Applied Psychological Measurement

<http://apm.sagepub.com/>

A Test-Length Correction to the Estimation of Extreme Proficiency Levels

David Magis, Sébastien Béland and Gilles Raïche

Applied Psychological Measurement 2011 35: 91 originally published online 28 September 2010

DOI: 10.1177/0146621610378289

The online version of this article can be found at:

<http://apm.sagepub.com/content/35/2/91>

Published by:

<http://www.sagepublications.com>

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: <http://apm.sagepub.com/cgi/alerts>

Subscriptions: <http://apm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations: <http://apm.sagepub.com/content/35/2/91.refs.html>

A Test-Length Correction to the Estimation of Extreme Proficiency Levels

Applied Psychological Measurement
35(2) 91–109
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621610378289
http://apm.sagepub.com

David Magis¹, Sébastien Béland², and Gilles Raïche²

Abstract

In this study, the estimation of extremely large or extremely small proficiency levels, given the item parameters of a logistic item response model, is investigated. On one hand, the estimation of proficiency levels by maximum likelihood (ML), despite being asymptotically unbiased, may yield infinite estimates. On the other hand, with an appropriate prior distribution, the Bayesian approach of maximum a posteriori (MAP) yields finite estimates, but it suffers from severe estimation bias at the extremes of the proficiency scale. As a first step, a simple correction to the MAP estimator is proposed to reduce this estimation bias. The correction factor is determined through a simulation study and depends only on the length of the test. In a second step, some additional simulations emphasize that the corrected estimator behaves like the ML estimator and outperforms the standard MAP method for extremely small or extremely large abilities. Although based on the Rasch model, the method could be adapted to other logistic item response models.

Keywords

item response theory, Rasch model, proficiency level, estimation, bias

In educational measurement, one is often faced with the issue of estimating the proficiency (or ability) level of examinees who have been assigned an aptitude test. Large assessment tests such as the SAT (Scholastic Aptitude Test) or the PISA (Program for International Student Assessment) are examples of major interest in that area. The accurate estimation of the proficiency levels of examinees is therefore a crucial issue, and several approaches have been proposed to deal with it. One important framework is item response theory (IRT), which aims at binding the proficiency level of one examinee to a set of item characteristics by means of an appropriate item response model. One of the early item response models proposed in the literature was introduced by Rasch (1960) and is nowadays referred to as the Rasch or one-parameter logistic (1PL) model.

¹University of Liège and K. U. Leuven, Belgium

²Université du Québec à Montréal, Canada

Corresponding Author:

David Magis, Department of Mathematics (B37), University of Liège, Grande Traverse 12, B-4000 Liège, Belgium

Email: David.Magis@ulg.ac.be

This model can be considered to be a particular case of the more general three-parameter logistic (3PL) model (Birnbaum, 1968):

$$P_j(\theta_i) = \Pr(Y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j, \chi_j) = \chi_j + (1 - \chi_j) \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]}, \quad (1)$$

where Y_{ij} is the response of subject i to item j (with values 0 and 1 for, respectively, an incorrect or a correct response); θ_i is the proficiency level of subject i ; and α_j , β_j , and χ_j denote the discrimination, the difficulty, and the pseudo-guessing levels of item j . Simplification to the two-parameter logistic (2PL) is achieved by fixing the pseudo-guessing parameter χ_j to zero, whereas the Rasch model is obtained by additionally restricting the discrimination parameter α_j to unity.

Estimation of both person and item parameters can be performed in several ways. A first possibility is to obtain joint estimates of the parameters, for instance by joint maximum likelihood (Lord, 1980; Lord & Novick, 1968). However, person proficiency levels are nuisance parameters and it is better to estimate item parameters unconditionally toward these. Accordingly, several techniques have been suggested, such as conditional maximum likelihood (Andersen, 1970, 1972), marginal maximum likelihood (Bock & Aitkin, 1981), and Bayesian modal estimation (Swaminathan & Gifford, 1985, 1986), among others (see also Lord, 1986). Once the item parameters are estimated, they are kept as fixed values, and the proficiency levels of the subjects can then be estimated conditionally on the item parameters. The latter step is sometimes called “conditional estimation” of proficiency level (Hambleton & Swaminathan, 1985), but this should not be confused with the terminology introduced by Andersen (1970) to denote the item parameter estimation technique. In the present study, the framework is restricted to conditional estimation of proficiency levels given that the item parameters are known in advance. Consequently, for better clarity, the term *conditional* is withdrawn from the forthcoming discussion.

However, it is quite difficult to obtain reliable estimates of proficiency levels if these levels are extremely large or extremely small. Most of the estimators exhibit a large bias for such extreme proficiency levels. For instance, the maximum likelihood estimates of proficiency level become infinite with extremely large or extremely small proficiency levels (Hambleton & Swaminathan, 1985). Making use of a prior distribution usually avoids this issue, but it also results in more severely biased estimates. Several authors (Firth, 1993; Samejima, 1993, 1998; Warm, 1989) have suggested alternative methods to reduce the overall bias of an estimator, without specifically focusing on the extreme proficiency levels (see also Wang, Hanson, & Lau, 1999).

The present approach is somewhat similar, but with the clear objective of reducing the estimation bias at the extreme values of the ability scale by applying a simple test-length correction factor to the well-known maximum a posteriori (MAP) estimator. The next section begins with a brief review of the usual likelihood-based methods of proficiency level estimation, together with their strengths and weaknesses (especially in the field of extreme proficiency level estimation). A bias-corrected estimator is then proposed to improve the reliability of the estimation of extreme proficiency levels. The derivation of the corrected estimator is then performed in the section “A Bias-Corrected MAP Estimator,” and its accuracy in reducing the bias of the extreme proficiency levels is assessed in the next section.

Estimation of Proficiency Level

When the item parameters are fixed, the estimation of proficiency level can be performed by several methods (for a review, see Baker, 1992, or Hambleton & Swaminathan, 1985). The present study focuses on two common likelihood-based approaches: the maximum likelihood (ML) estimation and the MAP estimation. Both are described below, in the general framework of any

dichotomous item response model characterizing the probability $P_j(\theta)$ of a correct response to item j (the person subscript i is removed to simplify the notations).

Maximum Likelihood Estimation

Let J be the length of the test (or equivalently, the number of items to be administered). The ML estimation of proficiency level θ is performed by determining the value $\hat{\theta}_{ML}$ of θ that maximizes the likelihood function

$$L(\theta) = \prod_{j=1}^J P_j(\theta)^{Y_j} Q_j(\theta)^{1-Y_j}, \quad (2)$$

where $Q_j(\theta) = 1 - P_j(\theta)$ is the probability of an incorrect response to item j . For mathematical ease, it is useful to determine $\hat{\theta}_{ML}$ by maximizing the log-likelihood function

$$\log L(\theta) = \sum_{j=1}^J \{Y_j \log P_j(\theta) + (1 - Y_j) \log Q_j(\theta)\}. \quad (3)$$

The maximum likelihood estimator $\hat{\theta}_{ML}$ has some desirable statistical properties: it is asymptotically unbiased and normally distributed (Cramer, 1946), and its asymptotic standard error is related to the information function (Baker, 1992; Samejima, 1977). Moreover, if the item response model is appropriately chosen, this estimator is unique. This is the case for the 2PL model and for the Rasch model. However, it has been shown (Samejima, 1973) that Birnbaum's 3PL model may lead to several local maxima, especially with short tests. This phenomenon seems to disappear as the length of the test increases, and Lord (1980) suggests a minimum test length of 20 items to avoid this happening. Note that Magis and Raïche (2010) proposed an iterative estimation method to determine the number and the values of the different local maxima of the likelihood function, if any.

The maximum likelihood approach has another drawback, whatever the logistic model under consideration. Examinees with an extremely low or extremely high proficiency level are more likely to provide only incorrect or only correct responses to the test, a situation that implies infinite ML estimates (Hambleton & Swaminathan, 1985). This can be directly observed from Equation 2: If the response pattern holds only correct responses (i.e., $Y_j = 1$ for all items), the likelihood function $L(\theta)$ becomes a product of strictly increasing functions and $\hat{\theta}_{ML}$ is therefore equal to $+\infty$ (similarly, only incorrect responses yield $-\infty$ as the value for $\hat{\theta}_{ML}$). To overcome this issue, one usually restricts the range of values for the estimated proficiency level, so that the maximum likelihood method provides finite estimates. However, it is important to notice that applying such a range restriction is equivalent to estimating the proficiency level with a Bayesian estimator and a uniform noninformative prior distribution on the selected range, as is indicated in the next section. So, strictly speaking, restricting the proficiency levels to a range of allowable values no longer provides a maximum likelihood estimate.

MAP Estimation

The second estimator of proficiency level comes from the Bayesian paradigm and can be described as follows. In addition to the likelihood function $L(\theta)$, one also provides a prior distribution of the proficiency level, say $f(\theta)$. The posterior distribution $g(\theta)$ of θ is defined as the product of the likelihood function and the prior distribution: $g(\theta) = f(\theta)L(\theta)$. Therefore, in a similar way to the ML estimate, the MAP estimator is the value $\hat{\theta}_{MAP}$ of θ that maximizes the posterior distribution $g(\theta)$, or its logarithm $\log g(\theta) = \log f(\theta) + \log L(\theta)$ for mathematical simplicity.

The choice of an accurate prior distribution $f(\theta)$ of the proficiency level is the central key point of MAP estimation. Noninformative priors, such as the uniform distribution (restricted to a finite interval), are frequently used in Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 2004). In this context, the uniform prior has an effect of limiting the range of allowable θ values and avoids the issue of an infinite estimate of proficiency, as indicated earlier. However, the usual selection of prior distribution is the standard normal distribution $N(0, 1)$, which is an informative prior. By using this prior distribution, it is assumed that the proficiency level is symmetrically distributed around the central value of zero with a standard deviation of one unit. The remainder of this study is focused on this standard normal distribution as the prior density for MAP estimation.

Using the standard normal distribution as the prior density implies that the estimated proficiency levels are shrunk toward the prior mean value (zero in this context) and that the MAP estimator is less variable (has lower standard errors) than the ML estimator (Baker, 1992). On the other hand, it also implies an increase in the estimation bias, especially at the extremes of ability level (Lord, 1983, 1986). Another advantage of MAP is that it avoids the issue of an infinite estimate, because the posterior distribution converges to zero as the proficiency level θ becomes infinite. It is, nevertheless, not guaranteed that multiple local maxima will not occur, because this depends on the appropriate selection of both the prior density and the selected model. For the 2PL model, the uniqueness of the MAP estimator is guaranteed (see, e.g., Magis & Raiche, 2010).

In sum, the MAP estimator usually has a lower standard error but a higher bias, especially for extremely large or extremely small proficiency levels. One possible correction to this bias would be to modify the prior normal distribution by shifting its mean toward larger or smaller values. Making use of a prior distribution derived from the observed sample data (Carlin & Louis, 1996; Mislevy, 1984) is another solution, but it requires both an appropriate method and a sufficiently large number of items to obtain improved results. The approach the authors want to present here is based on an ad hoc adjustment of the MAP estimator by reducing its bias, mainly at the extremes of the proficiency scale.

A Bias-Corrected MAP Estimator

Because the traditional MAP estimator (with prior standard normal density) always exists but is severely biased for extreme proficiency levels and for small to moderate test lengths, the relationship between the (empirical) bias of MAP and the length J of the test was first studied. Empirical findings were then used to correct the MAP estimator in order to reduce its bias. The Rasch model was focused on for the sake of simplicity, though all the developments discussed below could be easily adapted to the 2PL and the 3PL models (the adjustment would, however, certainly be different).

Intuitive Approach

Prior to evaluating the bias of the MAP estimator, several working assumptions are necessary. First, only the Rasch model is implied to generate the response patterns, as already mentioned. Second, the relationship between the bias of the MAP estimator $\hat{\theta}_{MAP}$ and the proficiency level θ is assumed to be linear in θ . In other words, the bias of $\hat{\theta}_{MAP}$, defined by $B(\hat{\theta}_{MAP}) = E(\hat{\theta}_{MAP} - \theta)$, is linked to the proficiency level θ by the relationship

$$B(\hat{\theta}_{MAP}) \approx a\theta + b, \quad (4)$$

where the slope a and the intercept b have to be determined. Lord (1986) noticed that $B(\hat{\theta}_{MAP})$ is curvilinear in θ rather than simply linear, but that the nonlinear part of this relationship

mostly occurs at the extreme values of the proficiency level. On the other hand, Lord (1983) also established that the (asymptotic) bias of $\hat{\theta}_{MAP}$ is equal to the (asymptotic) bias of $\hat{\theta}_{ML}$ plus a linear function of the proficiency level θ . Because the ML estimator is asymptotically unbiased, it is reasonable to assume a linear relationship between $B(\hat{\theta}_{MAP})$ and θ , as described in Equation 4. Nevertheless, it was decided to investigate this assumption throughout the simulations.

The final assumption, applying a standard normal prior distribution, considers that the bias of the MAP estimator decreases with θ and is approximately equal to zero for proficiency levels close to the prior mean value. This means that $\hat{\theta}_{MAP}$ overestimates negative proficiency levels and underestimates positive ones, as already pointed out by Lord (1983), among others. This assumption is also a consequence of the fact that MAP estimates are shrunk toward the prior mean value, as described earlier. However, it was planned to investigate its accuracy by means of the simulations. Relationship 4 can therefore be simplified by the following:

$$B(\hat{\theta}_{MAP}) \approx -a\theta, \quad a > 0. \quad (5)$$

Because $B(\hat{\theta}_{MAP}) = E(\hat{\theta}_{MAP} - \theta)$, it follows from Equation 5 that the average value of $\hat{\theta}_{MAP}$ is approximately equal to $(1 - a)\theta$, and an easy way to cancel the bias is to correct it by the multiplicative factor $1/(1 - a)$. This yields a corrected MAP estimator of the proficiency level, which will be further denoted by C-MAP:

$$\hat{\theta}_{C-MAP} = \frac{\hat{\theta}_{MAP}}{1 - a}. \quad (6)$$

Assuming that the parameter a is different from unity, the C-MAP estimator is well defined and shares the same properties of the MAP estimator: it always exists whenever using the standard normal prior distribution and it is unique for an accurate choice of the logistic model (1PL or 2PL, or 3PL with a sufficiently large test length J). The multiplicative correction is expected to reduce the bias of the MAP estimator, though it will simultaneously increase the standard error of the estimator.

Simulation Study

To derive an accurate value for parameter a , the following simulation study was conducted. R software (R Development Core Team, 2008) was used, especially the *ltm* package (Rizopoulos, 2006), to perform the simulations and the computations. The R code can be obtained by simple request to the authors.

First, the length of the test varied from 5 items to 100 items by increments of five units. For each test length, the linear relationship between $B(\hat{\theta}_{MAP})$ and θ , given by Equation 4, was evaluated by replicating the following process 100 times.

Seventeen values θ_t of the proficiency level were selected, from -4 to 4 , by increments of 0.50 . For each θ_t value, 1,000 sets of J item difficulties were drawn from a uniform distribution $U([-4; 4])$. For each set of drawn item parameters, a response pattern was randomly generated by drawing each response Y_{ij} from a Bernoulli distribution with the success probability $P_j(\theta_t)$ (where the discrimination and pseudo-guessing parameters are all fixed to one and zero, respectively). Hence, for each proficiency level θ_t , 1,000 response patterns were available, each pattern corresponding to a specific set of item difficulties. If $\hat{\theta}_{MAP,k}^t$ denotes the MAP estimate of the k th response pattern ($k = 1, \dots, 1000$), the bias of $\hat{\theta}_{MAP}$ is approximated for each proficiency level θ_t by

$$\hat{B}(\hat{\theta}_{MAP}) = \frac{1}{1000} \sum_{k=1}^{1000} (\hat{\theta}_{MAP,k} - \theta_t). \quad (7)$$

This process therefore provides 17 pairs $(\hat{B}(\hat{\theta}_{MAP}), \theta_t)$ of both the proficiency level and the empirical bias of MAP at this level. These 17 pairs are eventually summarized by a triplet of values. The first value of this triplet is the Pearson linear correlation between the proficiency levels θ_t and the estimated bias $\hat{B}(\hat{\theta}_{MAP})$. The other two values of the triplet are the intercept and the slope parameters estimated by a linear least squares regression of $\hat{B}(\hat{\theta}_{MAP})$ with respect to θ_t .

In sum, for each test length J , the 100 replications of the process described above yields 100 triplets of values (Pearson correlation intercept and slope), one triplet per replication. To illustrate this idea, three replications of the process were represented when the test length was fixed to 20 items. Figure 1 displays the three bias curves, each particular value corresponding to the empirical bias calculated with Equation 7. The upper right legend summarizes the three aforementioned quantities, namely, the Pearson correlation, the intercept, and the slope of the linear regression, for each of the three replications. The curves of Figure 1 are provided to better understand the simulation scheme; they should not be directly analyzed as such because they represent only a fraction of the full process.

The Pearson correlations are required to assess the quality of the linear relationship in Equation 4, which is one of the central assumptions of the present approach. The investigation of the empirical intercept parameters is also mandatory to ensure that the working hypothesis leading to Equation 5 is valid in this framework. Eventually, a careful description of the slope parameters will yield the requested correction of the MAP estimator, given in Equation 6. Each of these three quantities will be summarized by means of “box-and-whiskers” plots or “boxplots” (Tukey, 1977). Such boxplots are very useful to describe a one-dimensional distribution, by representing smaller and larger values together with the median, the first and the third quartiles of the empirical distribution.

Results

The results are discussed in three parts: the correlations between the empirical bias and the proficiency levels, the intercept parameters, and the slope parameters.

Pearson correlations. Figure 2 displays the distributions of the Pearson correlations for each test length. Three major conclusions can be drawn from Figure 2. First, all correlations are negative, which confirms that the slope in Equation 4 has to be negative. Second, these correlations are all very close to -1 . The linear approximation of the relationship between $B(\hat{\theta}_{MAP})$ and θ is therefore an acceptable assumption. Also, the correlations tend to depart from -1 as the length of the test increases, and more variability is observed in that case. However, all correlations remain smaller than $-.97$. The linear relationship is much more likely to be observed for very small sets of items, and the curvilinear trend described by Lord (1986) is more present with larger tests. In conclusion, the linear approximation given in Equation 4 is an acceptable working hypothesis.

Intercept parameters. The distributions of the intercept parameters are displayed in Figure 3, using boxplots in a similar way as for Figure 2. For any fixed J , variability is present in the distributions of intercepts. This variability, however, remains very small and the parameters are roughly centered around zero. The simulations therefore confirm the simplification from Equation 4 to Equation 5, which means that on average, the bias of $\hat{\theta}_{MAP}$ becomes closer to zero when the proficiency level θ also equals zero.

Slope parameters. Finally, Figure 4 proposes the slope parameters in a similar way to the intercept parameters from Figure 3. Note that the slopes have negative values because of the negative correlations of Figure 2, but their absolute values have been displayed in order to better represent

Figure 1. Example of relationship between $B(\hat{\theta}_{MAP})$ and θ

Figure 2. Boxplots of the linear correlations between the empirical bias of maximum a posteriori (MAP) estimator and the proficiency level

the relationship between J and the slope parameter a from Equation 4. There is a very modest variation in the slope parameters for each fixed J , because the displayed boxplots reduce almost to a single trait (actually the largest range of the slope parameters is equal to .007). Moreover, the slopes tend to converge toward zero as the test length increases, which is expected because the

Figure 3. Boxplots of the intercept parameters from the linear regression of $B(\hat{\theta}_{MAP})$ with respect to θ

Figure 4. Boxplots of the slope parameters from the linear regression of $B(\hat{\theta}_{MAP})$ with respect to θ

bias decreases with larger tests. Thus, the simulations emphasize that very little variation occurs in both the intercept and the slope parameters through the 100 simulated runs.

Because there is very little variability in the slope parameters for any test length J , a convenient relationship between a and J can be approximated as follows. First, it is of interest to derive

a simple formula for a , which depends on J only and fulfills both the requirements of Equations 5 and 6 and the empirical findings of Figure 4. It follows, therefore, that the family of functions

$$\wp = \left\{ f_c(J) = \frac{c}{c+J} : c > 0 \right\} \quad (8)$$

is a convenient choice for an approximation $\hat{a}$ of a . Indeed, for any positive constant c , the function $f_c(J)$ takes positive values between zero and one, the lower and upper bounds being reached, respectively, when the test size J grows infinitely or is equal to zero. Moreover, $f_c(J)$ is a decreasing function of J and follows a negative exponential trend, which is required based on Figure 4. Note that other families of functions could be considered instead, based for instance on an exponential or inverse square root relationship between a and J . Nevertheless, the family in Equation 8 was focused on because of its simpler form.

Finally, a nonlinear optimization routine was used to select a convenient value for the constant c , that is, by obtaining the best fit between the function $f_c(J)$ and the empirical slope parameters. It turned out that setting $c = 9.9996$ yields the best fit to the data, which can be reasonably approximated by the simple value of 10. In conclusion, the distribution of the slope parameters can be approximated by $\hat{a} = 10/(10 + J)$. This function is displayed in Figure 4 by means of triangles linked by a dashed line. The approximate trend described by $\hat{a}$ is not perfect. It is most suitable for small J , even though it tends to slightly underestimate the (absolute) slope values as J becomes large.

Summary

From the simulations, the corrected MAP estimator can now be rewritten by replacing a with $\hat{a}$ in Equation 6:

$$\hat{\theta}_{C-MAP} = \frac{J+10}{J} \hat{\theta}_{MAP}. \quad (9)$$

The correction factor $(J+10)/J$ is always larger than one, meaning that $\hat{\theta}_{C-MAP}$ provides larger proficiency level estimates (in absolute values) than $\hat{\theta}_{MAP}$. This allows a reduction in the bias of extreme proficiency levels, as explained earlier. Moreover, when the length of the test increases, the correction factor decreases to one, and hence, the correction to the MAP estimate is less significant. For very large tests, the MAP and C-MAP estimators provide very close estimates, whereas C-MAP differs substantially from MAP for small tests.

It is important to emphasize that the bias correction introduced in Equation 9 was derived by using a prior standard normal density for the proficiency levels. In practice, however, it is possible that some other prior distributions could be considered, based for instance on previous knowledge regarding the distribution of examinees' proficiencies. Thus, the present approach is mostly validated in the usual framework of a prior standard normal density. This issue is addressed in more detail at the end of the article.

The C-MAP estimator reduces the bias in $\hat{\theta}_{MAP}$. On the other hand, because the multiplicative correction factor is larger than one, the standard error $SE(\hat{\theta}_{C-MAP})$ of the C-MAP estimator is also a multiple of the standard error $SE(\hat{\theta}_{MAP})$ of $\hat{\theta}_{MAP}$:

$$SE(\hat{\theta}_{C-MAP}) = \frac{J+10}{J} SE(\hat{\theta}_{MAP}). \quad (10)$$

It is therefore necessary to investigate whether the C-MAP estimator is really useful in estimating proficiency levels. In other words, one has to evaluate whether the gain in bias reduction

compensates for the increase in the standard error. One easy way to determine this is to compute the mean squared error (MSE), given by

$$MSE(\hat{\theta}_{C-MAP}) = B(\hat{\theta}_{C-MAP})^2 + SE(\hat{\theta}_{C-MAP})^2, \quad (11)$$

and to determine whether C-MAP performs better in terms of MSE. It is expected that for extreme proficiency levels the bias reduction in MAP bias would be greater, so the C-MAP estimator would perform better than the MAP method.

Another issue of interest is to compare the efficiency of the C-MAP method to the usual ML approach. This is the topic of the next section.

Assessing the Accuracy of C-MAP

In the previous section, a bias correction to the MAP estimator was established on the basis of a large simulation study. Following from this, it was of interest to determine whether the C-MAP estimator outperforms the classical ML and MAP methods. Therefore, the comparative simulation study described below was conducted. The design of the study is similar to the first step of the simulation study described in the section "Simulation Study." The same notations are used in this section.

Method

Four test lengths were selected, ranging from very small ($J = 10$) to small ($J = 20$), moderate ($J = 40$), and large ($J = 60$) test lengths. The values of the proficiency level varied from -4 to 4 by increments of 0.50 . For each test length and each proficiency level θ_i , 1,000 replications were obtained by drawing the J item difficulty parameters from a uniform distribution each time over the range $[-4, 4]$. The response patterns were then generated by drawing binary responses Y_{ij} from a Bernoulli distribution with the success probability given by $P_j(\theta_i)$, using Equation 1 and the appropriate item parameters (i.e., fixing the discriminations at one and the pseudo-guessing parameters at zero). This design was used to obtain the same number of subjects for each value of proficiency level, and it enabled computation of the bias, the standard error, and the root mean square error (RMSE) of the methods with the same sample size for all θ_i values. This is, however, different from the usual simulation design where the proficiency levels are drawn from a standard normal distribution. This case is discussed later on.

The estimation of the proficiency level was performed with the ML, MAP, and C-MAP methods, and a maximum of 1,000 estimates were returned for each method (given J and the proficiency level). For the ML method, only finite estimates were retained, so the patterns with only-correct or only-incorrect responses were withdrawn from the simulation results. The bias was then estimated using Equation 7 (for ML, the mean was calculated on the set of finite estimates only), whereas the standard error of the estimator was approximated by computing the sample standard deviation of the estimates. Empirical values of the RMSE were then derived from Equation 11. In addition, the proportion of patterns with only-correct or only-incorrect responses was recorded as a supplementary empirical result.

Results

The cases of infinite ML estimates are first discussed followed by discussions of the empirical bias, the standard error, and the RMSE values.

Infinite ML estimates. The proportions of response patterns yielding infinite ML estimates of the proficiency levels are displayed in Table 1. Note that only extreme proficiency levels are

Table 1. Proportions of Only-Correct or Only-Incorrect Response Patterns for Extreme Proficiency Levels

Proficiency level										
Items	-4	-3.5	-3	-2.5	-2	2	2.5	3	3.5	4
10	.48	.35	.22	.11	.06	.06	.12	.22	.30	.49
20	.22	.12	.05	.01	.01	.00	.02	.05	.12	.23
40	.11	.04	.01	.00	.00	.00	.00	.01	.04	.10
60	.06	.01	.00	.00	.00	.00	.00	.00	.01	.05

considered because the values within the range $[-2, 2]$ almost never led to such patterns (except in a very small number of cases and only with a test length of 10 items).

As expected, the proportions of only-correct or only-incorrect response patterns diminish as the length of the test increases and the proficiency level gets closer to zero. The proportions are also rather symmetric around the prior average proficiency level of zero. Note that for small numbers of items, the proportions of only-correct and only-incorrect patterns are quite large, especially for proficiency levels of -4 and 4 .

Bias. Figure 5 displays the curves of the empirical bias. The ML, MAP, and C-MAP estimators are represented respectively by means of circles, triangles, and crosses. First, it can be observed, as expected, that the ML and C-MAP estimators are much less biased than the MAP estimator. Moreover, except for very short test lengths ($J = 10$), both the ML and C-MAP methods have comparable bias curves. With 10 items, C-MAP is even slightly less biased than ML at extreme values of the proficiency levels. As the length of the test increases, all three estimation methods tend to be almost unbiased, though MAP still presents some large bias values at the extreme values of θ .

Note also that the ML and MAP bias curves indicate the same overall trend, which is in contradiction to the usual findings regarding these two estimators (e.g., Lord, 1986). One possible explanation is that the empirical bias for the ML method is calculated on the set of finite estimates only, and Table 1 emphasizes that the proportion of finite ML estimates is very small at the extreme values of θ . In most common proficiency level estimation software, the estimated proficiency values are usually restricted to a particular range. For instance, the functions from the R package *ltm* (Rizopoulos, 2006) limit the ML estimates to the range $[-4, 4]$. With such a restriction, the ML estimates for only-correct and only-incorrect patterns are, respectively, equal to the upper and lower bounds of that range, with the direct consequence that the ML bias shows an opposite trend to that of the MAP method.

Standard errors. The curves of the standard errors, displayed in Figure 6, are also as expected. First, the standard errors are much smaller with MAP, and with ML, than with C-MAP, as predicted by Equation 10. The three curves are rather symmetric around the a priori average proficiency level of zero. Furthermore, with very short tests, C-MAP has the largest standard errors, whereas for longer tests, ML and C-MAP show similar values. Nevertheless, it is important to remember that the standard errors for ML were computed with fewer proficiency level estimates than those for C-MAP and MAP, especially for extreme proficiency levels. The three methods behave almost identically for large tests. Note that the MAP and C-MAP curves are roughly parallel, which is a consequence of the correction factor $J/(J + 10)$ in Equation 10.

RMSE. The findings regarding the bias and the standard error can be summarized by the examination of the RMSE values displayed in Figure 7. Both the ML and C-MAP methods are very close in terms of the mean squared error, which is obvious following what was observed for the bias and the standard error separately. Moreover, these estimation methods show much smaller

Figure 5. Empirical bias for ML (circles), MAP (triangles), and C-MAP (crosses) estimators, Rasch model
 Note: ML = maximum likelihood; MAP = maximum a posteriori; C-MAP = corrected maximum a posteriori.

RMSE values for extremely low and extremely high proficiency levels when $|\theta|$ is greater than 2. For proficiency levels within the $[-2, 2]$ range, the MAP estimator performs best, especially on small tests. This is because of the use of the standard normal distribution as the prior density. Accordingly, for θ values close to the prior mean, MAP is only slightly biased and is less variable, which results in smaller mean squared errors. For more extreme proficiency levels, the decrease in the standard error cannot compensate for the bias of the MAP estimation, and therefore, both the ML and C-MAP methods are more efficient.

Normally Distributed Parameters

The previous comparative study highlighted the fact that the C-MAP estimator performs very well in decreasing the estimation bias with respect to the ML and MAP estimators. However, several conceptual issues can arise from the design of the study. One major drawback of this study is that both the item difficulties and the proficiency levels were randomly drawn from

Figure 6. Empirical standard errors for ML (circles), MAP (triangles), and C-MAP (crosses) estimators, Rasch model

Note: ML = maximum likelihood; MAP = maximum a posteriori; C-MAP = corrected maximum a posteriori.

nonnormal distributions. For the item difficulties, the uniform distribution on the range $[-4, 4]$ was used. For the proficiency levels, the sequence of values from -4 to 4 by steps of one half was selected, and the same number of examinees was considered for each value of θ . In practice, however, very easy and very difficult items are much less frequent than items with average difficulty level. Moreover, it is not surprising to observe a large bias with the MAP method, because of the gap between the prior standard normal density and the discrete uniform distribution that was used to generate the response patterns. An additional issue comes from the fact that each particular value of the proficiency level was assigned a set of different item difficulty parameters, so no full data matrix was generated on the basis of the same items. In practice, this is unrealistic because the same test is given to a set of subjects with different proficiency levels.

To complete the present discussion, a smaller, but more realistic, additional simulation study was conducted. The data matrix consisted of 4,000 examinees who were assigned the same test of 40 items. Both the item difficulties and the proficiency levels were randomly drawn from

Figure 7. Empirical RMSE values for ML (circles), MAP (triangles), and C-MAP (crosses) estimators, Rasch model

Note: RMSE = root mean square error; ML = maximum likelihood; MAP = maximum a posteriori; C-MAP = corrected maximum a posteriori.

a standard normal distribution. With this large number of examinees, it is expected that the 100 least able examinees have a proficiency level smaller than -2 , and the 100 most able examinees have a proficiency level higher than 2 . The response patterns were generated using a similar approach to that described in the section “Assessing the Accuracy of C-MAP.” The proficiency levels were estimated using the ML, MAP, and C-MAP methods. Note that the generated data set did not hold any only-correct or only-incorrect response patterns, so the issue of infinite ML estimates does not arise.

Because of the random drawing of the true proficiency levels, the bias, standard error, and RMSE were computed for all θ belonging to some predefined intervals. The considered intervals are of the kind $[-\infty, -2.5]$, $[-\infty, -2]$, and so on, until $[-\infty, 0]$ (and similarly for positive values). Table 2 displays the empirical bias, standard error, and RMSE values for the three methods and according to these ranges of θ values. In addition, the frequencies of proficiency levels are provided within each range.

Table 2. Comparison of Bias, Standard Error (SE) and Root Mean Squared Error (RMSE) of ML, MAP, and C-MAP Estimators for the Simulated Data From Standard Normal Proficiency and Difficulty Distributions, With 40 Items and Various Ranges of the True Proficiency Level

θ	Frequency	Bias			SE			RMSE		
		ML	MAP	C-MAP	ML	MAP	C-MAP	ML	MAP	C-MAP
≤ -3	8	.120	.871	.291	.362	.185	.225	.382	.890	.368
≤ -2.5	41	-.051	.603	.055	.492	.316	.371	.495	.681	.375
≤ -2	119	-.039	.464	-.021	.535	.350	.416	.536	.581	.417
≤ -1.5	285	-.011	.351	-.063	.464	.342	.402	.465	.490	.406
≤ -1	655	-.017	.236	-.097	.428	.349	.407	.429	.421	.419
≤ -0.5	1231	-.007	.170	-.083	.406	.346	.408	.406	.385	.416
≤ 0	2037	-.013	.106	-.069	.391	.346	.406	.391	.362	.412
≥ 0	1963	.045	-.092	.096	.411	.348	.415	.414	.359	.426
≥ 0.5	1265	.056	-.138	.120	.434	.351	.420	.437	.378	.437
≥ 1	676	.074	-.202	.134	.468	.355	.424	.474	.408	.444
≥ 1.5	303	.103	-.287	.125	.525	.361	.430	.535	.461	.448
≥ 2	107	.136	-.404	.076	.580	.355	.422	.596	.538	.429
≥ 2.5	21	.266	-.575	-.008	.673	.389	.451	.724	.694	.451
≥ 3	5	.030	-.934	-.331	.464	.260	.281	.465	.969	.434

Note: ML = maximum likelihood; MAP = maximum a posteriori; C-MAP = corrected maximum a posteriori.

Let us consider first the ranges of extremely small and extremely large proficiency levels, for example, smaller than -2 and larger than 2 . At these extremes, the C-MAP method yields the smallest bias, except when the proficiency levels are smaller than -3 or larger than 3 . However, the relatively small number of examinees (eight with proficiency smaller than -3 and five with proficiency larger than 3) does not allow any reliable conclusion regarding the accuracy of the estimation methods, although they may be looked at more carefully for other reasons (such as fairness issues, for instance). Moreover, the bias of the ML method is somewhat close to that of the C-MAP technique and the MAP method is the most biased estimator at the extremes of the proficiency scale, as expected. On the other hand, the lowest standard errors are provided by the MAP method whereas the ML approach yields the largest standard errors. This confirms the results from the previous simulations. The standard errors of C-MAP lie between those of the MAP and the ML approaches and are usually closer to the values of the MAP estimator. Eventually, the conclusions regarding the RMSE values are identical to those for the bias: They are smallest for the C-MAP method and largest for the MAP estimator, whereas the ML approach stands in between.

When the ranges of proficiency levels tend to get closer to the mean value of zero, the ML method provides the smallest bias, but the C-MAP method remains less biased than the MAP method. However, the increase in standard errors for C-MAP is such that the values are now comparable to the standard errors of the ML estimator. The MAP alternative therefore remains the least variable approach for proficiency levels around the mean prior values, as expected. Globally, all three methods provide similar RMSE values. This confirms that near average proficiency levels, the gain in bias reduction is balanced by an increase in standard error and vice versa.

It is easy to notice the similarity of these conclusions to those of the previous study. The C-MAP method is most useful for reducing the bias at the extremes of the proficiency scale, without a large increase in its standard error. For average proficiency levels, however, C-MAP appears to behave like the other methods, and the gain in bias reduction is not sufficient to consider the C-MAP technique as the best candidate overall.

Discussion

The C-MAP estimator was built on the idea of reducing the bias of the MAP estimator of the proficiency level. Simulations confirm that this correction performs well. For extreme proficiency levels, the bias reduction is much greater than the resulting increase in standard error, which implies that C-MAP is a more efficient estimator of proficiency level in that case. Moreover, the simulations showed that C-MAP and ML perform quite similarly in terms of both bias and standard error. This is actually the case when only finite ML estimates of the proficiency level are taken into account. Restricting the range of θ to limit the range of the proficiency level estimates (which is commonly done in practice) would obviously modify this assessment. It is also important to remember that C-MAP always provides finite estimates of proficiency level.

Finally, the two simulation designs highlighted the same trend in the behavior of the C-MAP method with respect to the ML and MAP techniques. This provides evidence that the bias correction of the C-MAP method tends to effectively reduce the bias, especially for extreme proficiency levels, both when the prior standard normal distribution is realistic (second study) or unrealistic (first study) with respect to the simulated data.

Final Comments

The purpose of this study was to suggest a simple test-length correction of the MAP estimator of the proficiency level. The correction is made up of a multiplicative factor, which depends only on the number of items and was calibrated through a simulation study. The correction aims at decreasing (and, in fact, almost cancels) the bias of the MAP estimator, but it implies a corresponding inflating effect on the standard error of this estimator. For small tests, the correction in bias is rather large, whereas for larger tests both the MAP and C-MAP methods provide similar trends in estimating proficiency levels.

Because the C-MAP method provokes a large increase in the standard error for very short tests, one alternative method would be to modify the prior distribution of the proficiency levels, for instance, by increasing the prior variance. However, such an approach was not investigated here because the C-MAP estimator is based on an ad hoc test-length adjustment of the MAP method. It might nevertheless be useful to compare both approaches in a further comparative study.

It was established that the C-MAP method is most useful at extremely low and extremely high proficiency levels, where the bias reduction is much greater than the increased variability because of the multiplicative correction factor. In particular, response patterns with only-correct or only-incorrect responses, and which are typically produced by examinees with, respectively, very high and very low proficiency levels, can be estimated with the C-MAP method but not with the ML method. The simulations ultimately highlighted that ML and C-MAP perform similarly. Note, however, that the particular study of extreme proficiency levels could be carried out with a specific test that is designed for extreme proficiency values. Such a test might be more efficient than the C-MAP method, which aims at reducing the overall estimation bias on the whole proficiency scale. Another solution might be to apply the test-length correction only to extreme ability estimates, without affecting the average estimates of ability. This might improve the estimation at the extremes of the ability scale, without distorting the estimates around the average ability level. Further research, however, is necessary to develop such a method.

The correction factor introduced in the C-MAP estimator was calibrated through the intensive simulation study described in the section "A Bias-Corrected MAP Estimator." One fundamental hypothesis is that the standard normal distribution is considered as a prior density for the proficiency levels. If this is not the case, then the bias correction could be inappropriate. However,

a complete analysis of this issue does not seem feasible, because each particular prior distribution needs to be investigated. The present C-MAP estimator can therefore be seen as a first proposal in correcting the bias of the MAP method, and it performs very well provided that the prior standard normal density is under consideration. Some ongoing research will hopefully address this assumption by seeking to discover the best correction factor when other normal (but nonstandard) distributions are used as prior densities. Another issue of interest is to compare the efficiency of the C-MAP estimator when the average difficulty level is different from the average ability level.

To calibrate the bias correction in the section “A Bias-Corrected MAP Estimator,” the item difficulties were drawn from a uniform distribution in order to include easy, average, and difficult items. It would also be of interest to further investigate whether the C-MAP estimator still performs well when the test contains a different set of item characteristics, for instance, if the test is designed with difficult items only to target a high-ability sample of examinees. Nevertheless, the simulated example in the subsection “Normally Distributed Parameters,” where the items are drawn from a standard normal distribution, is more in line with conventional test designs and the current results showed that C-MAP was still efficient in reducing the bias at the extremes of the proficiency scale. Thus, there is some evidence that the C-MAP method performs well with almost any type of test design, although a careful investigation of this hypothesis needs to be performed.

Another issue of interest is to determine whether the C-MAP estimator would still be efficient in correcting the bias of the MAP method when other logistic models are considered. The simulations described in the section “Assessing the Accuracy of C-MAP” have been partly extended to the 2PL and the 3PL models. The primary results tend to indicate that the C-MAP estimator is still useful in reducing the bias of the extreme proficiency levels, especially with the 2PL model. However, because the C-MAP form of Equation 10 was derived under the Rasch model, a better approach would be to reproduce the study of the section “A Bias-Corrected MAP Estimator” with these extended logistic models. It is very likely that another correction factor would be suitable, though it seems that the advantages of the current C-MAP method are still present when the considered model is no longer the Rasch model.

Because the bias correction of the C-MAP estimator was derived from the whole range of proficiency levels, its main goal is to reduce the estimation bias on the whole range, rather than for a specific pattern. However, it is expected that the C-MAP method would be useful in the fields of psychometrics and educational measurement. A possible application could be considered in the framework of computerized adaptive testing, where proficiency level estimates are iteratively estimated and used for the selection of the next item to be administered. At the early stages of an adaptive test, highly biased estimators can result from small test-length response patterns with only-correct or only-incorrect responses. This could lead to the selection of inappropriate items and thus influence the remaining stages of the test (e.g., Chang & Yin, 2008, or Rulison & Loken, 2009). The use of the C-MAP method at the early stages of adaptive testing might improve the first-item selections and enhance the adaptive process. Other situations, such as application of short screening tests or subtests, might provide other fields of application for the C-MAP method.

In addition, and related to one aforementioned issue, the efficiency of the MAP and C-MAP methods could be compared when the distribution of the proficiency levels is actually not standard normal. If the proportion of subjects with extreme proficiency levels is higher than expected with a standard normal distribution of proficiencies, then the MAP estimator might fail to give accurate estimates for those subjects. However, the test-length correction of the C-MAP estimator could reduce the impact of prior misspecification and, thus, provide more reliable estimates of levels at the extremes of the proficiency scale.

Several extensions of the present study are currently under consideration. First, as mentioned earlier, the C-MAP method was calibrated by means of simulations involving the Rasch model only. It would be interesting to determine whether the 2PL or 3PL models would imply identical adjustments of the correction factor. A similar approach with the posterior mean estimator (or expected a posteriori; EAP) of the proficiency level (Bock & Mislevy, 1982), instead of the MAP estimator, is also under consideration. This estimation method is sometimes preferred to MAP, especially in the computerized adaptive testing framework. Usually, both the EAP and MAP methods yield close estimates of proficiency level, and it would be interesting to look for a potential bias correction of EAP and therefore a C-EAP alternative for estimating extreme proficiency levels.

Acknowledgments

The authors thank the editor and three anonymous referees for their insightful comments, which greatly improved the quality of the article. This research was partly funded by the research fund of the K.U. Leuven, Belgium, the research grant "Chargé de Recherches" from the National Funds for Scientific Research, Belgium, and by the Social Science and Humanities Research Council of Canada. Correspondence should be sent to David Magis, Department of Mathematics, University of Liège, Grande Traverse 12, B-4000 Liège, Belgium (david.magis@ulg.ac.be).

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

- Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society (Series B)*, 32, 283-301.
- Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society (Series B)*, 34, 42-54.
- Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.
- 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.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of the EM algorithm. *Psychometrika*, 37, 29-51.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.
- Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. London, England: Chapman & Hall.
- Chang, H. H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, 73, 441-450.
- Cramer, H. (1946). *Mathematical models for statistics*. Princeton, NJ: Princeton University Press.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates [Correction: 82, 667]. *Biometrika*, 80, 27-38.

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman & Hall.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48, 233-245.
- Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23, 157-162.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Magis, D., & Raîche, G. (2010). An iterative maximum a posteriori estimation of proficiency level to detect multiple local likelihood maxima. *Applied Psychological Measurement*, 34, 75-90.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49, 359-381.
- R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17, 1-25.
- Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, 33, 83-101.
- Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, 38, 221-223.
- Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 42, 193-198.
- Samejima, F. (1993). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, 58, 119-138.
- Samejima, F. (1998, April). *Expansion of Warm's weighted likelihood estimator of ability for the three-parameter logistic model to general discrete responses*. Paper presented at the annual meeting of the national Council on Measurement in Education, San Diego, CA.
- Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50, 349-364.
- Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51, 589-601.
- Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Wang, T., Hanson, B. A., & Lau, C.-M. A. (1999). Reducing bias in CT trait estimation: A comparison of approaches. *Applied Psychological Measurement*, 23, 263-278.
- Warm, T. A. (1989). Weighted likelihood estimation of in item response theory. *Psychometrika*, 54, 427-450.