k on item j , which we consider as a fixed constant in this article. From the definition of reliability, we can note that $\rho_{XX'}(\alpha)$ can also be expressed as

$$\rho_{XX'}(\boldsymbol{\alpha}) = \frac{\sigma_T^2(\boldsymbol{\alpha})}{\sigma_X^2(\boldsymbol{\alpha})} = \frac{\sigma_X^2(\boldsymbol{\alpha}) - \sigma_e^2(\boldsymbol{\alpha})}{\sigma_X^2(\boldsymbol{\alpha})} = 1 - \frac{\sigma_e^2(\boldsymbol{\alpha})}{\sigma_X^2(\boldsymbol{\alpha})}, \quad (14)$$

where, following Kim and Feldt (2010), $\sigma_e^2(\boldsymbol{\alpha})$ is the conditional error variance $\sigma_{e|\theta}^2(\boldsymbol{\alpha})$ averaged over the distribution of θ , i.e.,

$$\sigma_e^2(\boldsymbol{\alpha}) = \int_{-\infty}^{\infty} \sigma_{e|\theta}^2(\boldsymbol{\alpha}) g(\theta) d\theta, \quad (15)$$

where the conditional error variance $\sigma_{e|\theta}^2(\boldsymbol{\alpha})$ can be calculated by (Kim & Feldt, 2010)

$$\sigma_{e|\theta}^2(\boldsymbol{\alpha}) = \sum_{j=1}^J \sigma_{e_j|\theta}^2(\boldsymbol{\alpha}_j) = \sum_{j=1}^J \sigma_{X_j|\theta}^2(\boldsymbol{\alpha}_j) = \sum_{j=1}^J \left\{ \sum_{k=1}^{m_j} P_{jk}(\theta; \boldsymbol{\alpha}_j) W_{jk}^2 - \left[\sum_{k=1}^{m_j} P_{jk}(\theta; \boldsymbol{\alpha}_j) W_{jk} \right]^2 \right\}. \quad (16)$$

Consider a test that has possible observed scores from 0 to K . The variance of the observed score distribution, $\sigma_X^2(\boldsymbol{\alpha})$, can be calculated from the probabilities $r_i(\boldsymbol{\alpha})$ for each possible observed score, i.e.,

$$\sigma_X^2(\boldsymbol{\alpha}) = \sum_{i=0}^K r_i(\boldsymbol{\alpha}) x_i^2 - \left(\sum_{i=0}^K r_i(\boldsymbol{\alpha}) x_i \right)^2, \quad (17)$$

where x_i denotes the score value of the i th observed score. The probabilities $r_i(\boldsymbol{\alpha})$ can be calculated from the recursive formulas given for dichotomous items in Lord and Wingersky (1984) and for polytomous items in Thissen, Pommerich, Billeaud, and Williams (1995).

Asymptotic Variance of Item Response Theory Reliability Coefficient Estimators

Define $\hat{\boldsymbol{\alpha}}$ as the estimator of the item parameter vector and let n denote the sample size. In the following, it is assumed that asymptotic normality holds for the item parameter estimator $\hat{\boldsymbol{\alpha}}$. Under this condition, the asymptotic variance of the reliability coefficient estimators can be derived using standard asymptotic theory (Ferguson, 1996). Let $\Sigma_{\hat{\boldsymbol{\alpha}}}$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\alpha}}$.

Item Response Theory Marginal Reliability

The integral in Equation (13) is approximated by a sum using Gauss-Hermite quadrature and we thus obtain the estimator of $\rho_{\Theta}(\boldsymbol{\alpha})$ as

$$\hat{\rho}_{\Theta}(\boldsymbol{\alpha}) = \sum_{l=1}^L \frac{\sum_{j=1}^J I_j(\theta_l; \boldsymbol{\alpha})}{\sum_{j=1}^J I_j(\theta_l; \boldsymbol{\alpha}) + 1} \frac{e^{\frac{\theta_l^2}{2}}}{\sqrt{2\pi}} w_l, \quad (18)$$

where w_l is the weight for the l th Gauss-Hermite quadrature point θ_l . The function $\hat{\rho}_{\Theta}(\boldsymbol{\alpha})$ is continuous and differentiable with regard to the item parameter vector $\boldsymbol{\alpha}$. We may then apply the delta method to approximate the variance of $\hat{\rho}_{\Theta}(\boldsymbol{\alpha})$ and hence we retrieve, for large n ,

$$\sigma_{\hat{\rho}_{\Theta}(\boldsymbol{\alpha})}^2 \approx \frac{\partial \hat{\rho}_{\Theta}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \Sigma_{\hat{\boldsymbol{a}}} \left[\frac{\partial \hat{\rho}_{\Theta}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \right]', \quad (19)$$

where

$$\begin{aligned} \frac{\partial \hat{\rho}_{\Theta}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= \frac{\partial \left[\sum_{l=1}^L \frac{\sum_{j=1}^J I_j(\theta_l; \boldsymbol{\alpha}_j)}{\sum_{j=1}^J I_j(\theta_l; \boldsymbol{\alpha}_j) + 1} \frac{e^{\frac{\theta_l^2}{2}}}{\sqrt{2\pi}} w_l \right]}{\partial \boldsymbol{\alpha}} \\ &= \sum_{l=1}^L \left\{ \frac{\sum_{j=1}^J \frac{\partial I_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}}}{\sum_{j=1}^J I_j(\theta_l; \boldsymbol{\alpha}_j) + 1} - \frac{\sum_{j=1}^J I_j(\theta_l; \boldsymbol{\alpha}_j) \sum_{j=1}^J \frac{\partial I_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}}}{\left[\sum_{j=1}^J I_j(\theta_l; 2\boldsymbol{\alpha}_j) + 1 \right]^2} \right\} \frac{e^{\frac{\theta_l^2}{2}}}{\sqrt{2\pi}} w_l. \end{aligned} \quad (20)$$

In Equation (20), the expression of $\frac{\partial I_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}}$ for dichotomous IRT models is

$$\frac{\partial I_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}} = \sum_{k=1}^{m_j} \frac{\partial P_{jk}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \theta_l} \left[2 \frac{\partial^2 P_{jk}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \theta_l \partial \boldsymbol{\alpha}} - \frac{\partial P_{jk}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \theta_l} \frac{\partial P_{jk}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}} \right] \quad (21)$$

and for polytomous IRT with the GPCM is

$$\begin{aligned} \frac{\partial I_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}} &= \sum_{k=1}^{m_j} \left\{ \frac{\partial P_{jk}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}} a_j \sum_{c=1}^{m_j} c P_{jc}(\theta_l; \boldsymbol{\alpha}_j) + \frac{\partial a_j}{\partial \boldsymbol{\alpha}} P_{jk}(\theta_l; \boldsymbol{\alpha}) \sum_{c=1}^{m_j} c P_{jc}(\theta_l; \boldsymbol{\alpha}_j) \right. \\ &\quad \left. + P_{jk}(\theta_l; \boldsymbol{\alpha}_j) a_j \sum_{c=1}^{m_j} c \frac{\partial^2 P_{jc}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \theta_l \partial \boldsymbol{\alpha}} \right\}. \end{aligned} \quad (22)$$

For the 3-PL model,

$$\frac{\partial P_{jk}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}} = u_{jk} \frac{\partial P_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}} - (1 - u_{jk}) \frac{\partial P_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}} \quad (23)$$

where $\frac{\partial P_j(\theta_l; \boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}}$ is given in Hambleton and Swaminathan (1985) while for the GPCM the vector $\frac{\partial P_{jk}(\theta_l; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$ is given in Muraki (1992). The vector $\frac{\partial^2 P_{jk}(\theta_l; \boldsymbol{\alpha}_j)}{\partial \theta_l \partial \boldsymbol{\alpha}}$ is provided in the appendix.

Item Response Theory Test Reliability

The integral implicit in the numerator of Equation (14) is approximated by a sum using Gauss-Hermite quadrature and so are the integrals required to calculate $r_i(\alpha)$. We thus obtain the estimator of $\rho_{XX'}(\alpha)$ as

$$\hat{\rho}_{XX'}(\alpha) = 1 - \frac{\sum_{l=1}^L \sum_{j=1}^J \left\{ \sum_{k=1}^{m_j} P_{jk}(\theta_l; \alpha) W_{jk}^2 - \left[\sum_{k=1}^{m_j} P_{jk}(\theta_l; \alpha) W_{jk} \right]^2 \right\} \frac{e^{\frac{\theta_l^2}{2}}}{\sqrt{2\pi}} w_l}{\sum_{i=0}^K \hat{r}_i(\alpha) x_i^2 - \left[\sum_{i=0}^K \hat{r}_i(\alpha) x_i \right]^2}. \quad (24)$$

Again, $\hat{\rho}_{XX'}(\alpha)$ is a continuous and differentiable function of the item parameter vector α . Hence, the delta method can be used to approximate the variance of $\hat{\rho}_{XX'}(\alpha)$ and we retrieve, for large n ,

$$\sigma_{\hat{\rho}_{XX'}(\alpha)}^2 \approx \frac{\partial \hat{\rho}_{XX'}(\alpha)}{\partial \alpha} \Sigma_{\hat{\alpha}} \left[\frac{\partial \hat{\rho}_{XX'}(\alpha)}{\partial \alpha} \right], \quad (25)$$

where

$$\frac{\partial \hat{\rho}_{XX'}(\alpha)}{\partial \alpha} = \frac{1}{\hat{\sigma}_X^2(\alpha)} \left[\frac{\hat{\sigma}_e^2(\alpha)}{\hat{\sigma}_X^2(\alpha)} \frac{\partial \hat{\sigma}_X^2(\alpha)}{\partial \alpha} - \frac{\partial \hat{\sigma}_e^2(\alpha)}{\partial \alpha} \right]. \quad (26)$$

Note that

$$\begin{aligned} \frac{\partial \hat{\sigma}_e^2(\alpha)}{\partial \alpha} = & \sum_{l=1}^L \left\{ \sum_{j=1}^J \sum_{k=1}^{m_j} \frac{\partial P_{jk}(\theta_l; \alpha)}{\partial \alpha} W_{jk}^2 - 2 \sum_{j=1}^J \left[\sum_{k=1}^{m_j} P_{jk}(\theta_l; \alpha) W_{jk} \sum_{k=1}^{m_j} \frac{\partial P_{jk}(\theta_l; \alpha)}{\partial \alpha} W_{jk} \right] \right\} \frac{e^{\frac{\theta_l^2}{2}}}{\sqrt{2\pi}} w_l \end{aligned} \quad (27)$$

and

$$\frac{\partial \hat{\sigma}_X^2(\alpha)}{\partial \alpha} = \sum_{i=0}^K x_i^2 \frac{\partial \hat{r}_i(\alpha)}{\partial \alpha} - 2 \sum_{i=0}^K x_i \hat{r}_i(\alpha) \sum_{i=0}^K \frac{\partial \hat{r}_i(\alpha)}{\partial \alpha} x_i, \quad (28)$$

where the derivatives $\frac{\partial \hat{r}_i(\alpha)}{\partial \alpha}$ for the GPCM are given in Andersson (2016) and the derivatives $\frac{\partial \hat{r}_i(\alpha)}{\partial \alpha}$ for the 3-PL model are given in Ogasawara (2003).

Confidence Interval Estimation

For large sample sizes, approximate confidence intervals for the reliability coefficients can be constructed using the derived variances and a normal approximation.

Hence, approximate 95% confidence intervals for $\rho_{\Theta}(\alpha)$ or $\rho_{XX'}(\alpha)$ are given by $\hat{\rho}_{\Theta}(\alpha) \pm z_{(0.975)} \sqrt{\hat{\sigma}_{\rho_{\Theta}}^2(\alpha)}$ and $\hat{\rho}_{XX'}(\alpha) \pm z_{(0.975)} \sqrt{\hat{\sigma}_{\rho_{XX'}}^2(\alpha)}$, where $z_{(0.975)}$ denotes the 0.975 quantile of the standard normal distribution. The properties of the confidence intervals in finite samples are investigated in the next section using simulations.

Simulation Study

Simulations were conducted to study the finite sample properties of the derived asymptotic variances and confidence intervals. The values of the latent variable were drawn from the $N(0, 1)$ distribution and the responses to a test with nine GPCM items and the responses to a test with 40 3-PL items were simulated. The GPCM item parameters were taken from the study by Fischer, Tritt, Klapp, and Fliege (2011) and are given in Table 1. For the 3-PL model, the a -parameters were drawn from the $U(0.5, 2)$ distribution, the b -parameters from the $N(0, 1)$ distribution and the c -parameters from the $U(0.15, 0.20)$ distribution. The 3-PL parameters used in the study are given in Table 2. With the selected item parameters and $N(0, 1)$ as the reference distribution, for the GPCM the true marginal reliability coefficient was 0.8718 and the true test reliability coefficient was 0.8744 while for the 3-PL model the true marginal reliability coefficient was 0.8487 and the true test reliability coefficient was 0.8285. The item parameters were estimated with the R (R Development Core Team, 2016) package *mirt* (Chalmers, 2012) using marginal maximum likelihood with the assumption of a $N(0, 1)$ latent distribution. Since maximum likelihood estimation was used, no priors were specified for the item parameters. The observed information matrix (Yuan, Cheng, & Patton, 2013) was calculated after estimation by numerically differentiating the gradient of the observed log-likelihood function at the obtained maximum likelihood estimates to provide an estimate of the asymptotic covariance matrix, using newly written R code to improve the speed of calculation. The observed information matrix is appropriate in this setting since the models are correctly specified. After estimation, the reliability coefficients and their asymptotic variances were calculated using newly written R code. The number of Gauss-Hermite quadrature points used was 49 and 2,000 replications were used. The average bias, average asymptotic standard errors (ASE), Monte Carlo standard errors (MCSE), and empirical coverage rates of 95% confidence intervals were calculated. For the GPCM, sample sizes 250, 500, 1,000, 2,000, 4,000, and 8,000 were considered and for the 3-PL model sample sizes 1,000, 2,000, 4,000, and 8,000 were considered. The reason for only considering the larger sample sizes with the 3-PL model was that reliable estimation with the smaller sample sizes was not possible due to high non-convergence rates. With the GPCM, the nonconvergence rate was 0.45% (sample size 250) and with the 3-PL model the nonconvergence rates were 9.1% (sample size 1,000) and 0.7% (sample size 2,000). For all other settings, the nonconvergence rates were 0%.

Table 1. Generalized Partial Credit Model Item Parameters for the Nine-Item Scale Used in the Simulation Study.

Item	a	b_2	b_3	b_4
1	1.49	-1.14	0.64	0.68
2	2.73	-0.74	0.57	0.95
3	0.79	-1.52	-0.24	0.41
4	1.65	-1.62	0.20	0.40
5	0.71	-0.47	0.36	1.22
6	1.18	0.14	0.80	1.43
7	1.22	-0.70	0.77	1.01
8	0.80	0.37	1.26	1.73
9	1.40	0.89	1.92	2.26

Results

In Table 3 the results for the simulation with the GPCM are given. For the marginal reliability coefficient, there exists a small but statistically significant bias for sample sizes lower than 2,000. The asymptotic standard errors are accurate as an estimate of the sampling variability for all sample sizes. The empirical coverage rates of the 95% confidence intervals are slightly lower than the nominal level for sample sizes 250 and 500 but with sample sizes 1,000 and higher the empirical coverage rate is not statistically significantly different from 95%. For the test reliability coefficient, the bias is not statistically significantly different from zero for any sample size. The asymptotic standard error is accurate for all sample sizes considered. The empirical coverage rates of the 95% confidence intervals for the test reliability coefficient are not statistically significantly different to the nominal level for any sample size.

The results for the 3-PL model simulation are shown in Table 4. The estimator of the marginal reliability coefficient has a small but statistically significant bias under all settings considered. The asymptotic standard errors are accurate for all sample sizes but the empirical coverage rates of the confidence intervals are statistically significantly different from the nominal level of 95% with all sample sizes except the highest sample size of 8,000. For the test reliability coefficient, the bias is smaller and not statistically significantly different from zero with sample sizes 2,000 and higher. The asymptotic standard errors are accurate for all sample sizes and the empirical coverage rate is not statistically significantly different from the nominal level for any sample size.

Concluding Remarks

With the results presented in this article, empirical researchers and practitioners have access to methods that evaluate the variability of reliability coefficient estimators in IRT and with which confidence intervals for the reliability coefficients can be

Table 2. Three-Parameter Logistic Item Parameters Used in the Simulation Study.

Item	<i>a</i>	<i>b</i>	<i>c</i>	Item	<i>a</i>	<i>b</i>	<i>c</i>
1	1.62	−1.85	0.19	21	1.19	0.52	0.20
2	0.81	−0.06	0.15	22	1.25	−0.85	0.18
3	1.68	2.05	0.19	23	0.80	−0.62	0.17
4	1.50	−0.52	0.15	24	1.60	−1.17	0.16
5	0.73	−0.06	0.17	25	1.79	−0.67	0.18
6	1.37	−1.16	0.20	26	0.84	−0.40	0.20
7	1.00	0.79	0.18	27	1.49	−1.94	0.19
8	1.63	−0.40	0.17	28	0.80	0.77	0.17
9	1.55	−1.37	0.18	29	1.66	−0.86	0.16
10	0.70	−1.13	0.19	30	1.13	0.07	0.19
11	1.18	1.28	0.15	31	1.69	1.20	0.18
12	1.76	−2.36	0.19	32	0.62	0.30	0.20
13	0.75	−0.47	0.20	33	0.57	1.13	0.16
14	1.38	0.74	0.17	34	1.07	−0.38	0.17
15	0.54	−1.25	0.16	35	1.55	−0.74	0.17
16	0.92	−0.36	0.19	36	0.88	1.56	0.19
17	1.73	0.50	0.17	37	1.37	−2.30	0.18
18	0.63	−0.90	0.17	38	1.65	0.98	0.16
19	0.61	−1.34	0.20	39	0.95	−0.11	0.17
20	0.75	0.27	0.15	40	1.76	1.40	0.18

Table 3. Mean Bias ($\times 100$), Asymptotic and Monte Carlo Standard Errors ($\times 100$), and Coverage of 95% Confidence Intervals (%) for the GPCM Reliability Coefficient Estimators, With Estimated Standard Errors in Parentheses.

Reliability	Sample size	Mean bias	ASE	MCSE	CI coverage
Marginal	250	0.14 (0.02)	0.90 (0.00)	0.90 (0.02)	92.71 (0.58)
	500	0.06 (0.01)	0.65 (0.00)	0.66 (0.01)	93.45 (0.55)
	1,000	0.03 (0.01)	0.46 (0.00)	0.46 (0.01)	94.10 (0.53)
	2,000	0.02 (0.01)	0.32 (0.00)	0.33 (0.01)	94.20 (0.52)
	4,000	0.01 (0.01)	0.23 (0.00)	0.23 (0.00)	95.70 (0.45)
	8,000	0.00 (0.00)	0.16 (0.00)	0.16 (0.00)	95.05 (0.49)
Test	250	0.02 (0.03)	1.13 (0.00)	1.13 (0.02)	94.73 (0.50)
	500	−0.01 (0.02)	0.80 (0.00)	0.81 (0.01)	94.15 (0.52)
	1,000	−0.00 (0.01)	0.57 (0.00)	0.57 (0.01)	94.55 (0.51)
	2,000	−0.00 (0.01)	0.40 (0.00)	0.40 (0.01)	95.10 (0.48)
	4,000	0.00 (0.01)	0.28 (0.00)	0.28 (0.00)	95.80 (0.45)
	8,000	−0.00 (0.00)	0.20 (0.00)	0.20 (0.00)	95.30 (0.47)

Note. GPCM = generalized partial credit model; ASE = asymptotic standard error; MCSE = Monte Carlo standard error; CI = confidence interval.

estimated. The simulation study indicates that the estimated confidence intervals for the test reliability coefficient have good coverage properties with sample sizes as small as 250 with the GPCM and 1,000 with the 3-PL model. The estimator of the

Table 4. Mean Bias ($\times 100$), Asymptotic and Monte Carlo Standard Errors ($\times 100$), and Coverage of 95% Confidence Intervals (%) for the 3-PL Reliability Coefficient Estimators, With Estimated Standard Errors in Parentheses.

Reliability	Sample size	Mean bias	ASE	MCSE	CI coverage
Marginal	1,000	0.52 (0.01)	0.60 (0.00)	0.61 (0.01)	84.21 (0.86)
	2,000	0.28 (0.01)	0.43 (0.00)	0.44 (0.01)	89.48 (0.69)
	4,000	0.15 (0.01)	0.31 (0.00)	0.32 (0.01)	91.95 (0.61)
	8,000	0.08 (0.00)	0.22 (0.00)	0.22 (0.00)	94.70 (0.50)
Test	1,000	0.05 (0.02)	0.72 (0.00)	0.74 (0.01)	94.39 (0.54)
	2,000	0.02 (0.01)	0.51 (0.00)	0.51 (0.01)	95.12 (0.48)
	4,000	0.01 (0.01)	0.36 (0.00)	0.36 (0.01)	95.25 (0.48)
	8,000	-0.00 (0.01)	0.25 (0.00)	0.26 (0.00)	95.65 (0.46)

Note. 3-PL = three-parameter logistic; ASE = asymptotic standard error; MCSE = Monte Carlo standard error; CI = confidence interval.

marginal reliability coefficient is slightly biased with small samples when using the GPCM and with all sample sizes considered in this article when using the 3-PL model. With the GPCM, the estimated confidence intervals were however still largely accurate and had correct coverage with sample size 1,000 while with the 3-PL model the confidence intervals had correct coverage only with the largest sample size considered. If sufficient computational resources are available, a nonparametric bootstrap approach (Davison & Hinkley, 1997) can be used to estimate the finite sample bias of the reliability coefficient estimators. This bias estimate can then be used together with the results in this article to generate bias-adjusted confidence intervals, which will achieve an improved coverage rate.

The differences between the properties of the marginal and test reliability estimators can be attributed to the differences with regard to the estimation of item response functions and the estimation of expected information functions. In Ogasawara (2002), it was shown that item response functions were possible to estimate accurately in spite of unstable item parameter estimates while the expected information functions were more affected by unstable item parameter estimates. Since the marginal reliability is calculated from the expected information functions while the test reliability only uses the item response functions, the difference in the accuracy of the two different reliability estimators is consistent with the results of Ogasawara (2002).

Since the marginal and test reliability coefficients measure different concepts, a direct comparison between them is not very meaningful. However, this study does indicate that the estimator of the test reliability coefficient has slightly higher sampling variability but lower bias than the estimator of the marginal reliability coefficient when using the same item parameters.

When calculating and reporting the IRT reliability coefficient estimates, it is important to note that the reliability coefficients are not invariant with regard to the latent distribution. This means that the reliability coefficients will be different for

populations with different latent distributions even if the item parameters are invariant. If the distribution parameters of the population are known, the results of this paper can incorporate these without any changes to the derivations presented by suitably changing the approximation of the integrals needed. If the distribution parameters have been estimated, for example by using parametric (Mislevy, 1984), nonparametric (Bock & Aitkin, 1981), semi-parametric (Woods, 2006) or multiple group (Muthén & Lehman, 1985) methods, an extension to the methods in this article is required where the derivatives with respect to the distribution parameters are derived.

Appendix

With the GPCM, for each item j and each category $k \in \{1, \dots, m_j\}$, the required derivative vector is

$$\frac{\partial^2 P_{jk}(\theta_l; \alpha_j)}{\partial \theta_l \partial \alpha} = \begin{pmatrix} 0 & \dots & \frac{\partial^2 P_{jk}(\theta_l; \alpha_j)}{\partial \theta_l \partial a_j} & \frac{\partial^2 P_{jk}(\theta_l; \alpha_j)}{\partial \theta_l \partial b_{j2}} & \dots & \frac{\partial^2 P_{jk}(\theta_l; \alpha_j)}{\partial \theta_l \partial b_{jm_j}} & 0 & \dots & 0 \end{pmatrix}, \quad (29)$$

where

$$\begin{aligned} \frac{\partial^2 P_{jk}(\theta_l; \alpha_j)}{\partial \theta_l \partial a_j} &= P_{jk}(\theta_l; \alpha_j) \left(k - \sum_{c=1}^{m_j} c P_{jc}(\theta_l; \alpha_j) \right) + \frac{\partial P_{jk}(\theta_l; \alpha_j)}{\partial a_j} a_j \left(k - \sum_{c=1}^{m_j} c P_{jc}(\theta_l; \alpha_j) \right) \\ &\quad - P_{jk}(\theta_l; \alpha_j) a_j \sum_{c=1}^{m_j} c \frac{\partial P_{jc}(\theta_l; \alpha_j)}{\partial a_j} \end{aligned} \quad (30)$$

and, for $k^* = 2, \dots, m_j$,

$$\frac{\partial^2 P_{jk}(\theta_l; \alpha_j)}{\partial \theta_l \partial b_{jk^*}} = \frac{\partial P_{jk}(\theta_l; \alpha_j)}{\partial b_{jk^*}} a_j \left(k - \sum_{c=1}^{m_j} c P_{jc}(\theta_l; \alpha_j) \right) - P_{jk}(\theta_l; \alpha_j) a_j \sum_{c=1}^{m_j} c \frac{\partial P_{jc}(\theta_l; \alpha_j)}{\partial b_{jk^*}}. \quad (31)$$

With the 3-PL model, for each item j , the derivative vectors are

$$\frac{\partial^2 P_{j1}(\theta_l; \alpha_j)}{\partial \theta_l \partial \alpha} = \begin{pmatrix} 0 & \dots & -\frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial a_j} & -\frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial b_j} & -\frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial c_j} & 0 & \dots & 0 \end{pmatrix} \quad (32)$$

and

$$\frac{\partial^2 P_{j2}(\theta_l; \alpha_j)}{\partial \theta_l \partial \alpha} = \begin{pmatrix} 0 & \dots & \frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial a_j} & \frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial b_j} & \frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial c_j} & 0 & \dots & 0 \end{pmatrix} \quad (33)$$

where

$$\frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial a_j} = \frac{(1 - P_j(\theta_l; \alpha_j))(P_j(\theta_l; \alpha_j) - c_j)}{1 - c_j} + \frac{a_j(1 - P_j(\theta_l; \alpha_j))}{1 - c_j} \frac{\partial P_j(\theta_l; \alpha_j)}{\partial a_j} - \frac{a_j(P_j(\theta_l; \alpha_j) - c_j)}{1 - c_j} \frac{\partial P_j(\theta_l; \alpha_j)}{\partial a_j}, \quad (34)$$

$$\frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial b_j} = \frac{a_j(1 - P_j(\theta_l; \alpha_j))}{1 - c_j} \frac{\partial P_j(\theta_l; \alpha_j)}{\partial b_j} - \frac{a_j(P_j(\theta_l; \alpha_j) - c_j)}{1 - c_j} \frac{\partial P_j(\theta_l; \alpha_j)}{\partial b_j} \quad (35)$$

and

$$\frac{\partial^2 P_j(\theta_l; \alpha_j)}{\partial \theta_l \partial c_j} = - \frac{a_j(P_j(\theta_l; \alpha_j) - c_j)}{1 - c_j} \frac{P_j(\theta_l; \alpha_j)}{\partial c_j} + a_j(1 - P_j(\theta_l; \alpha_j)) \frac{\left(\frac{\partial P_j(\theta_l; \alpha_j)}{\partial c_j} - 1 \right) (1 - c_j) + (P_j(\theta_l; \alpha_j) - c_j)}{(1 - c_j)^2}. \quad (36)$$

The derivatives $\frac{\partial P_j(\theta_l; \alpha_j)}{\partial a_j}$, $\frac{\partial P_j(\theta_l; \alpha_j)}{\partial b_j}$ and $\frac{\partial P_j(\theta_l; \alpha_j)}{\partial c_j}$ can be found in Hambleton and Swaminathan (1985).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Tao Xin declared funding by the National Natural Science Foundation of China (Grant No. 31371047).

References

- Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT models. *Journal of Educational Measurement*, 53, 459-477.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29.
- Cheng, Y., Yuan, K.-H., & Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educational and Psychological Measurement*, 72, 52-67.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, England: Cambridge University Press.
- Ferguson, T. (1996). *A course in large sample theory*. London, England: Chapman & Hall.
- Fischer, H. F., Tritt, K., Klapp, B. F., & Fliege, H. (2011). How to compare scores from different depression scales: Equating the patient health questionnaire (PHQ) and the

- ICD-10-symptom rating (ISR) using item response theory. *International Journal of Methods in Psychiatric Research*, 20, 203-214.
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347-360.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.
- Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, 77, 153-162.
- Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, 11, 179-188.
- Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, 8, 452-461.
- Magis, D. (2015). A note on the equivalence between observed and expected information functions with polytomous IRT models. *Journal of Educational and Behavioral Statistics*, 40, 96-105.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49, 359-381.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational and Behavioral Statistics*, 10, 133-142.
- Ogasawara, H. (2002). Stable response functions with unstable item parameter estimates. *Applied Psychological Measurement*, 26, 239-254.
- Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika*, 68, 193-211.
- R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, 19, 39-49.
- Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, 11, 253-270.
- Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika*, 67, 251-259.
- Yuan, K.-H., Cheng, Y., & Patton, J. (2013). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, 79, 232-254.