British Journal of Mathematical and Statistical Psychology (2014), 67, 430–450
© 2013 The British Psychological Society



The British Psychological Society

www.wileyonlinelibrary.com

On the asymptotic standard error of a class of robust estimators of ability in dichotomous item response models

David Magis[1,2]*

[1]Department of Education, University of Liège, Belgium
[2]Department of Psychology, Katholieke Universiteit Leuven, Belgium

In item response theory, the classical estimators of ability are highly sensitive to response disturbances and can return strongly biased estimates of the true underlying ability level. Robust methods were introduced to lessen the impact of such aberrant responses on the estimation process. The computation of asymptotic (i.e., large-sample) standard errors (ASE) for these robust estimators, however, has not yet been fully considered. This paper focuses on a broad class of robust ability estimators, defined by an appropriate selection of the weight function and the residual measure, for which the ASE is derived from the theory of estimating equations. The maximum likelihood (ML) and the robust estimators, together with their estimated ASEs, are then compared in a simulation study by generating random guessing disturbances. It is concluded that both the estimators and their ASE perform similarly in the absence of random guessing, while the robust estimator and its estimated ASE are less biased and outperform their ML counterparts in the presence of random guessing with large impact on the item response process.

1. Introduction

One currently growing field of research in item response theory (IRT) is the detection of aberrant responses within a pattern or aberrant response patterns as a whole. 'Aberrant' is a generic term that encompasses different response attitudes towards the test. One may think of cheating, guessing, stress, tiredness, lack of time and inattention as processes yielding aberrant item responses. Because such abnormal behaviours jeopardize the idea of a fair measurement among respondents, and because they invalidate the simple use of the underlying IRT model, several indices have been developed to identify response patterns that exhibit some problematic responses. This research field is often referred to as person fit (see, for example, Karabatsos, 2003; Meijer & Sijtsma, 2001).

Person fit analysis is obviously mandatory to detect aberrant trends or behaviours, but it is a *post-hoc* analysis. Indeed, it often requires the prior estimation of item and person parameters, the latter being most probably strongly influenced by the presence of response disturbances. For instance, it is most likely that random guessing in a multiple-choice questionnaire might artificially inflate the total test score and lead to the impression of a higher-ability respondent than in reality. This is even reinforced by the fact that a large proportion of respondents might be tempted to cheat during the test

*Correspondence should be addressed to David Magis, Department of Education (B32), University of Liège, Boulevard du Rectorat 5, B-4000 Liège, Belgium (e-mail: david.magis@ulg.ac.be).

DOI:10.1111/bmsp.12027

(Hutton, 2006). This also has an impact on the statistical distributions of person fit statistics themselves (Molenaar & Hoijtink, 1990; Snijders, 2001).

This paper focuses on a different approach, the robust estimation of ability levels. By 'robust' is meant that the impact of aberrant item responses can be downweighted appropriately, in order to maintain an ability estimation process that is resistant (or robust) to response disturbances. Robust statistical methods are not new and their adaptation to the context of psychometrics has become common in past decades (for a review, see Magis & De Boeck, 2011). The basic goal of robust estimators is not to identify the type of response disturbance present, but rather to provide a robust ability estimate by getting rid of this disturbance, and thus to be less biased than the classical, non-robust estimators. The comparison between classical and robust estimates, however, can be informative. For instance, if the robust estimate is much smaller than its classical alternative (e.g., the maximum likelihood estimator), this may indicate that some item responses are unexpectedly correct and lead to an increase in test score and ability estimate (with non-robust methods). One may therefore think that the response pattern of interest is characterized by random guessing or cheating, and further investigation might be undertaken to identify the type of response disturbance. On the other hand, if both non-robust and robust estimators return similar ability estimates, one may conclude the absence of significant response disturbances.

Wainer and Wright (1980) were among the first to propose a robust approach to estimating ability levels in the Rasch model. Mislevy and Bock (1982) generalized Wainer and Wright's approach by introducing smoother weight functions. Recently, Schuster and Yuan (2011) improved Mislevy and Bock's robust estimator by selecting another weight function for technical estimation issues. Unfortunately, almost all these studies focused on the (robust) point estimation of ability only, and except in Mislevy and Bock (1982) none of them investigated the computation of related standard errors of estimates. This is, however, a key feature for statistical inference as it relates to the precision of the ability estimate, with straightforward consequences and applications in fields such as computerized adaptive testing (Van der Linden & Glas, 2009; Wainer, 2000), where robust estimation of ability after each item administration might improve the adaptive process to some extent. The need for an accurate formula to compute the standard error of robust estimators is therefore real.

The purpose of this paper is threefold: to briefly recall the methodology beyond robust estimation of ability in dichotomous item response models; to derive the asymptotic standard error (ASE) formula for of a broad class of robust estimators; and to assess in a simulation study the formulas derived. Robust estimators, as proposed by Mislevy and Bock (1982), require the selection of both a residual measure that models the gap between the response (or the ability level) and the item characteristics (i.e., parameters), and a weight function that depends on the residual measure. Rather than focusing on a single choice for the weight function and the residual measure, this paper considers a broad class of robust estimators with mild constraints on the residual measure. This makes the ASE formula quite general and applicable for many types of robust estimators, including the proposals of Mislevy and Bock (1982) and Schuster and Yuan (2011).

2. Ability estimation

The following framework is considered. Let θ be the ability level of some respondent assigned a test consisting of n items. The items are dichotomously

scored 0 for an incorrect response and 1 for a correct response. Set X_i as the response to item i ($i = 1,\ldots,n$) and let $P_i(\theta)$ be the probability of answering the item correctly:

$$P_i(\theta) = \Pr(X_i = 1|\theta, \mathbf{p}_i), \tag{1}$$

where $\mathbf{p}_i$ is the set of item parameters that arise from the underlying IRT model. In this paper, the item parameters are assumed to be known values from previous calibrations of the test. This is consistent with the standard framework of person fit (Meijer & Sijtsma, 2001) and robust ability estimation (Schuster & Yuan, 2011). Define $Q_i(\theta) = 1 - P_i(\theta)$ as the probability of an incorrect response to that item. Finally, set $\mathbf{X} = (X_1,\ldots,X_n)$ as the response pattern of the respondent. Two methods of ability estimation, conditionally upon the response pattern and the item parameters, are considered: the maximum likelihood and the robust estimation methods.

2.1. Maximum likelihood estimation

The maximum likelihood (ML) estimator of ability is obtained as the value $\hat{\theta}_{\mathrm{ML}}$ of θ that maximizes the likelihood function

$$L(\theta) = \prod_{i=1}^{n} P_i(\theta)^{X_i} Q_i(\theta)^{1-X_i}, \tag{2}$$

or its logarithm,

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \{X_i \log P_i(\theta) + (1 - X_i) \log Q_i(\theta)\}. \tag{3}$$

Equivalently, $\hat{\theta}_{\mathrm{ML}}$ is obtained by solving the following non-linear relationship:

$$\frac{dl(\theta)}{d\theta} = \sum_{i=1}^{n} \frac{dl_i(\theta)}{d\theta} = 0, \tag{4}$$

with

$$l_i(\theta) = X_i \log P_i(\theta) + (1 - X_i) \log Q_i(\theta). \tag{5}$$

Note that the dependence of $l_i(\theta)$ on X_i is not directly stated for the sake of brievity, but will play a central role in forthcoming developments.

2.2. Robust estimation

As already pointed out, the ML estimator of ability is sensitive to aberrant item responses that may arise from various response disturbances. To circumvent this undesirable phenomenon and its impact on ability estimation, robust methods have been developed. They basically involve weighting the item contributions to equation (4) in order to lessen the $\hat{\theta}_{\mathrm{ROB}}$ impact of aberrant responses to the estimation process. More specifically, the robust estimator of θ satisfies the equation

$$\sum_{i=1}^{n} = \omega[r_i(\theta)] \frac{dl_i(\theta)}{d\theta} = 0, \tag{6}$$

where $r_i(\theta)$ is an appropriate residual measure for item i and $\omega(r)$ is an appropriate weight function. The weight function $\omega(r)$ has the following characteristics: it takes values in [0,1], it equals 1 whenever $r = 0$ and it converges towards 0 when r departs from zero (whether taking positive or negative values). Intuitively, the contribution of the derivative of $l_i(\theta)$ to equation (6) is maximized when the residual measure is 0 and it decreases when this residual becomes large (whether positive or negative). Wainer and Wright (1980) proposed that $\omega[r_i(\theta)]$ equals 1 or 0 depending on whether item i is included in the estimation process or not. Mislevy and Bock (1982) recommended using the so-called Tukey biweight (or bisquare) weight function (Mosteller & Tukey, 1977),

$$\omega(r) = \begin{cases} \left[1 - \left(\frac{r}{B}\right)^2\right]^2 & \text{if } |r| \leq 4 \\ 0 & \text{if } |r| > 4, \end{cases} \tag{7}$$

while Schuster and Yuan (2011) advocated the Huber-type weight function,

$$\omega(r) = \begin{cases} 1 & \text{if } |r| \leq 1 \\ \frac{1}{|r|} & \text{if } |r| > 1. \end{cases} \tag{8}$$

The thresholds 4 and 1 in equations (7) and (8) respectively are actually tuning constants that were fixed by the authors. Note that all these weight functions satisfy the conditions listed above to define $\omega(r)$ appropriately.

In this paper, the residual measures $r_i(\theta)$ are assumed to be independent of the item responses X_i and may depend only on the ability level θ and the item parameters. Such residual measures belong to the so-called Mallows class of residuals (for further details, see Carroll & Pederson, 1993). This is an important assumption for forthcoming developments. Moreover, this is in line with Mislevy and Bock (1982) and Schuster and Yuan (2011), who considered the same residual function under the two-parameter logistic (2PL) model (Birnbaum, 1968):

$$r_i(\theta) = a_i(\theta - b_i), \tag{9}$$

with a_i and b_i being respectively the item discrimination and difficulty levels. Note that some residual measures were developed to take item responses into account directly (Pregibon, 1982). Schuster and Yuan (2011) discuss both types of measures and advocate the use of residual measures of the Mallows class. Therefore, the class of robust estimators that is considered in this paper is defined by (6) with a residual measure that belongs to the Mallows class.

3. Asymptotic standard errors

Estimating the ASE of an ability estimator is probably as important as obtaining the point ability estimate itself. For the ML estimator, the standard error formula is well known and corresponds to the reciprocal of the inverse of the information function (Birnbaum, 1968;

Hambleton & Swaminathan, 1985). For robust estimators, Mislevy and Bock (1982) gave an approximate formula and Schuster and Yuan (2011) did not discuss the issue.

The true ASE of both the ML and the robust estimators, as defined above, can actually be derived from the general framework of estimating equations (Yuan & Jennrich, 1998). Estimating equations take the generic form

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta; X_i), \tag{10}$$

where $g_i(\theta; X_i)$ is a stochastic function of the item response X_i. In the current one-dimensional framework, ability estimators that satisfy equation (10) are broadly defined as solutions $\hat{\theta}$ of

$$G_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\theta}; X_i) = 0. \tag{11}$$

In the present case, and given (4) and (6), the ML and robust ability estimators can be seen as solutions of (11) by setting

$$g_i(\theta; X_i) = \omega_i(\theta)\frac{dl_i(\theta)}{d\theta}, \tag{12}$$

with $\omega_i(\theta) = 1$ for the ML estimator and $\omega_i(\theta) = \omega[r_i(\theta)]$ for the robust estimator.

The strength of the estimating equations approach is that the random variables X_i need to be independent but not necessarily identically distributed. This is central in the IRT context, since item responses are considered to be locally independent (i.e., conditional upon the true ability level) but their distribution is not identical, since X_i actually has a Bernoulli distribution with success probability $P_i(\theta)$ and the latter usually varies across items. Moreover, it is worth noticing that estimating equations often refer to a multivariate framework, so that the covariance matrix of the asymptotic distribution of the estimator is derived. This covariance matrix is known as the sandwich estimate of the true covariance matrix (Huber, 1981; Maronna, Martin, & Yohai, 2006; Rousseeuw & Leroy, 1987; Stefanski & Boos, 2002; Zeileis, 2006). In the present paper, however, one-dimensional IRT models are considered so the derivations will be simpler than under a general matrix notation context.

3.1. Estimating equations and standard errors

Let us first develop the estimating equations framework to derive the ASE of an ability estimator $\hat{\theta}$ of the unknown univariate parameter θ for independent random variables X_i ($i = 1, \ldots, n$) that are distributed as Bernoulli variables with success probabilities $P_i(\theta)$. The asymptotic context refers to a sample size (or a test length in this case) n that converges to infinity. It is assumed, moreover, that: the ability levels are bounded on a continuous scale; the success probabilities $P_i(\theta)$ are bounded away from 0 and 1; and the successive derivatives of $P_i(\theta)$ exist and are bounded away from 0. These assumptions are standard in theoretical IRT developments (see Lord, 1983; Warm, 1989). Moreover, set θ_0 as the true (unknown) parameter value, to be estimated.

The asymptotic distribution of the estimator $\hat{\theta}$ can be easily derived provided that a small set of assumptions are fulfilled. These assumptions arise from Yuan and Jennrich (1998) and are reported below in their univariate form for convenience (see also Inagaki, 1973; Yuan, 1997). $E(\cdot)$ and $V(\cdot)$ refer to the expectation and the variance of the stochastic term in parentheses, while $f'(\theta)$ is standard notation for the first derivative of the function f with respect to θ. Furthermore, $g_i(\theta; X_i)$ is simply written as $g_i(\theta)$ for the sake of brevity.

Assumption 1 *For any* $i = 1, \ldots, n$,

$$E(g_i(\theta_0)) = 0, V(g_i(\theta_0)) = V_i \text{ and } \bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i \rightarrow V, \tag{13}$$

where V *is a positive constant.*

Assumption 2 *There exist positive numbers* δ *and* B *such that, for any* $i = 1, \ldots, n$,

$$E\left(\left|\frac{g_i(\theta_0)^2}{1+V_i}\right|^{1+\delta}\right) \leq B, \tag{14}$$

where $V_i = V(g_i(\theta_0))$ *as in Assumption 1.*

Assumption 3 *For any* $i = 1, \ldots, n$, *the function* $g_i(\theta_0)$ *is twice continuously differentiable in a neighbourhood* Θ of θ_0.

Assumption 4 *For any* $\theta \in \Theta$,

$$G_n'(\theta) \xrightarrow{P} G'(\theta), \tag{15}$$

where $\xrightarrow{P}$ denotes convergence in probability, and $G(\theta)$ *is a non-stochastic function of* θ *such that* $A = G'(\theta_0)$ *is different from zero.*

Assumption 5 *For any* $\theta = \Theta$,

$$E(G_n'(\theta)) \rightarrow G'(\theta). \tag{16}$$

Assumption 6 *For any* $\theta = \Theta$ *and any* $i = 1, \ldots, n$,

$$\|g_i'(\theta)\| \leq T_i \text{ and } \Pr\left(\max_{i \geq 1} T_i < \infty\right) = 1. \tag{17}$$

Yuan and Jennrich (1998) established that, for any estimator $\hat{\theta}$ that is a solution of the estimating equations (11) and such that Assumptions 1–6 are fulfilled,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A^{-2}V), \tag{18}$$

where $\xrightarrow{d}$ denotes convergence in distribution, and A and V are given in Assumptions 4 and 1, respectively. In other words, $\hat{\theta}$ is asymptotically normally distributed with mean θ_0 and variance $A^{-2}V/n$.

3.2. The case of ML and robust estimators

Both the robust estimator (6) and the ML estimator (4), obtained as solutions of the estimating equations (11) with $g_i(\theta)$ functions set as in (12), fulfil the six assumptions listed above. Assumptions 1, 3 and 5 are straightforward to assess and are established below. Proofs that Assumptions 2, 4 and 6 are fulfilled are sketched in the Appendix.

Let us focus on the robust estimator first. Using (5) and (12), the $g_i(\theta)$ functions can be written as

$$g_i(\theta) = \omega_i(\theta)\frac{[X_i - P_i(\theta)]P_i'(\theta)}{P_i(\theta)Q_i(\theta)}, \tag{19}$$

where $P_i'(\theta)$ refers to the first derivative of $P_i(\theta)$ with respect to θ. Then

$$E(g_i(\theta)) = \omega_i(\theta)\frac{P_i'(\theta)}{P_i(\theta)Q_i(\theta)}\ E[X_i - P_i(\theta)] = 0, \tag{20}$$

and

$$V(g_i(\theta)) = \omega_i(\theta)^2\frac{P_i'(\theta)^2}{P_i(\theta)^2Q_i(\theta)^2}\ V(X_i) = \omega_i(\theta)^2\frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{21}$$

due to the Bernoulli distribution of X_i and the fact that the residuals $r_i(\theta)$ belong to the Mallows class. Note that

$$\frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} = I_i(\theta), \tag{22}$$

is the item information function (Birnbaum, 1968; Lord, 1980), and is bounded for any θ because of the aforementioned assumptions on the item response model. It follows that

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V(g_i(\theta_0)) = \frac{1}{n}\sum_{i=1}^{n} \omega_i(\theta_0)^2 I_i(\theta_0). \tag{23}$$

In sum, $\bar{V}$ is the arithmetic mean of n bounded values $\omega_i(\theta_0)^2 I_i(\theta_0)$ and is therefore bounded for any n, so that its limiting value is computed (for large n) as

$$V = \frac{1}{n}\sum_{i=1}^{n} \omega_i(\theta_0)^2 I_i(\theta_0), \tag{24}$$

and is positive by definition of both the weight and the information functions. This ensures that Assumption 1 is fulfilled.

Assumption 3 is trivially fulfilled according to (19) and the aforementioned assumptions on the item response model. Moreover,

$$g_i' = [X_i - P_i(\theta)]\left(\omega_i(\theta)\frac{P_i'(\theta)}{P_i(\theta)Q_i(\theta)}\right)' - \omega_i(\theta)\frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{25}$$

whence

$$E(g_i'(\theta)) = -\omega_i(\theta)I_i(\theta), \tag{26}$$

according to (22) and the Bernoulli distribution of X_i. It follows that

$$E(G_n'(\theta)) = \frac{1}{n}\sum_{i=1}^{n} E(g_i'(\theta)) = -\frac{1}{n}\sum_{i=1}^{n}\omega_i(\theta)I_i(\theta), \tag{27}$$

and hence

$$A = -\frac{1}{n}\sum_{i=1}^{n}\omega_i(\theta_0)I_i(\theta_0), \tag{28}$$

since (27) is the arithmetic mean of n bounded functions $\omega_i(\theta)I_i(\theta)$. Assumption 5 is consequently fulfilled for the robust estimator. Note that these results also hold for the ML estimator since it is a special case of the robust estimator by setting $\omega_i(\theta) = 1$ for all items. Recall that the assessment of the remaining assumptions is displayed in the Appendix.

Consequently, both estimators are asymptotically normally distributed with mean θ_0 (i.e., they are consistent estimators) and their ASE is computed from (18) by using (24) and (28):

$$A^{-2}V = \frac{\frac{1}{n}\sum_{i=1}^{n}\omega_i(\theta_0)^2 I_i(\theta_0)}{\left(\frac{1}{n}\sum_{i=1}^{n}\omega_i(\theta_0)I_i(\theta_0)\right)^2} = \frac{n\sum_{i=1}^{n}\omega_i(\theta_0)^2 I_i(\theta_0)}{\left(\sum_{i=1}^{n}\omega_i(\theta_0)I_i(\theta_0)\right)^2}. \tag{29}$$

Consequently, using (18) and (29), the true ASE of the robust estimator is equal to

$$ASE(\hat{\theta}_{\mathrm{ROB}}) = \frac{\sqrt{\sum_{i=1}^{n}\omega_i(\theta_0)^2 I_i(\theta_0)}}{\sum_{i=1}^{n}\omega_i(\theta_0)I_i(\theta_0)}, \tag{30}$$

and can be approximated by plugging in the robust estimate of ability $\hat{\theta}_{\mathrm{ROB}}$:

$$\widehat{ASE}(\hat{\theta}_{\mathrm{ROB}}) = \frac{\sqrt{\sum_{i=1}^{n}\omega_i(\hat{\theta}_{\mathrm{ROB}})^2 I_i(\hat{\theta}_{\mathrm{ROB}})}}{\sum_{i=1}^{n}\omega_i(\hat{\theta}_{\mathrm{ROB}})I_i(\hat{\theta}_{\mathrm{ROB}})}. \tag{31}$$

For the ML estimator, all weights $\omega_i(\theta)$ equal 1, so the true ASE simplifies to

$$ASE(\hat{\theta}_{\mathrm{ML}}) = \frac{\sqrt{\sum_{i=1}^{n} I_i(\theta_0)}}{\sum_{i=1}^{n} I_i(\theta_0)} = \frac{1}{\sqrt{\sum_{i=1}^{n} I_i(\theta_0)}}, \tag{32}$$

and is approximated by

$$\widehat{ASE}(\hat{\theta}_{\mathrm{ML}}) = \frac{1}{\sqrt{\sum_{i=1}^{n} I_i(\hat{\theta}_{\mathrm{ML}})}}. \tag{33}$$

Equations (32) and (33) are well-known results for the ML estimator (Birnbaum, 1968; Hambleton & Swaminathan, 1985).

4. Simulation study

Given the true ASEs (29) and (31) of the ML and robust ability level estimators, it would be interesting to determine how they can be accurately estimated by plugging in the ability estimates instead of the true levels. Moreover, this estimation of the ASE is worth investigating in the absence or presence of response disturbances. For these reasons, the following simulation study was conducted. It follows closely the one by Schuster and Yuan (2011), but with two main differences. First, the test length was varied in this study, while it was kept to 44 items in the previous simulation. Second, item responses were generated under a 2PL model, while strictly speaking Schuster and Yuan (2011) made use of a 1PL model with regular increases in item difficulty, which is somewhat unrealistic in practical testing situations.

4.1. Design

Three design factors were manipulated: the test length; the type of data generation process (i.e., with or without a response disturbance mechanism); and the true ability levels. Tests with 20, 40 or 60 items were considered. Moreover, only one type of response disturbance mechanism was selected, namely, the random guessing process (Schuster & Yuan, 2011). Although several other processes exist (see St-Onge, Valois, Abdous, & Germain, 2011), we selected this particular mechanism for several reasons. First, it is easy to understand and model, as will be shown later on. Second, Schuster and Yuan (2011) did not highlight major differences between random guessing and transcription errors in the results of their simulation study. Finally, random guessing mostly affects low-ability respondents, which allows the range of ability levels which are considered in the study to be restricted. In the absence of response disturbances, true ability levels θ_0 were selected from -3 to 3 in steps of one unit. In case of presence of random guessing, ability levels from -3 to 0 only were selected, as higher levels would almost never be affected by random guessing.

4.2. Item parameters

One set of 60 pairs of item difficulties and discriminations was generated and used throughout the study, as follows. Let a_i and b_i be the discrimination and difficulty level of item i, respectively. Set $a_i^* = \log a_i$. The a_i and b_i parameters were generated in three steps. First, parent distributions for a_i^* and b_i were fixed: normal distributions with mean 0 and standard deviation 2 (for b_i), and mean 0 and standard deviation 0.1225 (for a_i^*), respectively. The latter value is commonly used to generate item discriminations (Donoghue & Allen, 1993; Penfield, 2003; Zwick, Donoghue, & Grima, 1993). Moreover, both parameters were required to be correlated with a correlation coefficient of .8. Second, 60 pairs of parameters (b_i, a_i^*) were drawn from the bivariate normal distribution BVN(μ, Σ) with

$$\mu = (0, 0)^T \text{ and } \Sigma = \begin{pmatrix} 4 & 0.196 \\ 0.196 & 0.015 \end{pmatrix}. \quad (34)$$

The variances 4 and 0.015 are the squares of the standard deviations 2 and 0.1225, while the covariance 0.196 is derived such that both vectors have correlation .8. Finally, item difficulty levels b_i are set as the first components of the pairs of coefficients, while item discriminations a_i are derived by $a_i = \exp a_i^*$.

The final 60 item parameters are displayed in Table 1. Note that the same set of parameters was used for generating shorter tests of 20 items and 40 items. More precisely, tests of 20 (40) items were generated by using the first 20 (40) item parameters of Table 1.

Summary statistics for the item parameters for each test length are displayed in Table 2. These statistics include, separately for the item difficulty and discrimination parameters, the minimum and maximum values, the 25th and 75th percentiles (P25 and P75), as well as the mean and the median values. In addition, correlations between difficulty and discrimination parameters are given. As expected, item discriminations are centred roughly around 1 with a maximal range from 0.749 to 1.416. Item difficulties, on the other hand, range from −4.703 to 4.271 with a slightly positive average value. Finally, the correlations range from .784 (for tests of 40 items) to .832 (for tests of 20 items), in agreement with the predefined value of .8 at the generation step. Note that a broad range of difficulty levels was chosen to cover the whole range of ability levels, and thus increase the precision of ability estimation at the extremes of this scale.

Table 1. Item discrimination (*a*) and difficulty (*b*) parameters used to generate the samples of 20 items (items 1–20), 40 items (items 1–40) and 60 items (items 1–60)

Item	*a*	*b*	Item	*a*	*b*	Item	*a*	*b*
1	0.799	−1.961	21	1.317	4.271	41	1.049	−0.665
2	1.192	0.561	22	0.782	−2.557	42	0.848	−3.889
3	1.100	3.106	23	1.115	3.125	43	0.901	−1.711
4	0.910	−2.243	24	1.081	3.775	44	1.129	3.732
5	1.032	−0.113	25	0.842	−0.199	45	0.992	1.162
6	0.950	0.187	26	0.915	−4.726	46	1.416	4.171
7	0.915	1.226	27	0.994	0.892	47	1.046	−0.522
8	0.942	−0.524	28	0.924	−1.219	48	1.160	0.008
9	1.214	3.972	29	1.046	1.546	49	0.951	−0.408
10	0.920	−0.362	30	1.059	0.615	50	0.869	−2.428
11	0.968	0.747	31	1.215	1.627	51	0.853	−1.772
12	1.126	1.993	32	1.175	0.796	52	1.192	4.104
13	0.953	−0.798	33	1.009	2.038	53	0.984	−1.079
14	0.938	−2.036	34	0.876	−0.696	54	1.254	2.677
15	1.178	3.555	35	0.826	−1.694	55	0.947	−2.040
16	0.749	−4.703	36	0.857	−1.308	56	0.861	−3.884
17	1.203	1.878	37	0.990	−3.263	57	1.068	−0.529
18	1.066	0.144	38	0.919	−1.802	58	1.195	1.978
19	1.195	2.123	39	0.888	−1.196	59	1.123	2.286
20	0.933	0.730	40	0.932	−0.549	60	1.102	3.266

Table 2. Summary statistics of the item parameters of Table 1: minimum (Min), 25th percentile (P25), median, mean, 75th percentile (P75), maximum (Max), and correlation between discrimination and difficulty parameters (Cor)

Items	Par.	Min	P25	Median	Mean	P75	Max	Cor
20	*a*	0.749	0.930	0.961	1.014	1.139	1.214	.832
	b	−4.703	−0.593	0.374	0.374	1.907	3.972	
40	*a*	0.749	0.915	0.961	1.001	1.103	1.317	.784
	b	−4.726	−1.241	0.166	0.174	1.690	4.271	
60	*a*	0.749	0.915	0.991	1.016	1.124	1.416	.806
	b	−4.726	−1.405	−0.052	0.190	1.982	4.271	

Table 3. Proportions of response patterns affected by random guessing (%patterns) and proportions of item responses modified by random guessing (%resp)

		$\theta_0 = -3$		$\theta_0 = -2$		$\theta_0 = -1$		$\theta_0 = -0$	
Items	p^*	% Patterns	% Resp	% Patterns	% Resp	% Patterns	% Resp	% Patterns	% Resp
	.05	0.958	0.138	0.854	0.085	0.763	0.064	0.517	0.032
20	.10	0.977	0.163	0.936	0.122	0.763	0.064	0.801	0.070
	.15	0.983	0.176	0.968	0.148	0.878	0.092	0.801	0.070
	.05	0.995	0.115	0.974	0.079	0.930	0.059	0.759	0.032
40	.10	0.998	0.140	0.993	0.109	0.950	0.065	0.904	0.051
	.15	0.999	0.153	0.997	0.122	0.988	0.092	0.948	0.065
	.05	1.000	0.108	0.994	0.081	0.986	0.063	0.934	0.039
60	.10	1.000	0.141	0.999	0.105	0.988	0.067	0.980	0.060
	.15	1.000	0.155	1.000	0.114	0.998	0.090	0.989	0.069

4.3. Generation of response patterns

For any combination of test length (and related item parameters in Table 1) and true ability level, 5,000 response patterns were generated. Each pattern was obtained by randomly drawing the item responses as Bernoulli random variables with success probabilities given by the 2PL model, using the true ability level and the item parameters. In the absence of random guessing, response patterns were left unmodified.

The introduction of random guessing effect was performed as follows. First, a cut-off value p^* was selected as equal to .05, .10 or .15. Then, any item response with true response probability less than p^* was replaced by a random draw from a Bernoulli distribution with success probability .20. This process was used by Schuster and Yuan (2011) and translates the fact that when the probability of answering an item correctly is very low (i.e., below the cut-off p^*), the mechanism of random guessing with a .20 success probability is triggered. This motivates the restriction to very low ability levels in this case, since higher ability levels would almost never lead to true success probabilities less than p^*.

Table 3 provides a summary of the aberrant response generation process for each test length, true ability level, and cut-off value p^*. Two proportions are reported: the proportion of response patterns (among the 5,000 generated) affected by random guessing (i.e., such that at least one item response was modified), and the overall proportion of item responses that were modified due to random guessing. As expected,

both proportions increase with the cut-off value p^*, conditionally upon the test length and the true ability level. Moreover, lower ability levels are more often affected by random guessing than medium ability levels. Finally, increasing the test length usually increases the proportion of response patterns affected by random guessing (since more items are expected to be affected), but the overall proportion of affected item responses remain quite stable, except with very low ability levels.

4.4. Summary statistics

The design of this simulation study leads to 57 settings: 21 in the absence of random guessing (three test lengths and seven true ability levels) and 36 in the presence of random guessing (three test lengths, four true ability levels and three cut-off values). As mentioned earlier, 5,000 response patterns were drawn in each design setting.

For each pattern, four statistics were computed: the ML and robust ability level estimates $\hat{\theta}_{ML}$ and $\hat{\theta}_{ROB}$, and the estimated ASE of these estimators $\widehat{ASE}(\hat{\theta}_{ML})$ and $\widehat{ASE}(\hat{\theta}_{ROB})$. The robust estimates were obtained by using the Huber-type weight function (8). These values were eventually summarized by computing the averaged signed bias (ASB) and the root mean squared error (RMSE) of the ability level estimators on the one hand, and approximate ASB and RMSE values of the estimated ASEs on the other hand.

More precisely, let $\hat{\theta}_{ML,i}$ be the ith ML estimate of ability ($i = 1, \ldots, 5{,}000$). Then, the ASB and the RMSE of the ML estimator were computed as

$$ASB = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_{ML,i} - \theta_0), \tag{35}$$

and

$$RMSE = \sqrt{\frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_{ML,i} - \theta_0)^2}. \tag{36}$$

Corresponding values were obtained for the robust estimator by replacing the $\hat{\theta}_{ML,i}$ estimates by the respective $\hat{\theta}_{ROB,i}$ values.

The ASE values, on the other hand, should ideally be compared to the true (finite-sample) SE value, that is, the standard deviation of the full-sample distribution of the ability estimates. Because it is impossible to compute the full-sample distribution of these estimates (due to the large number of response patterns), the true SE was approximated as follows:

$$\widehat{SE}(\hat{\theta}_{ML}) = \sqrt{\frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_{ML,i} - \bar{\theta}_{ML})^2}, \text{ with } \bar{\theta}_{ML} = \frac{1}{5000}\sum_{i=1}^{5000}\hat{\theta}_{ML,i}. \tag{37}$$

This comes from the fact that the simulation design yields 5,000 ability estimates from the same set of item parameters, which constitutes a large sample of the distribution of the estimators. Moreover, set $\widehat{ASE}_i(\hat{\theta}_{ML})$ as the ith estimated ASE of the ML estimator (i.e., computed by equation (33) with the ith ML ability estimate). Then, the approximate ASB of the ASE of the ML estimator was computed as

$$ASB_{\text{approx}} = \frac{1}{5000}\sum_{i=1}^{5000}\left[\widehat{ASE}_i(\hat{\theta}_{\text{ML}}) - \widehat{SE}(\hat{\theta}_{\text{ML}})\right], \tag{38}$$

and

$$RMSE_{\text{approx}} = \sqrt{\frac{1}{5000}\sum_{i=1}^{5000}\left[\widehat{ASE}_i(\hat{\theta}_{\text{ML}}) - \widehat{SE}(\hat{\theta}_{\text{ML}})\right]^2}, \tag{39}$$

and similarly for the robust estimator.

5. Results

The results are displayed in four distinct parts, relating respectively to the ability level estimators in the absence of random guessing, the ability level estimators in the presence of random guessing, the estimated ASEs in the absence of random guessing and the estimated ASEs in the presence of random guessing.

5.1. Ability estimation in the absence of random guessing

Table 4 displays the ASB and RMSE values of both ability level estimators in the absence of response disturbances. As expected, both the ASB and the RMSE decrease with the test length, but the decrease is slower at the extremes of the ability scale (especially in the $\theta_0 = -3$ case). This is because most items are informative for average ability levels and fewer informative items are available at the extremes of the scale. Moreover, both estimators underestimate the true ability level when it is truly negative, and overestimate it when it is truly positive. This is a known trend for the ML estimator (Lord, 1983). Finally, the ASB values are comparable between the two estimators but

Table 4. ASB and RMSE values of the ML and robust ability level estimators in the absence of random guessing

		20 items		40 items		60 items	
θ_0	Stat	ML	ROB	ML	ROB	ML	ROB
−3	ASB	−0.051	−0.037	−0.051	−0.043	−0.038	−0.037
	RMSE	0.774	0.817	0.596	0.618	0.499	0.519
−2	ASB	−0.124	−0.118	−0.054	−0.054	−0.031	−0.033
	RMSE	0.826	0.862	0.556	0.583	0.438	0.457
−1	ASB	−0.072	−0.078	−0.025	−0.028	−0.010	−0.012
	RMSE	0.710	0.743	0.477	0.497	0.383	0.398
0	ASB	−0.004	−0.004	0.005	0.006	0.008	0.011
	RMSE	0.592	0.617	0.430	0.448	0.353	0.369
1	ASB	0.015	0.017	0.005	0.006	0.001	0.001
	RMSE	0.583	0.605	0.426	0.440	0.358	0.371
2	ASB	0.023	0.019	0.019	0.020	0.008	0.009
	RMSE	0.608	0.636	0.446	0.472	0.368	0.391
3	ASB	0.029	0.020	0.027	0.022	0.015	0.012
	RMSE	0.627	0.653	0.492	0.518	0.396	0.418

the RMSE is slightly larger for the robust estimator. This is not surprising as the robust method is mainly designed to correct for response disturbances, which are not present in this design. In the absence of response disturbances, ML estimation is expected to perform better than any alternative robust method. The overall superiority of the ML estimator (even if only slightly) in the absence of such disturbances is therefore not surprising.

5.2. Ability estimation in the presence of random guessing

The ASB and RMSE values of the ability level estimators in the presence of random guessing are displayed in Table 5. The ASB values are quite constant with the test length (conditionally upon the cut-off value p^*) but they clearly increase with the cut-off value p^*. The same trend can be observed for the RMSE but with a clear decrease with the test length. Moreover, the robust estimator is much less biased than the ML method, especially for very low ability levels, as expected. With higher ability levels the gap between the two methods is smaller in terms of bias, as the impact of random guessing is being reduced. Note that all ASB values are positive, indicating a general overestimation of the ability levels, as expected from the random guessing mechanism. The RMSE values confirm this trend, returning smaller values for the robust estimator with low ability levels, and comparable values for both estimators with higher ability levels. The bias reduction of the robust estimator (with respect to the ML method) is however important enough to ensure that the robust method outperforms the ML approach, in terms of overall lower RMSE values. Note that these findings are in agreement with Schuster and Yuan (2011) who used the same random guessing mechanism but a simpler underlying IRT model with fixed item parameters.

Table 5. ASB and RMSE values of the ML and robust ability level estimators in the presence of random guessing

			$\theta_0 = -3$		$\theta_0 = -2$		$\theta_0 = -1$		$\theta_0 = 0$	
Items	p^*	stat.	ML	ROB	ML	ROB	ML	ROB	ML	ROB
20	.05	ASB	1.539	0.646	0.765	0.192	0.453	0.098	0.218	0.067
		RMSE	1.797	1.192	1.150	0.880	0.877	0.760	0.672	0.638
20	.10	ASB	1.648	0.788	0.926	0.337	0.453	0.098	0.353	0.163
		RMSE	1.882	1.304	1.255	0.956	0.877	0.760	0.744	0.704
20	.15	ASB	1.690	0.858	1.006	0.432	0.510	0.149	0.353	0.163
		RMSE	1.921	1.375	1.308	1.018	0.902	0.780	0.744	0.704
40	.05	ASB	1.343	0.436	0.757	0.200	0.457	0.122	0.241	0.076
		RMSE	1.499	0.794	0.954	0.612	0.694	0.526	0.528	0.466
40	.10	ASB	1.447	0.557	0.900	0.303	0.484	0.138	0.311	0.122
		RMSE	1.590	0.894	1.071	0.670	0.714	0.535	0.573	0.496
40	.15	ASB	1.486	0.612	0.941	0.345	0.547	0.191	0.344	0.149
		RMSE	1.624	0.939	1.107	0.700	0.760	0.566	0.597	0.517
60	.05	ASB	1.297	0.382	0.786	0.200	0.526	0.141	0.313	0.102
		RMSE	1.399	0.652	0.922	0.496	0.686	0.438	0.506	0.398
60	.10	ASB	1.448	0.536	0.904	0.279	0.545	0.151	0.400	0.157
		RMSE	1.539	0.779	1.022	0.546	0.701	0.443	0.574	0.437
60	.15	ASB	1.483	0.588	0.930	0.302	0.600	0.195	0.423	0.175
		RMSE	1.572	0.823	1.044	0.559	0.743	0.472	0.591	0.449

5.3. ASEs in the absence of random guessing

Let us focus on the ASE values of these estimators. Table 6 returns the approximate ASB and RMSE values for these estimated ASEs in the absence of random guessing, similarly to Table 4. Overall, both approximate ASB and RMSE values are smaller in the middle of the ability scale, where more informative items are located, and larger at the extremes of the scale. The approximate ASB values are quite similar between both estimators, though smaller for the ML estimator with small tests. All approximate ASB values are positive (except for a few that are slightly negative and very close to 0), meaning that estimated ASEs are larger than or very close to their (estimated) SE values. Similarly, the approximate RMSE values are overall smaller for the ML estimator. Note, however, that the gap between the two estimated methods is quite small, especially when the test length increases. Eventually, both the approximate ASB and the RMSE decrease with the test length, as expected.

5.4. ASEs in the presence of random guessing

Finally, in the presence of random guessing, the approximate ASB and RMSE values of the ASEs are obtained as given in Table 7, displayed similarly to Table 5 for the ability level estimators. First, the approximate ASB is negative overall for both the ML method and the robust method. The magnitude of the approximate ASB decreases as the ability levels increase towards 0, which is expected since the impact of random guessing becomes smaller. Moreover, the robust ASE exhibits smaller approximate bias values than the ASE of the ML estimator overall. The bias reduction (from ML estimation to robust estimation) is more important with very low ability levels, while the approximate ASB values are comparable with higher ability levels. As for the ability estimators themselves, the approximate ASB tends to increase negatively (i.e., to become even more negative) with the cut-off value, but this mostly occurs at the extreme of the ability scale only. Finally, similar trends can be inferred for the approximate RMSE values, with an overall superiority of the robust ASE estimator with very low ability levels and quite comparable values at

Table 6. Approximate ASB and RMSE values of the ASE of the ML and robust ability level estimators in the absence of random guessing

		20 items		40 items		60 items	
θ_0	Approx. Stat.	ML	ROB	ML	ROB	ML	ROB
−3	ASB	0.246	0.288	0.058	0.080	0.018	0.032
	RMSE	0.308	0.363	0.097	0.123	0.054	0.069
−2	ASB	0.009	0.031	−0.010	−0.007	−0.007	−0.004
	RMSE	0.154	0.191	0.052	0.062	0.031	0.036
−1	ASB	−0.031	−0.03	−0.007	−0.006	−0.004	−0.005
	RMSE	0.09	0.112	0.027	0.033	0.015	0.017
0	ASB	0.000	−0.003	−0.001	−0.003	0.004	0.001
	RMSE	0.031	0.036	0.011	0.013	0.005	0.004
1	ASB	−0.007	−0.009	−0.004	−0.002	−0.001	0.001
	RMSE	0.014	0.019	0.007	0.008	0.003	0.005
2	ASB	−0.002	−0.002	0.002	−0.002	0.003	0.000
	RMSE	0.036	0.045	0.019	0.024	0.008	0.009
3	ASB	0.058	0.073	0.012	0.017	0.004	0.003
	RMSE	0.088	0.11	0.037	0.046	0.016	0.018

Table 7. Approximate ASB and RMSE values of the ASE of the ML and robust ability level estimators in the presence of random guessing

Items	p^*	Approx. Stat.	$\theta_0 = -3$		$\theta_0 = -2$		$\theta_0 = -1$		$\theta_0 = 0$	
			ML	ROB	ML	ROB	ML	ROB	ML	ROB
20	.05	ASB	−0.195	−0.057	−0.158	−0.033	−0.121	−0.065	−0.050	−0.023
		RMSE	0.240	0.227	0.196	0.171	0.138	0.117	0.056	0.041
20	.10	ASB	−0.192	−0.122	−0.165	−0.091	−0.121	−0.065	−0.073	−0.075
		RMSE	0.233	0.251	0.197	0.187	0.138	0.117	0.076	0.082
20	.15	ASB	−0.199	−0.168	−0.163	−0.133	−0.118	−0.082	−0.073	−0.075
		RMSE	0.238	0.278	0.192	0.209	0.133	0.125	0.076	0.082
40	.05	ASB	−0.152	−0.031	−0.097	−0.029	−0.074	−0.030	−0.045	−0.017
		RMSE	0.161	0.090	0.103	0.061	0.077	0.043	0.045	0.020
40	.10	ASB	−0.153	−0.080	−0.106	−0.056	−0.077	−0.036	−0.056	−0.038
		RMSE	0.161	0.117	0.112	0.077	0.080	0.047	0.057	0.040
40	.15	ASB	−0.154	−0.099	−0.110	−0.071	−0.083	−0.055	−0.063	−0.053
		RMSE	0.161	0.130	0.116	0.089	0.086	0.063	0.064	0.054
60	.05	ASB	−0.113	−0.025	−0.094	−0.019	−0.076	−0.027	−0.043	−0.015
		RMSE	0.118	0.059	0.096	0.037	0.076	0.031	0.043	0.015
60	.10	ASB	−0.118	−0.075	−0.094	−0.039	−0.076	−0.029	−0.056	−0.038
		RMSE	0.122	0.092	0.096	0.050	0.077	0.033	0.056	0.038
60	.15	ASB	−0.119	−0.090	−0.093	−0.042	−0.075	−0.044	−0.058	−0.044
		RMSE	0.122	0.105	0.095	0.052	0.076	0.046	0.058	0.044

higher ability levels. One noticeable exception occurs for very small tests, very low ability levels and large cut-off values. In this configuration, the approximate RMSE of the robust method is slightly larger than that of the ML method.

6. Discussion

The simulation study highlighted several important facts. First, the ML and robust estimators perform similarly in the absence of random guessing, with a slight advantage to the ML estimator in terms of RMSE. This was expected, as the robust estimator is not designed to outperform the standard estimators when there is no random guessing. On the other hand, the robust method is less biased and has lower RMSE values in the presence of random guessing, especially at low ability levels, for which the impact of guessing is maximal. This was also expected from Schuster and Yuan (2011), but was obtained under a more general IRT framework in the present study.

The main novelty arising from this study is the comparison of estimated ASEs with the true SEs, which were estimated on the basis of a large sample of response patterns. In the absence of random guessing, both estimated ASEs tend to overestimate the true SEs, with larger bias at the extremes of the ability scale where the test is less informative. The gap between the two ASEs is quite small, although that for the ML estimator tends to be slightly less biased and with smaller RMSE than the corresponding ASE for the robust estimator. In the presence of random guessing, however, the estimated ASE for the robust estimator is less biased than that for the ML estimator, especially at extremely low ability levels for which the impact of guessing is maximal. With ability levels closer to 0, both ASEs behave quite similarly and are almost unbiased. A similar trend can be observed for the RMSE values.

It can be conjectured that when the impact of random guessing is important, the robust estimator is overall less biased and performs better than the ML estimator, as expected, and this has a straightforward effect on its ASE which is estimated with lower bias too. In the absence of random guessing, however, both estimators behave similarly and the same holds for their ASEs. In such cases the estimated ability levels are probably very close together, which is an indirect indicator of absence of random guessing, while a larger gap in ability level estimates would reflect the presence of random guessing (and by extension, the presence of some response disturbance process).

The main aim of this paper is to provide a general formula for computing the ASE of the robust estimator of ability in dichotomous IRT models. The derivation of this formula is quite general and the simulation study highlighted its accuracy in the presence of random guessing. This study, however, has a limited scope and further extensions could be considered to complete the empirical findings. Four such extensions of interest are listed below.

First, this paper focused on one type of aberrant response mechanism, the random guessing process. Schuster and Yuan (2011) also considered transcription errors as an alternative process, but no major differences between the two mechanisms were highlighted. Other methods were proposed by, for example, St-Onge *et al.* (2011), and considering other aberrant response mechanisms might be a good way to generalize this study.

A second interesting extension of the study would examining the three-parameter logistic (3PL) model as the underlying IRT model for generating the data. However, the effect of random guessing would be downweighted by the lower asymptote levels of the 3PL model, as even very low ability levels do not lead to almost zero probability of a correct answer. Back to the random guessing mechanism of this study, if all lower asymptotes are fixed to .20 say, then any cut-off value p^* smaller than .20 will be useless since the true response probabilities will never become smaller than the cut-off value, and hence no random guessing will be introduced. This implies that another generating process for random guessing should be considered. Once again, alternative generation processes, such as the proposals by St-Onge *et al.* (2011), might be interesting competitive approaches.

Third, the simulation study considered the item parameters as fixed and the true values were considered for ability level estimation. In practice, however, these parameters should first be estimated from the data, or obtained from previous calibrations, at the risk of increasing estimation bias in the presence of response disturbances. Investigating how both ML and robust estimators behave in such a framework of ability level estimation with estimated item parameters would be very useful for practical research purposes.

Finally, there is no theoretical barrier to extending these derivations to ability level estimators in polytomous IRT models, provided that the latter remain one-dimensional. Only the structural form of the $g_i(\theta)$ functions will be modified and obviously more complex, but up to the possible addition of some regularity hypotheses on the structure of the polytomous models, the present framework should still be valid and lead to remarkably simple ASE formulas.

Acknowledgements

The authors wish to thank Stas Kolenikov and Timothy Johnson, Associate Editors, and two anonymous reviewers for fruitful suggestions that greatly improved the presentation of this paper. This research was funded by a postdoctoral grant 'Chargé de recherches' of the National

Fund for Scientific Research (FNRS), Belgium, the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy), and the Research Fund of the KU Leuven, Belgium.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Carroll, R. J., & Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B*, *55*, 693–706.

Donoghue, J. R., & Allen, N. L. (1993). Thin vs. thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics*, *18*, 131–154. doi:10.2307/1165084

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

Huber, P. J. (1981). *Robust statistics*. New York, NY: Wiley.

Hutton, P. (2006). Understanding student cheating and what educators can do about it. *College Teaching*, *54*, 171–176. doi:10.3200/CTCH.54.1.171-176

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating functions and the maximum likelihood estimators. *Annals of the Institute of Statistical Mathematics*, *25*, 1–26. doi:10.1007/BF02479355

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298. doi:10.1207/S15324818AME1604_2

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233–245. doi:10.1007/BF02294018

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research*, *46*, 733–755. doi:10.1080/00273171.2011.606757

Maronna, R. A., Martin, D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Chichester, UK: Wiley.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135. doi:10.1177/01466210122031957

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, *42*, 725–737. doi:10.1177/001316448204200302

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 75–106. doi:10.1007/BF02294745

Mosteller, F., & Tukey, J. (1977). *Exploratory data analysis and regression*. Reading, MA: Addison-Wesley.

Penfield, R. D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta Journal of Educational Research*, *49*, 231–243.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, *38*, 485–498. doi:10.2307/2530463

Rao, M. M. (1984). *Probability theory with applications*. New York: Academic Press.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: Wiley.

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, *36*, 720–735. doi:10.3102/1076998610396890

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342. doi:10.1007/BF02294437

Stefanski, L. A., & Boos, D. D. (2002). The calculus of M-estimation. *American Statistician*, *56*, 29–38. doi:10.1198/000313002753631330

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, *35*, 419–432. doi:10.1177/0146621610391777

Van der Linden, W., & Glas, C. A. W. (2009). *Elements of adaptive testing*. New York, NY: Springer.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, *45*, 373–391. doi:10.1007/BF02293910

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, *54*, 427–450. doi:10.1007/BF02294627

Yuan, K.-H. (1997). A theorem on uniform convergence of stochastic functions with applications. *Journal of Multivariate Analysis*, *62*, 100–109. doi:10.1006/jmva.1997.1674

Yuan, K.-H., & Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, *65*, 245–260. doi:10.1006/jmva.1997.1731

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*(9), 1–16.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233–251. doi:10.1111/j.1745-3984.1993.tb00425.x

Received 18 February 2013; revised version received 25 July 2013

Appendix

This section establishes that Assumptions 2, 4 and 6 are fulfilled for both the ML and robust estimators defined by (11) and (12).

Assumption 2 will be established by fixing $\delta = 1$. From (19) and (21), we have

$$\left|\frac{g_i(\theta_0)^2}{1+V_i}\right|^{1+\delta} = \left(\frac{\omega_i(\theta_0)^2}{1+\omega_i(\theta_0)^2 I_i(\theta_0)}\right)^2 \left(\frac{P_i'(\theta_0)}{P_i(\theta_0)Q_i(\theta_0)}\right)^4 (X_i - P_i(\theta_0))^4. \tag{A1}$$

Set $Y_i = [X_i - P_i(\theta)]^4$. Then, by definition of X_i, the random variable Y_i takes two possible values, $Q_i(\theta)^4$ with probability $P_i(\theta)$ and $P_i(\theta)^4$ with probability $Q_i(\theta)$. Hence,

$$E(Y_i) = P_i(\theta)Q_i(\theta)[P_i(\theta)^3 + Q_i(\theta)^3], \tag{A2}$$

and thus

$$E\left(\left|\frac{g_i(\theta_0)^2}{1+V_i}\right|^{1+\delta}\right) = \left(\frac{\omega_i(\theta_0)^2}{1+\omega_i(\theta_0)^2 I_i(\theta_0)}\right)^2 \frac{P_i'(\theta_0)^4}{P_i(\theta_0)^3 Q_i(\theta_0)^3}[P_i(\theta_0)^3 + Q_i(\theta_0)^3], \tag{A3}$$

which is bounded for any $i = 1, \ldots, n$. Assumption 2 is therefore fulfilled.

Now, to establish that Assumption 4 is fulfilled, we need the following result due to Chebyshev (see Rao, 1984, p. 57): if $Y_1, \ldots, Y_n$ is a sequence of independent random variables and if $S_n = \sum_{i=1}^n Y_i$ is such that

$$\frac{V(S_n)}{n^2} \to 0 \text{ as } n \to \infty, \tag{A4}$$

then S_n/n converges in probability to its expected value $E(S_n/n)$. In the present framework, set $Y_i = g_i'(\theta)$, so that

$$S_n = \sum_{i=1}^{n} g_i'(\theta) \text{ and } \frac{S_n}{n} = G_n'(\theta). \tag{A5}$$

According to (25), $g_i'(\theta)$ functions are linear combinations of the X_i variables and are therefore independent. Moreover, it follows from (12) that

$$g_i'(\theta) = \omega_i'(\theta)\frac{dl_i(\theta)}{d\theta} + \omega_i(\theta)\frac{d^2 l_i(\theta)}{d\theta^2}, \tag{A6}$$

so that

$$V(S_n) = \sum_{i=1}^{n}\left[\omega_i'(\theta)^2 V\left(\frac{dl_i(\theta)}{d\theta}\right) + \omega_i(\theta)^2\ V\left(\frac{d^2 l_i(\theta)}{d\theta^2}\right)\right], \tag{A7}$$

since $\omega_i(\theta)$ does not depend on X_i but $l_i(\theta)$ and its derivatives do. Now, according to (19) and (21),

$$V\left(\frac{dl_i(\theta)}{d\theta}\right) = I_i(\theta). \tag{A8}$$

Also, it is straightforward to show that

$$\frac{d^2 l_i(\theta)}{d\theta^2} = [X_i - P_i(\theta)]\left\{\frac{P_i''(\theta)}{P_i(\theta)Q_i(\theta)} - \frac{P_i'(\theta)^2[Q_i(\theta) - P_i(\theta)]}{P_i(\theta)^2 Q_i(\theta)^2}\right\} - I_i(\theta), \tag{A9}$$

(where $P_i''(\theta)$ stands for the second derivative of $P_i(\theta)$ with respect to θ), whence

$$V\left(\frac{d^2 l_i(\theta)}{d\theta^2}\right) = P_i''(\theta) - \frac{P_i'(\theta)^2[Q_i(\theta) - P_i(\theta)]}{P_i(\theta)Q_i(\theta)} = P_i''(\theta) - I_i(\theta)[Q_i(\theta) - P_i(\theta)], \tag{A10}$$

which is a bounded function of θ. In sum,

$$V(S_n) = \sum_{i=1}^{n} T_i, \tag{A11}$$

with

$$T_i = \omega_i'(\theta)^2 I_i(\theta) + \omega_i(\theta)^2\{P_i''(\theta) - I_i(\theta)[Q_i(\theta) - P_i(\theta)]\}, \tag{A12}$$

that is, a bounded function of θ for any item. It follows that

$$0 \leq \frac{1}{n^2} V(S_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{n} \leq \frac{1}{n} \sum_{k=1}^{n} \frac{T_k}{k} = \frac{1}{n} \sum_{k=1}^{n} A_k. \tag{A13}$$

Because T_k is bounded for any k, the sequence of variables A_k converges to 0 as $k \to \infty$, so the right-hand side of (A13) also converges to zero by Cesàro's theorem, which is sufficient to ensure that (A4) is fulfilled. Thus, $G_n'(\theta)$ converges in probability to its expected value, given by (27). This ensures that Assumption 4 is fulfilled.

Finally, Assumption 6 is assessed by noticing that

$$\|g_i'(\theta)\| = \left| [X_i - P_i(\theta)] \left(\omega_i(\theta) \frac{P_i'(\theta)}{P_i(\theta)Q_i(\theta)} \right)' - \omega_i(\theta) \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \right| \tag{A14}$$

and

$$\left(\omega_i(\theta) \frac{P_i'(\theta)}{P_i(\theta)Q_i(\theta)} \right)' = \omega_i(\theta) \frac{P_i'(\theta)}{P_i(\theta)Q_i(\theta)} + \omega_i(\theta) \frac{P_i'' P_i(\theta)Q_i(\theta) - P_i'(\theta)^2 [Q_i(\theta) - P(\theta)]}{P_i(\theta)^2 Q_i(\theta)^2}, \tag{A15}$$

and all terms in (A14) and (A15) are bounded by definition of X_i and due to the assumptions on the item response model.