Efficient Standard Errors in Item Response Theory Models for Short Tests

Lianne Ippel¹ and David Magis²

Abstract

In dichotomous item response theory (IRT) framework, the asymptotic standard error (ASE) is the most common statistic to evaluate the precision of various ability estimators. Easy-to-use ASE formulas are readily available; however, the accuracy of some of these formulas was recently questioned and new ASE formulas were derived from a general asymptotic theory framework. Furthermore, exact standard errors were suggested to better evaluate the precision of ability estimators, especially with short tests for which the asymptotic framework is invalid. Unfortunately, the accuracy of exact standard errors was assessed so far only in a very limiting setting. The purpose of this article is to perform a global comparison of exact versus (classical and new formulations of) asymptotic standard errors, for a wide range of usual IRT ability estimators, IRT models, and with short tests. Results indicate that exact standard errors globally outperform the ASE versions in terms of reduced bias and root mean square error, while the new ASE formulas are also globally less biased than their classical counterparts. Further discussion about the usefulness and practical computation of exact standard errors are outlined.

Keywords

item response theory, ability estimation, asymptotic standard error, exact standard error

¹Maastricht University, Maastricht, the Netherlands
²University of Liège, Liège, Belgium

Corresponding Author:
Lianne Ippel, Institute of Data Science, Maastricht University, Maastricht, Postbus 616, Maastricht, 6200 MD, Netherlands.
Email: lianne.ippel@maastrichtuniversity.nl
Introduction

Over the past decades, item response theory (IRT) models have received increased attention (De Boeck & Wilson, 2004; DeMars, 2010; Embretson & Reise, 2000; Li & Lissitz, 2004; van der Linden & Hambleton, 1997). Models for unidimensional or multidimensional latent traits (Reckase, 2009) for dichotomously or polytomously scored items (Ostini & Nering, 2006), model calibration techniques (Baker & Kim, 2004; Breslow & Clayton, 1993), and ability estimation methods were developed. In this article, we focus on the simple framework of unidimensional dichotomous IRT models.

Once an IRT model is calibrated and item parameters are fixed, the estimation of person ability levels can be performed. Several estimators were developed and many studies compared the asymptotic properties of these estimators, such as unbiasedness, uniqueness, asymptotic normality, and consistency (e.g., Lord, 1986; Magis & Verhelst, 2017; Schuster & Yuan, 2011; Warm, 1989). One aspect, however, that has received far less attention is the precision of these estimators, namely, the computation of their standard errors (SEs; Yang, Hansen, & Cai, 2011). Usually, formulas derived from asymptotic considerations are provided to compute the so-called asymptotic standard errors (ASEs). These ASEs are then considered as a measure of precision (of variability) of the estimation techniques. ASEs are readily available from referenced textbooks (Baker & Kim, 2004; Hambleton & Swaminathan, 1985) and in IRT software (e.g., Magis & Barrada, 2017; Zimowski, Muraki, Mislevy, & Bock, 1996). SEs, however, are not only useful for precision measurement; they can also be used as elements of item selection and stopping rule in computerized adaptive testing (CAT) scenarios of test administration (Magis, Yan, & von Davier, 2017; van der Linden & Glas, 2010).

Recent studies have given reasons to question the accuracy of the classical ASE formulas in realistic test scenarios. Since these ASEs are derived from an asymptotic framework, it is assumed that the number of items increases toward infinity. When many items have been administered, the resulting ability-level distribution can be treated as continuous, and ASEs are easily derived. However, it is unclear when the number of items satisfies this assumption (Lord, 1986). Given that tests usually consist of a limited number of items, the resulting ability-level distribution is discrete rather than continuous. Additionally, controversy arose when new formulations of the ASE were derived for two commonly used ability estimators, namely Bayes modal (BM) and weighted likelihood (WL; Magis, 2016).

In an attempt to obtain more efficient SE estimates, Magis (2014a) suggested computing exact standard errors, instead of asymptotic quantities, for the maximum likelihood (ML) and WL estimators. Exact standard errors are computed using the full-sample distribution of the response patterns and associated ability levels. Exact SEs can be obtained from a generic approach and thus for any specific estimator. Exact SEs are especially efficient when there is a small number of items (i.e., when the ability-level distribution deviates from a continuous density) and the generation of the sample distribution is neither technically nor computationally intensive.
The purpose of this article is to perform a comparison of different ways to compute the standard errors: exact, classical, and with the newly suggested formulations. In an intensive simulation study, we evaluate each of the formulations of the SE for dichotomous IRT models with five commonly used IRT ability-level estimators. The design is set such that it encompasses those SE formulations described in former related studies (Magis, 2014a, 2016) but with more general design factors and in a unique global approach.

This article is organized as follows. In the next section, notations are fixed and the IRT ability estimators and related standard errors are described. The general simulation study that compares all kinds of SEs and the results are presented in the subsequent section, together with a brief description of the results in the penultimate section. Eventually, perspectives for linear and adaptive tests are outlined in the final section.

**Ability Estimators and Standard Errors**

Throughout this article, the following notation is used. A test consists of \( n \) dichotomous items, each item \( i \) \((i = 1 \ldots n)\) with a specific set of parameters. The two-parameter logistic model (2PL) is considered in this study. With this model, the probability of a correct response is as follows:

\[
p_i(\theta) = \Pr(X_i = 1|\theta, a_i, b_i) = \frac{\exp [a_i(\theta - b_i)]}{1 + \exp [a_i(\theta - b_i)]},
\]

where \( X_i \) is the response to a single item \( i \) (coded as zero for an incorrect response and one for a correct response); \( \theta \) is the ability level of the test taker; and \( a_i \) and \( b_i \) are the two-item parameters, respectively, the discrimination level and the difficulty level. Under the simpler one-parameter logistic model (1PL), the discriminations \( a_i \) are set to a common constant across all items, usually 1. Item parameters are assumed to be known from previous calibration, so focus is on ability estimation only.

**Ability Estimation**

Five ability estimators are considered: maximum likelihood (ML; Lord, 1980), weighted likelihood (WL; Warm, 1989), robust estimator (ROB; Schuster & Yuan, 2011), Bayes modal (BM; Birnbaum, 1969) and expected a posteriori (EAP; Bock & Mislevy, 1982). They are briefly described below (see Table 1 for more details). Further details can be found in the aforementioned references.

ML estimator \( \hat{\theta}_{ML} \) is obtained by maximizing the likelihood \( L(\theta) \) or the log-likelihood \( l(\theta) \) of the response pattern \( X = (X_1, \ldots, X_n) \) for the test taker of interest. Equivalently, \( \hat{\theta}_{ML} \) is computed by equating the derivative of \( l(\theta) \) (with respect to \( \theta \)) equal to zero:
Table 1. Ability Estimation Methods and Their Corresponding ASEs.

| Method | Estimator | ASE classical | ASE new |
|--------|-----------|---------------|---------|
| ML     | $\hat{\theta}_{ML} = \frac{dl(\theta)}{d\theta} = 0$ | $\frac{1}{\sqrt{|I(\hat{\theta}_{ML})|}}$ | — |
| WL     | $\hat{\theta}_{WL} = \frac{dl(\theta)}{d\theta} + \frac{J(\theta)}{2I(\theta)} = 0$ | $\frac{1}{\sqrt{|I(\hat{\theta}_{WL})|}}$ | $\sqrt{|I(\hat{\theta}_{WL}) + F(\hat{\theta}_{WL})|}$ |
| ROB    | $\hat{\theta}_{ROB} = \sum_{i=1}^{n} \omega \frac{dI(\theta)}{d\theta} \frac{dl(\theta)}{d\theta} = 0$ | $\sqrt{\sum_{i=1}^{n} \omega \frac{dI(\theta)}{d\theta} \frac{dl(\theta)}{d\theta} + \sum_{i=1}^{n} \omega \frac{\frac{d^2 log f(\theta)}{d\theta^2}}{I(\theta)}} \frac{d\theta}{d\theta}$ | — |
| BM     | $\hat{\theta}_{BM} = \frac{dlog f(\theta)}{d\theta} + \frac{dl(\theta)}{d\theta} = 0$ | $\sqrt{\sum_{i=1}^{n} \omega \frac{dI(\theta)}{d\theta} \frac{dl(\theta)}{d\theta} + \sum_{i=1}^{n} \omega \frac{\frac{d^2 log f(\theta)}{d\theta^2}}{I(\theta)}} \frac{d\theta}{d\theta}$ | — |
| EAP    | $\hat{\theta}_{EAP} = \frac{\int \theta f(\theta) I(\theta) \frac{d\theta}{d\theta}}{\int f(\theta) I(\theta) \frac{d\theta}{d\theta}}$ | $\sqrt{\int \frac{\theta - \hat{\theta}_{EAP}}{f(\theta)} \frac{d\theta}{d\theta}}$ | — |

Note. ASE = asymptotic standard error; ML = maximum likelihood; BM = Bayes modal; WL = weighted likelihood; EAP = expected a posteriori; ROB = robust estimator.

\[
\hat{\theta}_{ML} = \frac{dl(\theta)}{d\theta} = 0. 
\] (2)

WL estimator $\hat{\theta}_{WL}$ was developed to correct the bias of the ML estimator. This bias correction is a function of the test information function $I(\theta)$ and an accurate $J(\theta)$ function that depends on the first and second derivatives of the item probability functions (see Warm, 1989, for further details):

\[
\hat{\theta}_{WL} = \frac{dl(\theta)}{d\theta} + \frac{J(\theta)}{2I(\theta)} = 0, 
\] (3)

where

\[
J(\theta) = \sum_{i=1}^{n} \frac{1}{p_i(\theta)(1 - p_i(\theta))} \frac{dp_i(\theta)}{d\theta} \frac{d^2 p_i(\theta)}{d\theta^2}. 
\] (4)

Robust estimator $\hat{\theta}_{ROB}$ is the solution of an equation that involves the derivatives of the log-likelihood components (one per item), each component being weighted by an accurate weight function ($\omega_i$) to reduce the impact of misfitting observations onto the estimation process:

\[
\hat{\theta}_{ROB} = \sum_{i=1}^{n} \omega \frac{dI(\theta)}{d\theta} \frac{dl_i(\theta)}{d\theta} = 0. 
\] (5)

Schuster and Yuan (2011) advocated using the Huber weight function (which was further considered in this study):
\[
\omega_i = \begin{cases} 
1 & \text{if } |r_i| \leq c \\
\frac{c}{|r_i|} & \text{if } |r_i| > c
\end{cases}
\] (6)

where \( r_i = a_i(\theta - b_i) \) is the residual function for item \( i \) and \( c \) is a tuning constant, commonly set to one.

BM estimator performs similarly to ML estimator, however, it maximizes the posterior distribution \( f(\theta)L(\theta) \) instead of the likelihood function, where \( f(\theta) \) is the prespecified prior distribution (usually the standard normal density). Thus, the BM estimate \( \hat{\theta}_{BM} \) can be obtained by equating the derivative of the log-posterior \( \log f(\theta) + l(\theta) \) to zero:

\[
\hat{\theta}_{BM} = \frac{d\log f(\theta)}{d\theta} + \frac{dl(\theta)}{d\theta} = 0.
\] (7)

This solving equation has a similar form to that of the WL estimator, and natural relationships between both estimators were highlighted when the prior density is taken as Jeffreys’s prior (Magis & Raïche, 2012).

Eventually, EAP estimator focuses on the posterior mean rather than the posterior mode (that is, the BM estimate). The EAP value \( \hat{\theta}_{EAP} \) is obtained by computing the average of the posterior distribution \( f(\theta)L(\theta) \), usually by numerical integration over the continuous posterior density:

\[
\hat{\theta}_{EAP} = \frac{\int \theta f(\theta)L(\theta) d\theta}{\int f(\theta)L(\theta) d\theta}.
\] (8)

**Asymptotic Standard Errors**

Once ability estimates are obtained from a given set of items and a response pattern, the computation of related ASEs can be performed. Classical ASE formulas appear in common IRT textbooks, such as Hambleton and Swaminathan (1985) and Lord (1980) for the ML estimator and Wainer (2000) for the BM estimator; or in estimator-specific literature, such as Wainer (2000) for the EAP estimator, Warm (1989) for the WL estimator, and Magis (2014b) for the ROB estimator. These classical ASEs are also listed in Table 1.

This classical way to derive ASEs was recently reconsidered (Magis, 2016) and new ASE formulas were derived from an asymptotic framework common to all (but EAP) estimators. It turned out that new ASEs for ML and ROB estimators are identical to their classical counterparts, while new versions were obtained for the BM and WL estimators. They are listed in the last column of Table 1. Note that for WL method, \( F \) is a function that depends on the test information function \( I(\theta) \), the tuning function \( J(\theta) \), and their derivatives with regard to \( \theta \). The precise formula can be found in Magis (2016).

For BM method, Magis (2016) established that the new ASE always returns smaller values than its classical counterpart (this can also be directly observed from
Table 1). Such a direct relationship does not exist for WL. Yet new ASEs highlighted lower bias and variability than their classical counterparts, especially with short tests (the gap reducing greatly with longer tests). Note that both ASE versions (of BM and of WL estimator) were compared in a simple design involving the 1PL model only. The present study extends the work of Magis (2016) in several ways: 2PL model is included, both classical and newly developed ASEs are compared with exact SE, and more ability estimators are included than in the Magis (2014b) article.

Table 1 provides more details on the computation of the five aforementioned estimators as well as the ASEs.

**Exact Standard Errors**

When the test length is not long enough, for instance at early stages of computerized adaptive testing (CAT), the asymptotic framework is inaccurate and ASE formulas (either the classical or the new version) might become invalid for measuring the precision of ability estimates. This is due to the assumption that the ability distribution is continuous while in practice only a small number of ability levels can occur when the number of items is limited. Exact SEs can theoretically overcome this problematic situation, however, they require a bigger computational effort, since instead of relying on the continuous distribution, a distribution of potential ability levels given the number of items is used. Since the general approach to compute the exact SEs is identical for all ability estimators, in the following developments, the subscript referring to the ability estimator is dropped.

Let $\hat{\theta}$ be the point ability estimate computed using the response pattern $X = (X_1, \ldots, X_n)$. By definition, the exact SE is the standard deviation (SD) of the full sample distribution of all ability estimates for the same set of item parameters. To obtain this sample distribution, one first has to list all potential response patterns $X^{(t)} : t = 1, \ldots, T$ where $T = 2^n$, for instance in case of a 2PL model and three test items, $X^{(t)}$ consists of the following potential responses:

- $t = 1: 0, 0, 0;
- t = 2: 1, 0, 0;
- t = 3: 0, 1, 0;
- t = 4: 0, 0, 1;
- t = 5: 0, 1, 1;
- t = 6: 1, 0, 1;
- t = 7: 1, 1, 0;
- t = 8: 1, 1, 1.$

For each response pattern, we compute the corresponding estimated ability level $\hat{\theta}$ and the likelihood of observing this particular response pattern, $p^{(t)}(\hat{\theta}) = \prod_{i=1}^{n} p_i^{X_i}(1 - p_i)^{(X_i-1)}$; the SD of the resulting distribution of possible $\hat{\theta}$ is the exact SE:

$$SE_{exact}(\hat{\theta}) = \sqrt{\sum_{t=1}^{T} p^{(t)}(\hat{\theta})(\hat{\theta}^{(t)} - \bar{\theta})^2},$$

(9)

where $\bar{\theta} = \sum_{t=1}^{T} p^{(t)}(\hat{\theta}) \hat{\theta}^{(t)}$ is the average ability estimate of the sample distribution.
The most important effort consists in listing the $T = 2^n$ potential response patterns and computing the $2^n$ associated ability estimates. With less than 10 items, this remains feasible. However, exponential increase of computational effort with the test length would discourage the use of exact SEs with longer tests. One main exception is under the 1PL model, since the sample distribution reduces to $n + 1$ values only, one for each possible test score, and applying the Lord–Wingersky algorithm to derive ability probabilities (Magis, 2014a).

Globally, exact SEs return lower bias and lower variability than the ASEs, although tests of 10 items or more do not really exhibit major differences between both approaches. However, the only study on this topic available so far (Magis, 2014a) focused on ML and WL estimators only, and for the latter only the classical ASE was considered. There is need for a broader comparison, involving all estimators and all types of ASE formulas (either classical or new formulations).

**Simulation Study**

**Design**

The aim of this article is to compare the exact SE with their respective asymptotic counterparts in dichotomous IRT models. In the case of WL and BM, this implies both the traditional as well as the more recent formulations of ASE.

Three design factors were manipulated: (1) the underlying model, either 1PL or 2PL; (2) the test length, either 5 items or 10 items; and (3) the true ability level, ranging from $-3$ to 3 by steps of one half unit. We do not consider longer test lengths because the computational effort quickly grows with the test length, while the added value only increases marginally (Magis, 2014a).

For each combination of IRT model and test length, 1,000 sets of item parameters were generated. Under the 2PL model, the discrimination parameter values of these items were drawn from a lognormal model with a mean of zero and a standard deviation of 0.1225 (e.g., Penfield, 2001), while under the 1PL they were all fixed to one. The difficulty parameter values were drawn from a standard normal distribution. Now, for each set of item parameters and each true ability level, 1,000 response patterns were drawn, each item response being generated from a Bernoulli distribution with success probability equal to the probability of answering the item correctly. Hence, per single set of items 13,000 response patterns were drawn (1,000 patterns times 13 true ability levels), yielding 13 millions patterns per IRT model and test length, thus, 52 millions patterns in total.

**Summary Statistics**

Summary statistics were computed separately for each ability estimator (ML, WL, ROB, BM, and EAP), IRT model, test length, and true ability level. Let $\theta$ be the true ability level from which the response pattern was generated. In line with the related work of Magis (2014a, 2016), we use the true ability level as criterion to estimate bias...
of the various SE formulas. Additionally, because \( \hat{\theta} \) might suffer from bias or imprecision, it is impossible to distinguish from the total bias, the contribution related to the ability estimation and that related to the SE.

Consider the \( k \)th replicated set of item parameters \((k = 1, \ldots, K)\) and let \( \hat{\theta}_{mk} \) \((m = 1, \ldots, M)\) be the \( m \)th ability estimate obtained with the \( m \)th generated response pattern (where \( M = 1,000 \)) under the \( k \)th replication (where \( K = 1,000 \)). For a given set of item parameters, and a given combination of ability estimator, IRT model, test length, and true ability three (or four for WL and BM) SE values are computed:

1. The true SE value \( SE_k(\theta) \), set as the exact SE with true underlying ability \( \theta \)
2. The classical ASE value \( ASE_k(\hat{\theta}_{mk}) \), using the formula in Table 1 and with estimated ability \( \hat{\theta}_{mk} \)
3. The exact SE value \( SE_k(\hat{\theta}_{mk}) \) with estimated ability \( \hat{\theta}_{mk} \)
4. The new ASE value \( ASE_{new,k}(\hat{\theta}_{mk}) \) (for BM and WL estimators only), using the formulas from last column of Table 1 and with estimated ability \( \hat{\theta}_{mk} \)

Then, in line with the study of Magis (2014a, 2016) we compute the summary statistics: average signed bias (ASB) for each replication \( k \) (and written below for the exact SE):

\[
ASB_k(SE_k(\hat{\theta}_{mk})) = \frac{1}{M} \sum_{m=1}^{M} (SE_k(\hat{\theta}_{mk}) - SE_k(\theta)),
\]

and where the exact \( SE_k(\hat{\theta}_{mk}) \) can be replaced by the classical ASE or the more recent formulation of the ASE to get the corresponding ASB values. The second summary statistic is the root mean square error (RMSE):

\[
RMSE_k(SE_k(\hat{\theta}_{mk})) = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (SE_k(\hat{\theta}_{mk}) - SE_k(\theta))^2}.
\]

Eventually, ASB and RMSE were averaged out across the \( K = 1,000 \) replications to provide the final summary statistics, per estimator, IRT model, test length, and true ability level:

\[
ASB(SE_k(\hat{\theta}_{mk})) = \frac{1}{K} \sum_{k=1}^{K} ASB_k(SE_k(\hat{\theta}_{mk}))
\]

and

\[
RMSE(SE_k(\hat{\theta}_{mk})) = \frac{1}{K} \sum_{k=1}^{K} RMSE_k(SE_k(\hat{\theta}_{mk})).
\]
Recall that in Equations (10) to (13), the exact SE value $SE_k(\hat{\theta}_{mk})$ can be replaced by the classical ASE or the new ASE to derive corresponding statistics.

The simulation study was conducted in programming environment R (R Core Team, 2018), using some implementation from the functions of the catR package (Magis & Barrada, 2017).

**Results**

First, the results of the 1PL and 2PL models are highly similar for each estimation method (maximum difference in summary statistics between 1PL and 2PL models is less than .01), except for the robust method. Therefore, we chose to present the results for the robust method of both models, and for the remaining methods we only present the 1PL model. Furthermore, the two Bayesian techniques, BM and EAP, resulted in very similar results as well, maximum difference in summary statistics (between BM and EAP) being less than .007. We therefore chose to only present the results of BM.

The general result across test lengths and estimation methods is that especially at the extremes of the ability scale, the ASE is more biased compared with the exact SE. Figures 1, 2, and 3 display the results of ML, WL, and BM methods, respectively, and each figure consists of four panels. The figures are organized as follows: The two panels on the left contain the results of the condition with 5 items and the two panels on the right contain the results of the condition with 10 items. The first row then contains ASB values and the second row RMSE values. In all three figures, the lines with squares represent the ASEs, the lines with circles the exact SEs. Note that the axis ranges differ from one method to another.

The results presented in Figure 1 for ML are in line with the results of Magis (2014a): In the five-item test length condition, the ASE overestimates the SE for those test takers with abilities at the either end of the scale, resulting in high ASB and RMSE. On the other hand, for these extreme test takers the exact SE is only marginally too small. The panel for the 10-item test length condition shows a similar but less pronounced pattern, again comparable to the results presented by Magis (2014a).

For the results presented in Figure 2 for WL, the dotted line with squares represents the new formulation of the ASE, while the solid line with squares represents the results of the traditional formulation of the ASE. Interestingly, the new ASE performs better (in terms of ASB and RMSE) than the classical ASE in both 5- and 10-item tests. Furthermore, ASB of the exact SE is small over the entire range of the ability scale. Exact SE thus outperforms both versions of ASE in this short-test context.

In Figure 3, the results of BM are presented. The new formulation of the ASE performs as well as the exact SE, even in the condition with only five items. Furthermore, the exact SE overestimates the true SE at the interval’s extremes while in Figure 1 the exact SE underestimated the true SE. As mentioned before, similar conclusions can be drawn for the EAP method.

Last, Figure 4 presents the results for the robust method. It deviates from the other figures, because the 1PL and 2PL models yielded slightly different results for this
estimation method. The lines with the open squares and open circles present the 1PL model. The lines with the squares and circles with a “×” present the 2PL model. Both for the 5-item as well the 10-item condition the ASE for the 1PL model yields more bias than the exact SE, either because the ASE is too large (for the extreme abilities), or because it is too small (for the average ability). Note that for the two left panels, ASB and RMSE values for the exact SE show reversed results between the 1PL and 2PL models. This is probably due to the fact that, while the exact SE for 1PL is on average unbiased (or at least less biased than for the 2PL model), the exact SE for 1PL is somewhat more variable than under the 2PL, resulting in larger RMSE for 1PL.

Figure 1. Simulation results of maximum likelihood (ML) for the one-parameter logistic (1PL) model, averaged over 1,000 test takers per ability level, and 1,000 tests.
Discussion

The purpose of this study was to compare the efficiency of exact standard errors in contrast to their commonly used asymptotic counterparts, using standard IRT ability estimators under 1PL and 2PL models. Asymptotic standard errors can be expected to perform poorly for short-test lengths, given that the asymptotic framework conditions are not satisfied. The exact standard errors, however, are especially suitable for short tests, as they are less biased and computationally still feasible. The results of this study are relevant, among others, for CAT scenarios where decisions have to be made in a short time frame with only limited information, or when the routing module of a multistage testing design contains a small number of items (Magis et al., 2017).

Figure 2. Simulation results of weighted likelihood (WL) for the oe-parameter logistic (1PL) model, averaged over 1,000 test takers per ability level, and 1,000 tests.
While we show, in a simulation study, that exact standard errors generally outperform the traditional asymptotic variants, this result is not equally pronounced for each estimation method. For instance, the two new formulations of the ASE presented by Magis (2016) for WL and BM show less biased results compared with the traditional formulations. Moreover, the gain in efficiency of using the exact SE is especially clear at the extreme ends of the ability scale, those levels of ability which were targeted poorly by the test. Even when the test only consists of 10 items, the advantage of exact standard error over asymptotic standard error, which is simpler to compute, is negligible for the midrange test takers.

It would be interesting to study under exactly which conditions it is beneficial to use the exact standard error compared with the asymptotic standard error. Furthermore, so far we only focused on 1PL and 2PL models for dichotomous IRT.
models. It might be worth investigating whether the 3PL model, including guessing, and especially the 4PL model, including inattention, return similar results with regard to the difference between ASEs and exact SEs. Additionally, it is worthwhile to compare ASEs and exact SEs in the context of polytomous IRT models. However, the technical feasibility rapidly decreases for polytomously scored items, since the full distribution will consist of “qn” patterns where q is the number of response categories and n the number of test items.

To summarize, we highlighted that ASEs of five commonly used ability estimators are outperformed in terms of bias by the exact SEs, especially in a scenario with only five items. Furthermore, the newly suggested formulations of the ASE by Magis (2016) are less biased than their classical counterparts. With tests of more than 10

Figure 4. Simulation results of robust estimator (ROB) for the one-parameter logistic (1PL) and two-parameter logistic (2PL) models, averaged over 1,000 test takers per ability level, and 1,000 tests.
items, making use of the new ASE formulations for BM and WL methods is definitely an asset when compared with the classical formulas that are still commonly used nowadays in psychometric research. With short tests or at early stages of CAT, exact SEs would be the optimal preference in terms of reduced bias and RMSE, whatever the estimation method.

**Authors’ Note**

David Magis is currently affiliated to IQVIA Belux.

**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: This research was funded by the Incentive Grant for Scientific Research MIS F.4505.17 of the Fonds de la Recherche Scientifique-FNRS, Belgium.

**ORCID iD**

Lianne Ippel https://orcid.org/0000-0001-8314-0305

**References**

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*, 258-276.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9-25.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

DeMars, C. (2010). *Item response theory*. Oxford, England: Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*, Boston, MA: Kluwer.

Li, Y. H., & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement, 41*, 85-117.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Magis, D. (2014a). Accuracy of asymptotic standard errors of the maximum and weighted likelihood estimators of proficiency levels with short tests. *Applied Psychological Measurement, 38*, 105-121.

Magis, D. (2014b). On the asymptotic standard error of a class of robust estimators of ability in dichotomous item response models. *British Journal of Mathematical and Statistical Psychology, 67*, 430-450.

Magis, D. (2016). Efficient standard error formulas of ability estimators with dichotomous item response models. *Psychometrika, 81*, 184-200.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, 76*(Code Snippet 1). Retrieved from https://www.jstatsoft.org/article/view/v076c01

Magis, D., & Raiche, G. (2012). On the relationships between Jeffreys modal and weighted likelihood estimation of ability under logistic IRT models. *Psychometrika, 77*, 163-169.

Magis, D., & Verhelst, N. (2017). On the finiteness of the weighted likelihood estimator of ability. *Psychometrika, 82*, 637-647.

Magis, D., Yan, D., & von Davier, A. A. (2017). Computerized adaptive and multistage testing (using R packages catR and mstR). New York, NY: Springer.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*, 235-259.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics, 36*, 720-735.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). New York, NY: Routledge/Taylor and Francis.

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

Yang, J. S., Hansen, M., & Cai, L. (2011). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement, 72*, 264-290.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.