Running head: EVALUATE RASCH ITEM PARAMETER RECOVERY

Evaluate Rasch item parameter recovery in MML and JML estimations by ACER ConQuest Software

Luc T Le

Australian Council for Education Research

and

Raymond J Adams

University of Melbourne & Australian Council for Education Research

Correspondence: Luc Le, ACER, 19 Prospect Hill Road (Private Bag 55), Camberwell VIC 3124, Australia. Email: le@acer.edu.au

Abstract

This study used Monte Carlo simulations to evaluate the item parameter recovery from ACER ConQuest 3 software (Adams, Wu, & Wilson, 2012) for the dichotomous Rasch model. Our primary focus was the comparison of its estimation methods, joint maximum likelihood (JML), marginal maximum likelihood (MML) with a normal distribution assumption and MML with a discrete distributions assumption when the populations were in fact non-normal. The simulation data sets were generated with two test lengths (10 and 50 items) and four alternative *true* population distributions for the abilities: normal, bimodal, uniform, and chi-square. As expected, results showed that MML-Normal was the best method when the assumption of ability distribution was matched, regardless the test length. However, the accuracy or MML-Normal decreased with the violation level of the assumption of normal distribution of the latent ability. The MML-Discrete estimation could overcome well the weakness of the MML-Normal when the normality of the ability distribution was violated. The estimates of the corresponding standard errors produced by ACER ConQuest 3 were also being examined and discussed.

Introduction

This paper is concerned with comparing the outcomes of using joint maximum likelihood estimation (JML) and marginal maximum likelihood estimation (MML) as estimation methods for Rasch measurement models (Rasch, 1960/1980). In this particular paper we will be limiting ourselves to an examination of the properties of JML and MML for Rasch's simple logistic model. Our particular interest is in comparing JML and MML when the assumptions required by MML are violated.

We begin by introducing the simple logistic model. Suppose that a sample of N examinees indexed $n = 1, \dots, N$ responds to a set of K test items indexed $i = 1, \dots, K$; and the items are scored dichotomously so that the response of student n to item i can be denoted x_{ni} which takes the value '1' for a correct response and '0' for an incorrect response, then the model can be written as:

$$Pr(x_{ni}; \delta_i, \theta_n) = \frac{exp[x_{ni}(\theta_n - \delta_i)]}{1 + exp(\theta_n - \delta_i)}, \qquad (1)$$

where θ_n is referred to as the case parameter, it is the location of case n on the latent continuum and δ_i is referred to as the item parameter, it is the location of item i on the latent continuum.

JML and MML are among the most popular estimation methods available for item response models. The JML method, as developed by Birnbaum (1968) and Wright and Pachapakesan (1969), and has been widely used (Hambleton & Swaminathan, 1985; Baker, 1992). Under the JML method all item parameters and all person parameters are regarded as fixed unknowns to be estimated. Therefore, the parameters involved in the estimation procedure of this method are all of the case parameters, the θ_n for $\mathrm{n} = 1, \dots, \mathrm{N}$ and all of the item parameters, the δ_i for $\mathrm{i} = 1, \dots, \mathrm{K}$.

JML requires maximisation of the likelihood:

$$\Lambda(\Theta, \Delta; X) = \prod_{n=1}^{N} \prod_{i=1}^{K} Pr(x_{ni}; \delta_i, \theta_n), \qquad (2)$$

with respect to Θ and Δ. Θ represents all the case parameters, Δ all the item parameters and ***X*** the data.

The MML method was developed by Bock and Lieberman (1970), and Bock and Aitken (1981). When using the MML method it is assumed that individual's positions on the latent variable are sampled from distributions of possible values. In the simplest applications of MML the model then becomes:

$$Pr(x_{ni}; \delta_i | \theta_n) = \frac{exp[x_{ni}(\theta_n - \delta_i)]}{1 + exp(\theta_n - \delta_i)}, \qquad (3)$$

where $\theta_n \sim g(\alpha)$, that is the case locations are distributed independently according to the probability density function g which has parameters α.

Rather than estimating the location of each case, the parameters α, of the distribution, g, from which the cases are sampled, are estimated. Under this method, item parameters are considered as "structural", while ability parameters are "incidental". As a result, in its estimation procedure, the MML includes the item parameters, the δ_i for $\mathrm{i} = 1, \ldots, \mathrm{K}$ and the population parameters α but not case parameters.

MML involves the maximisation of the likelihood:

$$\Lambda(\alpha, \Delta; X) = \prod_{n=1}^{N} \left[\int_{-\infty}^{\infty} \prod_{i=1}^{K} Pr(x_{ni}; \delta_i, \theta_n) d\, \theta_n \right], \qquad (4)$$

with respect to α and Δ. α represents all the case distribution (or population) parameters, Δ all the item parameters and ***X*** the data.

Practically, JML is relatively easy to implement and has been applied in many widely used computer programs. These include CALFIT (Wright & Mead, 1975), BICAL (Wright, Mead, & Bell, 1979), CREDIT (Masters, Wright, & Ludlow, 1980), FACETS (Linacre, 1989), Quest (Adams & Khoo, 1993) and Winsteps (Linacre 2007) to name a few. These implementations of JML for Rasch Models have been accompanied by a wide array of simulation studies (Wright & Douglas, 1977a, 1977b; Wright, Mead, & Bell, 1979; Masters, 1980) that have produced impressive results.

From a theoretical perspective however JML has some shortcomings. Proofs of the asymptotic properties of maximum likelihood estimators (Cramèr, 1946; Wald, 1949) assume that the number of parameters to be estimated is fixed and finite and does not changes as more independent observations are made. For the Rasch Model, however, the number of parameters to be estimated increases as the length and/or sample size increases. Neyman and Scott (1948) showed that when the number of parameters increases with the observations it is possible for maximum likelihood estimates to lack the usual properties of consistency, efficiency and asymptotic normality. Andersen (1973) showed that JML estimates of the item parameters for Rasch models are not consistent if the number of items is fixed and the size $N \rightarrow \infty$.

Examining the properties of JML in more detail Haberman (1977) showed that the JML estimates of the simple Rasch model are consistent when $N \rightarrow \infty$, $K \rightarrow \infty$ and $\log K/N \rightarrow 0$, and asymptotically multivariate normal when $N \rightarrow \infty$, $K \rightarrow \infty$ and $(\log K)^2/N \rightarrow 0$. Haberman's results were derived for the simple logistic model and we are not aware of extensions of the results to JML estimates of more general Rasch models. The key requirement in Haberman's proof is that the probability of inestimable parameters approach zero. Inestimable case parameters result when a case obtains a perfect or zero score and inestimable item parameters occur when a response category is not used – for the dichotomous model this reduces to the same requirements as for case parameter estimates. Therefore, Haberman's proof would suggest that the parameter estimates for more general models would be consistent provided the probability of unused categories and perfect and zero case scores approaches zero.

To deal with the bias in JML, Wright and Douglas (1978) proposed a correction of, (K-1)/K, where K is the number of items. They argued that this correction removed most of the bias for K>20 and this finding was supported by Wright (1988). For tests of fewer than 10 to 15 items, van den Wollenberg, Wierda, and Jansen (1998) suggest that this bias correction is inappropriate since the bias is dependent not only on the number of items, but also on the skewness of the item difficulty distribution. This correction has commonly been applied in JML software.

A second potential shortcoming of JML is that in many of its potential applications the goal is to make inferences concerning populations. For example the interest might be in the variance of a latent variable in a specific population, or the correlation between two latent variables in a specific population. In such contexts, if JML is used for estimating the measurement model then a two-step analysis is required. First the case parameters are all estimated with JML and then the population parameters are estimated from individual case estimates. A number of researchers (Adams, 1989; Adams, Wilson, & Wu, 1997; Mislevy, 1984) have illustrated that the use of case parameter estimates as though they were true values in a two-step analysis can lead to quite misleading outcomes. This problem is at its most serious when there are few items in a measurement. In general, as mentioned by Mislevy (1984) “The distribution of estimates of individual subjects’ parameters may then depart radically from the distribution of the parameters themselves, thereby invalidating any analyses that would treat the estimates as if they were the parameters they represent” (p. 359).

The MML method overcomes these disadvantages of the JML method, but it does so at the expense of making an additional assumption concerning the distribution for the latent variable. Although the distribution can be of any type, with a limit on the number of parameters, normal densities are most frequently used (see Bock & Lieberman, 1970; Bock & Aitkin, 1981; Thissen, 1982; Mislevy, 1984; Adams & Wu, 2007).

If MML is used, population parameters are estimated directly from the observed responses; that is without estimating a location parameter for each case. This avoids the problems associated with estimating population characteristics using fallible case parameter estimates in a two-step process. Secondly, if both the item response models and the assumed population distributions are correct the MML item parameter estimates are consistent for any fixed *K* (Bock & Aitkin, 1981; Harwell, Baker, & Zwarts, 1988).

From a theoretical perspective it should be noted that in assuming a distribution for the latent variable, MML is not just an alternative method of estimation – it fits a different model. Following the convention of all relationships between fixed quantities functional and relationships between random quantities structural (Kendall & Stuart, 1979), de Leeuw & Verhelst (1986) have called the model a structural Rasch Models if it is assumes that the cases are some from some distribution and a functional Rasch model if no distributional assumptions are made. The structural model that is fitted whenever MML is applied is a model with more assumptions than the functional model assumed when estimating with JML. The advantage of this is that, should the distributional assumptions be correct then the MML item parameter estimates will be consistent and will have a smaller mean squared error than their JML counterparts. The disadvantage is that when the distributional assumptions are not correct the parameter estimates may not be consistent and may have less desirable characteristics than JML estimates.

Engelen (1987), using simulated data to compare joint, marginal, and conditional maximum likelihood methods, as well as Bayesian methods, minimum chi-square methods, and paired comparison estimation, confirmed that MML was the best procedure when its assumptions where met. However, the application of the marginal estimation approach is often restricted to the assumption of a normal distribution for the population when this may not be a desirable assumption. Some empirical studies demonstrate that MML estimators loose accuracy and efficiency when the prior assumption of the latent distribution is violated. Specifically, factors showing effects on the accuracy and estimation error for parameter estimates could be the degree of skewness and kurtosis of the true underlying examinee parameter distribution, the match of the prior distribution to this underlying distribution, the variance of the prior distribution, sample size, test length and the number of parameters whose true values are extreme (Yen, 1987; Drasgow, 1989; Zwinderman & van den Wollenberg, 1990; Seong, 1990; Harwell & Janosky, 1991; Stone, 1992; Kirisci, Hsu, & Yu, 2001).

While MML is a most commonly used with an assumption of normality for the latent variable, this need not be the case. For example, Adams and Wilson (1997) discuss the use of a discrete distribution where a fixed set of grid points is assumed and a weight is estimated at each grid point.

This study is primarily concerned with the question of the accuracy in item parameter recovery by the MML method, when compared to that of the JML method when the distributional assumptions of MML are violated. We also examine the accuracy of estimation of the population variance. We consider four distributions: normal, chi-square with five degrees of freedom, a bimodal mixture of two normal distributions and a uniform distribution. We use samples of size 2000, two test lengths (10 items and 50 items) and we estimate the models using JML and MML with both a normal population assumption and with a discrete distribution assumption.

Method

Data generation

This study is concerned with item parameter recovery for the dichotomous Rasch model. Our primary focus is on comparing JML and MML when the assumptions of MML are violated, that is the abilities are not sampled from the distribution that is assumed in the estimation. We therefore generate data that conforms to the dichotomous Rasch model using four alternative *true* population distributions for the abilities. We then use the ACER ConQuest 3 software (Adams et *al*., 2012) to recover Rasch model parameter estimates using JML and MML. For the MML estimation we consider two alternative distribution assumptions. First, we assume a normal population distribution, the variance of which is estimated, this will be referred to as MML-Normal. Second, we assume a discrete population distribution, under which a set of 15 nodes uniformly spaced between –6.0 and 6.0 is assumed and densities at each node are estimated, this will be referred to as MML-Discrete.

For the simulation study a number of factors that can be varied need to be considered. The characteristics of the population distribution, the size of the ability sample, the characteristics of the item distribution and the length of the tests. For the sake of simplicity and to ensure focus on the shape of the population distribution, eight distinct combinations of the above listed factors were considered – four population distributions (to be described below), a single sample size of 2000 examinees, a single uniform U[–3,3] item distribution and two test lengths (10 and 50 items). The item difficulties of 10 and 50 items were randomly generated from a uniform distribution U[–3,3] and then transformed to ensure constrained as a mean of zero. These values were fixed and considered as the generated values for all replications. For each of the eight combinations of factors 1000 replications was undertaken.

The central variable in this investigation was the shape of the population distribution. The four distributions used in this study are shown in Figure 1: normal, bimodal, uniform and chi-square. For comparison purposes all four distributions had a mean of zero and standard deviation of one. The normal distribution was *N*(0,1). The uniform distribution was $U[-\sqrt{3},+\sqrt{3}]$. The bimodal distribution was a combination of two normal distributions with means of –0.8 and 0.8 respectively, and standard deviation of $\sqrt{0.6}$, $N\left(-0.8,\sqrt{0.6}\right)$and $N\left(+0.8,\sqrt{0.6}\right)$. The chi-square distribution was a standardisation of a chi-square distribution with five degree of freedom. This distribution was positively skewed (skewness of 1.26), and the other three were symmetric (skewness of zero).

More specifically, as can be seen from Figure, relative to the normal, the uniform distribution (Kurtosis= –1.20) has light tails, a flat centre, and heavy shoulders; the bimodal distribution (Kurtosis=3.79) has two peaks, light tails, a deep centre, and heavy shoulders; the chi-square distribution (Kurtosis=2.40) has a heavy right tail, a peaked centre, and light shoulders.

Insert FIGURE 1 about here

For each randomly drawn sample a set of simulated dichotomous data were generated using the fixed item difficulties. The data were generated using the ACER ConQuest 3 *generate* command so that they conformed to Rasch's simple logistic model.

Analysis

Item calibrations based on MML-Normal, MML-Discrete and JML methods were also implemented by ACER ConQuest 3 using the following *estimate* command.

“estimate! iterations=1000, converge=0.00001, fit=no, stderr=quick, *method=gauss*”

In this estimate command, the convergence criterion was set as 0.00001, the maximum number of iterations was 1000, and MML with a normal distribution was used as it is the default method of estimation. The fit=no option was used so that estimation time was reduced. Further, as the model was identified by setting the mean of the latent distribution at zero the item parameter estimates are independent (Adams, 1989) so that the stderr=quick option was expected to provide appropriate estimates of the standard errors.

For MML-Discrete the estimate command above with an option, *distribution=discrete*.

“estimate! iterations=1000, converge=0.00001, fit=no, stderr=quick, *distribution=discrete*”

In this estimation the default number of nodes (15) and the default node range (–6.0 to 6.0) was used. However, in the cases of study here only some of these nodes would be expected to have a non-negligible density. It can be seen from Figure 1 that the uniform distribution is covered by only five of the nodes (–1.714 to 1.714). The chi-square distribution is covered by only nine of the nodes (–0.857 to 6.000). Among those, three nodes (4.286, 5.143, 6.000) would rarely be used with the chi-square distribution. Similarly, six nodes (–6.000, –5.143, –4.286, 4.286, 5.143, 6.000) would rarely be used with the normal or the bimodal distributions. The normal distribution is likely to use the most number of nodes while the uniform distribution would use the least number of nodes in the estimation procedure.

The estimate command above with an option, *method=jml*, was used for item calibrations based on JML method:

“estimate! iterations=1000, converge=0.00001, fit=no, stderr=quick, *method=jml*”

In the ACER ConQuest 3 implementation of JML, the correction factor (*K*-1)/*K* is applied. The nature and number of estimated parameters differs amongst MML-Normal, MML-Discrete and JML. While for each method either 10 or 50 item parameters are estimated, the situation is quite different for the case or population parameters. For MML-Normal there is one estimated distribution (or person) parameter, the variance. For MML-Discrete, there are 15 estimated distribution (or person) parameters, the densities at each of the 15 nodes points. For JML, there are 1999 estimated person parameters, the location of each case on the latent dimension, but with a degree of freedom lost due to the identification constraint.

Since our primary focus is on the effect of violating the population distribution assumption on item parameter estimation and because it is only item parameters that are common to both estimation methods, we focus primarily on the parameter recovery for the item parameters. We also consider estimation of the population variance.

The accuracy of parameter recovery is shown by computing bias and root mean square error (RMSE) statistics for each of the estimated parameters. Bias for an item difficulty parameter or the variance parameter was computed as the mean difference, across the replications, between the estimated values and the true values.

$$Bias(\delta_i) = \left(\sum_{k=1}^{1000} \hat{\delta}_i^k \right) / 1000 - \delta_i \, , \qquad (5)$$

where δ_i denoted the generating difficulty value of item *i*, and $\hat{\delta}_i^k$ denoted its estimate in the *k*-th replication. An analogous approach was used for the variance parameter for MML-discrete.

RMSE was the square root of the average squared difference between the true and estimated values:

$$RMSE(\delta_i) = \sqrt{\sum_{k=1}^{1000} (\hat{\delta}_i^k - \delta_i)^2 / 1000} \quad (6)$$

.

Additionally, together with assessing the accuracy of parameter recovery of item difficulty parameter estimates obtained by the three estimation methods, the corresponding standard error (SE) of these estimates was also evaluated by the ratios of average error variance over sampling variance.

$$Ratio(SE_i) = \frac{\sum_{k=1}^{1000} \left(SE_i^{\,k}\right)^2 / 1000}{\sum_{k=1}^{1000} (\hat{\delta}_i^k - \bar{\delta}_i)^2 / 1000} \quad , \quad (7)$$

where $\bar{\delta}_i$ denotes the average of $\hat{\delta}_i^k$ estimates of difficulty value of item *i*, and SE_i^k denotes the standard error of the estimate in the *k*-th replication. If the standard error estimate, SE (produced by ACER ConQuest 3), was accurate, the ratio of the average error variance estimate over the sampling variance (equation 6) would approach unity. Otherwise, if the ratio was larger than unity, the standard error was overestimated. On the other hand, if the ratio was smaller than unity, the standard error was underestimated.

Results

Bias of item difficulty estimate

Table 1 gives the mean and standard deviation of the absolute value of item difficulty bias (across the items in each of the two tests) from the MML-Normal, MML-Discrete and JML estimators over the 1000 replications, for each of the four ability distributions. Additionally, Figure 2 shows the magnitude of the bias for individual items by each estimation method plotted against the generating item difficulty.

Insert TABLE 1 about here

Table 1 shows that, with the 10-item test, the bias was negligible for MML-Normal. In this case, the mean of the absolute bias value was only 0.003. The value increased to 0.006, which was still small when the distribution was bimodal and to 0.010 and 0.034 when the distribution was uniform and chi-square respectively. The bias in the MML-Normal estimators for the 50-item test is less than that for the 10-item test for all three non-normal distributions. However, the bias was negligible for three of the ability distributions: normal, bimodal and uniform, where the mean of the absolute bias value was only 0.002—0.003. The value was 0.009 when the ability distribution was chi-square.

Part (a) of Figure 2 demonstrates that when the abilities are normally distributed, MML-Normal has an almost zero bias for all generating values. For the bimodal and uniform distributions there was evidence of a linear bias resulting in underestimation of the difficult of easy items and over estimation in the difficulty of harder items, while for the chi-square the shape of the bias as a function of item difficulty is arc downwards. In the chi-square case there was underestimation of the difficult of both very easy and very hard items and there was over estimation in the difficulty of middle difficult items. These bias patterns are more evident for the longer test, part (b) of Figure 2, although the actual magnitude of the bias is less for the longer test than it was for the shorter test.

For JML the bias was larger than that for MML-Normal when the ability distributions was normal, smaller when the distribution was chi-square and similar for the bimodal and uniform distributions.

While the bias for the JML estimation for the long test was negligible for all distributions (the mean of the absolute bias value was only 0.003—0.005) Part (d) of Figure 2 demonstrates that there was general trend of underestimation of the difficult of easy items and over estimation in the difficulty items.

The MML-Discrete method produced estimates superior to MML-Normal for the three non-normal ability distributions and superior to JML for all ability distributions. In the 10-item test, the bias in the item difficulty parameter estimates from this method was very consistently small. The mean of the absolute bias was only 0.002 to 0.004. The mean of the absolute bias was 0.003 to 0.008 in the 50-item test. Part (e) and part (f) of Figure 2 indicate that the accuracy of MML-Discrete estimator was superior to the MML-Normal estimator when the ability distribution was chi-square. The MML-Discrete estimator did however have larger bias for uniform distributions than for the other three distributions. In that case, the difficult of easy items tended to be under-estimated while the difficult of harder items tended to be over-estimated. This probably happened due to the fact that in the computation procedure, only the middle five of the 15 quadrature nodes were utilised for the uniform distribution while more of the quadrature nodes were utilised for the other three distributions.

Insert FIGURE 2 about here

An explanation for the shape of the bias function for MML-Normal can be found by reviewing the cumulative density functions (CDF) for each of the distributions The CDFs, which are plotted in Figure 3 show that when using a normal distribution to approximate the chi-square distribution, there would be a substantially greater proportion of examinees answering the items correctly than expected for easy items (<–1.1 logits, for example) or hard items (>1.5 logits, for example). Therefore, the difficulty of these items would be underestimated.

Furthermore, there would be a substantially lower proportion of examinee correctly answering the items in middle range of difficulty (–0.8 to 0.8 logits) than expected. As a consequence, there was an over-estimation of the difficulty for these items. Similarly, Figure 3 suggests that using a normal distribution to approximate the uniform distribution, would result in underestimation for very easy items and over estimation for very hard items Finally, the CDF shape of the bimodal is closer to the CDF shape of the normal distribution. This could explain why the bias from the MML-Normal estimation in this distribution was smaller than that in the chi-square and the uniform distributions.

Insert FIGURE 3 about here

RMSE of item difficulty estimate

Table 2 gives the mean and standard deviation of the root mean square error, RMSE, of item difficulty estimates from the three estimation methods, for each of the two test lengths and for each of the four ability distributions. Additionally, the magnitude of the RMSE for individual items by each estimation methods is plotted against the generating item difficulty in Figure 4.

Firstly, according to Table 2, the mean value of RMSE from the MML-Normal estimator was the largest in the short test when the ability distribution was chi-square (0.071), and it was consistently smaller in other symmetric distributions (0.061—0.062). The mean value of RMSE from the MML-Normal estimator with the chi-square distribution was reduced to 0.065 in the long test, but it was not with other three symmetric distributions (0.061, 0.061 and 0.062 compared to 0.063, 0.063 and 0.062, respectively).

Secondly, the mean value of RMSE from the JML estimator in the short test was largest when the ability distribution was normal (0.074) and second largest when the ability distribution was chi-square (0.064). The RMSE mean value decreased in the long test to 0.063 and 0.063 respectively. The mean RMSE in bimodal and uniform distributions increased very slightly from the short test (0.061 and 0.061) to the long test (0.063 and 0.062, respectively).

Insert TABLE 2 about here

Furthermore, the mean value of the RMSE from the MML-Discrete method in the short test was consistently small and similar for all four ability distributions (0.060—0.061). The value increased slightly in the long test to 0.062—0.063. The small increase of the RMSE mean (from the short test to the long test here (and in some cases above in MML-Normal and JML) could be due to the fact that the actual standard deviation of generated item difficulty in the short test (SD=1.786) was smaller than that in the long test (SD=1.803).

Additionally, regarding the RMSE for individual items, Figure 4 shows that in general, the more the generating item difficulty differed from zero (middle difficulty) the larger the RMSE was, regardless of the estimation methods. However, in the case of the MML-

Normal and with the chi-square distribution, the shape of the RMSE pattern tended to be slight arc downwards at a middle interval of the ability distribution.

Insert FIGURE 4 about here

Standard error estimates

Table 3 provides a comparison of the between replication variation in the parameter estimates and the estimates of the standard errors. The ratios are plotted against the item parameters in Figure 5. The outcomes from MML-Discrete are not provided because the ACER ConQuest 3 implementation of MML-Discrete estimation provided clearly inappropriate estimates of the standard errors.

Insert TABLE 3 about here

The table shows that this ratio was closer to one for the MML-Normal than for the JML estimators, in every case. This suggests that the standard error estimates from MML-Normal, as produced by ACER ConQuest 3, were more appropriate than those estimated for JML. For both estimation methods, the standard errors estimated for the long test were more accurate than those estimated for the short test. There was no clear difference in the ratio value for the different ability distributions. The standard errors from JML were slightly overestimated in the short test with the three non-normal distributions. However, this did not happen with the JML for the long test.

Additionally, no clear systematic patterns were found in the plots of Figure 5, meaning that the ratios were independent of the item parameters. The standard errors from none of the combinations showed substantial under- or over-estimation.

Insert FIGURE 5 about here

Bias of ability variance estimate

As discussed above, in addition to item parameter estimates, the MML-Normal estimation yields estimates of the population variance while JML provides individual location estimates for every student. Under JML, estimates of population characteristics, such as the variance, can only be obtained via two-step procedures. The first step is the estimation of the person parameters and a second step is an estimation of the variance from those estimated person parameters.

In this section we compare the MML-Normal estimates of the variance with their generating values and similarly we compare the two-step estimates of the variance from JML and MML-Discrete with the generating values. Note that for the person parameter estimates under JML and MML-Discrete we used weighted likelihood estimates (WLE; Warm, 1982), since they are well known to be less biased than their unweighted counterparts (Roberts & Adams, 1997). The two-step estimates from MML-Normal were also compared to the generating values.

Insert TABLE 4 about here

Table 4 provides a comparison of the variance estimates obtained from each of the estimation methods with the generating values. The table also includes the RMSE values. The bias values from each case are plotted in Figure 6 and the RMSE is plotted in Figure 7.

Insert FIGURE 6 about here

Insert FIGURE 7 about here

It can be seen from Table 4 and Figures 6-7 that when the ability distribution was normal (a match with the assumption of the model estimation), the bias in the variance estimates for MML-Normal was very small (≤0.003) regardless of test length. The bias increased, however, when there was a violation in the normality assumption. When the ability distribution was bimodal or uniform, the estimate of the sample variance was overestimated. When the ability distribution was chi-square, the MML-Normal estimate of the sample variance was underestimated.

For all eight combinations in the study, the bias of the sample variance estimate obtained through individual person parameter WLE (two-step estimate) in JML, MML-Discrete and MML-Normal was similar. In each case the estimate of the sample variance was clearly an overestimate and the estimation bias reduced with increased test length. Moreover, the bias magnitudes for those methods were larger than the corresponding bias magnitudes from the MML-Normal direct estimation.

Conclusion and discussion

Several conclusions can be drawn from the simulations in this study. First, with a 50-item test, the three methods tended to produce similar results with small or negligible bias in item parameter estimates, although MML-Normal provided more accurate estimates than JML and MML-Discrete when the assumption of ability distribution was matched.

Second, while the accuracy of JML was dependent on test length this was not the case for MML-Normal. MML-Normal provided very reliable estimates in a 10-item test when the assumption of ability distribution was matched. However, the accuracy or MML-Normal was decreased when there was a violation of the assumption of a normal distribution of the latent ability. This method appeared to produce the largest bias when the ability distribution was skewed.

Third, MML-Discrete overcame the weaknesses of the MML-Normal when the normality of the ability distribution was violated. This method provided less bias than both

MML-Normal and JML, especially in a short test and when normality of the ability distribution was violated. The bias of item difficulty estimates from this method was consistently small. However, the accuracy of MML-Discrete estimator is probably dependent on the choice of nodes. In the case of this study, the MML-Discrete estimator had larger bias for uniform distributions than for the other three distributions and this corresponds to the case where there is the largest number of redundant nodes.

Regarding RMSE, when the sample size was large, increasing test length did not always help to reduce the mean value of RMSE in item difficulty recovery. Moreover, as expected, the more the generating item difficulty differed from zero (middle difficulty) the larger the RMSE was, regardless of the estimation methods.

Additionally, the MML-Normal and JML estimators from ACER ConQuest 3 provided good estimates for the standard errors of item difficulties under the Rasch model. The accuracy of the standard errors in both methods was substantially increased by the test length. Moreover, in all combinations examined in this study, the standard error produced by the MML-Normal tended to be more accurate than the standard error produced by the JML.

Finally, as a consequence of the fact that the population variance is directly estimated in the MML-Normal estimation model but not in the JML or the MML-Discrete, the estimation of the variance parameter was far more accurate in MML-Normal than in other two methods even when normality of the ability distribution was violated. When the assumption was matched the bias of MML-Normal estimate of the variance parameter was negligible. The two-step (indirect) estimates of the ability variance from the three methods were similar to each other and well and truly over-estimated.

Findings from this study suggest that to calibrate a short test, the MML-Normal should be in used if the ability distribution is approximately normal. Otherwise, the MML-Discrete should be considered, particularly when the normality assumption of ability distribution is likely to be markedly violated. MML-Discrete works well regardless of the shape of the ability distribution provided the nodes are well chosen to cover the range of the underlying ability distribution.

With a longer test, the three methods tend to produce similar results, although the MML-Normal provides more accurate estimates than the JML and the MML-Discrete when the assumed ability distribution is matched. However, the JML or the MML-Discrete should be recommended ahead of the MML-Normal when the assumption of ability distribution is severely violated (for example, chi-square distributions against normal distributions).

In brief, this study focussed on comparing the accuracy of item parameter recovery for MML and JML estimation methods with different ability distributions. Specifically, the study focussed on the effects of test length and the violation of the normality assumption of the ability distribution on the MML-Normal estimation and compared it to JML and MML-Discrete estimation. Consistent with the findings from a number of previous studies (e.g., Yen, 1987; Drasgow, 1989; Harwell & Janosky, 1991; Stone, 1992; Kirisci, Hsu, & Yu, 2001), it was found that the accuracy of MML-Normal estimators decreased when ability distribution was very skewed. Furthermore, the bias was differentially affected by not only the direction of skewness but also the kurtosis of the distribution. With the chi-square distribution, for example, the bias shape of the MML-Normal estimators tended to arc downwards. There was underestimation of the difficult of both very easy and very hard items, where there was over estimation in the difficulty of some middle difficult items.

Moreover, in this study, for easier comparison purposes, all generated ability samples had the same mean and variance. The study, therefore, did not include the effect of the variance of ability distribution on item parameter recovery of MML and JML. This remains a topic for future research.

Finally, findings from this study also suggest value in a more careful examination of MML-Discrete. The lower bias of MML-Discrete in the case of short test when the normality assumption is violated is quite a promising finding and should motivate further application of this method. One immediate area of valuable further investigation would be the impact of the number of nodes and the node range on the efficacy of the parameter estimation.

References

Adams, R. J. (1989). *Estimating measurement error and its effect on statistical analysis*. Doctoral Dissertation, University of Chicago.

Adams, R. J. and Khoo, S. (1993). *Quest : The Interactive Test Analysis System.* Melbourne, ACER: Australia for Educational Research.

Adams, R. J., and Wilson, M. R. (1996). A random coefficients multinomial logit: A generalized approach to fitting Rasch models. In *Objective Measurement III: Theory into Practice*, G. Engelhard and M. Wilson (eds.), pp 143–166. Norwood, New Jersey: Ablex.

Adams, R. J., Wilson, M. R., and Wang, W. C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement, 21,* 1–24.

Adams, R. J., Wilson, M. R. and Wu, M. L. (1997) Multilevel item response modelling: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficient multinomial logit model: A generalized form of the Rasch model. In M. v. Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57–76): Springer Verlag.

Adams, R.J., Wu, M.L., & Wilson, M.R. (2012) ACER ConQuest 3.0. [computer program]. Melbourne: ACER.

Andersen, E. B. (1973). Conditional inference in multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31–44.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Technique*. New York: Marcel Dekker, Inc.

Birnbaum, A. (1968). Some Latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.*) Statistical Theories of Mental Scores* (pp. 397–472). Reading, MA: Addition-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for dichotomously scored items. *Psychometrika*, *35*, 179–187.

Cramèr, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

De Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch Models. *Journal of Educational Statistics*, *11*, 183–196.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1, 1–38.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*, 77–90.

Engelen, R. J. H. (1987). A Review of Different Estimation Procedures in the Rasch Model. *Research Report 87-6*. Twente University, Enschede, The Netherlands.

Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5,* 815–841.

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

Harwell, M. R., Baker, F.B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and the EM algorithm: A didactic. *Journal of Educational Statistics*, *13*, 243–271.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279–291.

Kendall, M., & Stuart, A. (1979). *The Advanced Theory of Statistics: Volume 2 Inference and Relationship.* (4 ed). London: Charles Griffin.

Kirisci, L., Hsu, T.-C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146–162.

Linacre, J. M. (1989). FACETS: *A computer program for many-faceted Rasch measurement*. Chicago: MESA Psychometric Laboratory. Department of Education, University of Chicago.

Linacre, J. M. (2007). *User's guide and program manual to WINSTEPS: Rasch model computer programs*. Chicago: MESA Press.

Lord, F. M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. *Research Bulletin 75–33*. Princeton, NJ: Educational Testing Service.

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

Masters, G.N. (1980). *A Rasch model for rating scales*. Doctoral Dissertation, University of Chicago.

Masters, G. N., Wright, B. D., & Ludlow, L. (1980). *CREDIT: A computer program for analysing data in ordered response categories*. Chicago: MESA Psychometric Laboratory. Department of Education, University of Chicago.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observation. *Econometrika*, *16*, 1–32.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* (expanded ed). Chicago. The University of Chicago Press. (Original work published 1960).

Roberts, D., & Adams, R.J. (1997). Bias in weighted likelihood estimation when using the Rasch model. M. Wilson, G. Engelhard & K. Draney (Eds.), *Objective measurement IV Theory into practice*. Norwood, (pp 279–296) NJ: Ablex.

Swaminathan, H. and Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In Weiss (Ed.), *New Horizons in Testing*. New York: Academic Press.

Seong, T. (1990). Sensitivity of Marginal Maximum Likelihood Estimation of Item and Ability Parameters to the Characteristics of the Prior Ability Distributions. *Applied Psychological Measurement* , *14*, 299–311.

Thissen, D. (1982). Marginal Maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1–16.

van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1998). Consistency of Rasch Model Parameter Estimation: A Simulation Study. *Applied Psychological Measurement*, *12*, 307–313.

Wald, A. (1949). A note on the consistency of the maximum likelihood estimates. *Annals of Mathematical Statistics, 20*, 595–601.

Wright, B. D., & Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23–48.

Wright, B. D., & Douglas, G. A. (1977a). Conditional versus unconditional procedures for sample free item analysis. *Educational and Psychological Measurement, 37*, 47–60.

Wright, B. D., & Douglas, G. A. (1977b). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*, 281–295.

Wright, B. D., & Douglas, G. A. (1978). Better procedures for sample-free item analysis. *MESA Memorandum No. 20.* Chicago, IL: University of Chicago, Department of Education.

Wright, B. D., & Mead, R. J. (1975). *CALFIT: A computer program for analysing data with the Rasch Model. Chicago: MESA Psychometric Laboratory*. Department of Education, University of Chicago.

Wright, B. D., Mead, R. J. & Bell, S (1979). BICAL:Calibrating items and scales with the Rasch Model. *Research Memorandum 23*.Chicago: MESA Psychometric Laboratory. Department of Education, University of Chicago.

Wright, B. D. (1988). The Efficacy of Unconditional Maximum Likelihood Bias Correction: Comment on Jansen, van Wollenberg, and Wierda, *Applied Psychological Measurement*, *12*, 315–323.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*, 275–291.

Zwinderman, A. H., & Van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, *14,* 1, 73–81.

Table 1

Statistical summary of absolute bias in item estimates

Ability distribution	Minimum	Maximum	Mean	SD
MML-Normal and the short test				
Normal	0.000	0.006	0.003	0.002
Bimodal	0.002	0.020	0.007	0.006
Uniform	0.001	0.029	0.010	0.010
Chi-square	0.004	0.069	0.034	0.020
MML-Normal and the long test				
Normal	0.000	0.010	0.002	0.002
Bimodal	0.000	0.013	0.004	0.003
Uniform	0.000	0.015	0.005	0.003
Chi-square	0.001	0.035	0.015	0.009
JML and the short test				
Normal	0.005	0.058	0.028	0.017
Bimodal	0.000	0.019	0.005	0.005
Uniform	0.000	0.010	0.005	0.004
Chi-square	0.004	0.056	0.015	0.015
JML and the long test				
Normal	0.001	0.018	0.009	0.005
Bimodal	0.000	0.014	0.006	0.004
Uniform	0.000	0.013	0.006	0.004
Chi-square	0.000	0.014	0.006	0.003
MML-Discrete and the short test				

Normal	0.000	0.009	0.004	0.003
Bimodal	0.000	0.010	0.003	0.003
Uniform	0.000	0.013	0.004	0.005
Chi-square	0.000	0.008	0.002	0.003
MML-Discrete and the long test				
Normal	0.000	0.013	0.003	0.003
Bimodal	0.000	0.013	0.003	0.003
Uniform	0.001	0.022	0.008	0.006
Chi-square	0.000	0.017	0.003	0.003

Table 2

Statistical summary of RMSE in item estimates

Ability distribution	Minimum	Maximum	Mean	SD
MML-Normal and the short test				
Normal	0.049	0.085	0.061	0.014
Bimodal	0.049	0.087	0.061	0.015
Uniform	0.048	0.090	0.062	0.016
Chi-square	0.056	0.110	0.071	0.019
MML-Normal and the long test				
Normal	0.048	0.099	0.063	0.013
Bimodal	0.048	0.101	0.063	0.013
Uniform	0.048	0.100	0.062	0.013
Chi-square	0.051	0.109	0.065	0.015
JML and the short test				
Normal	0.051	0.117	0.074	0.025
Bimodal	0.049	0.086	0.061	0.015
Uniform	0.049	0.088	0.061	0.014
Chi-square	0.052	0.101	0.064	0.017
JML and the long test				
Normal	0.049	0.101	0.063	0.013
Bimodal	0.048	0.100	0.063	0.013
Uniform	0.049	0.098	0.062	0.013
Chi-square	0.048	0.098	0.063	0.013
MML-Discrete and the short test				

Normal	0.049	0.086	0.061	0.014
Bimodal	0.049	0.085	0.060	0.014
Uniform	0.048	0.086	0.060	0.014
Chi-square	0.047	0.085	0.060	0.014
MML-Discrete and the long test				
Normal	0.048	0.100	0.063	0.013
Bimodal	0.048	0.098	0.062	0.013
Uniform	0.049	0.101	0.063	0.014
Chi-square	0.048	0.102	0.063	0.014

Table 3

Statistical summary of ratios of average error variance over sampling variance

Ability distribution	Minimum	Maximum	Mean	SD
MML-Normal and the short test				
Normal	0.944	1.119	1.040	0.054
Bimodal	0.970	1.152	1.058	0.056
Uniform	0.970	1.115	1.054	0.051
Chi-square	0.944	1.218	1.043	0.090
MML-Normal and the long test				
Normal	0.917	1.107	1.010	0.045
Bimodal	0.891	1.121	1.008	0.046
Uniform	0.961	1.123	1.018	0.037
Chi-square	0.903	1.118	1.014	0.051
JML and the short test				
Normal	0.793	1.109	0.963	0.112
Bimodal	1.022	1.234	1.134	0.063
Uniform	1.047	1.209	1.129	0.056
Chi-square	1.044	1.218	1.147	0.054
JML and the long test				
Normal	0.926	1.119	1.024	0.046
Bimodal	0.903	1.144	1.027	0.048
Uniform	0.970	1.152	1.039	0.039
Chi-square	0.932	1.114	1.032	0.048

Table 4

Statistical summary of bias and RMSE of sample variance estimates

	Bias				
Ability distribution	Minimum	Maximum	Mean	SD	RMSE
	MML-Normal and the short test				
Normal	-0.160	0.186	0.002	0.060	0.060
Bimodal	-0.134	0.226	0.039	0.059	0.071
Uniform	-0.112	0.267	0.059	0.058	0.083
Chi-square	-0.241	0.133	-0.056	0.064	0.085
	MML-Normal and the long test				
Normal	-0.105	0.123	0.003	0.037	0.037
Bimodal	-0.074	0.119	0.012	0.032	0.034
Uniform	-0.064	0.116	0.019	0.028	0.034
Chi-square	-0.186	0.075	-0.040	0.044	0.060
	JML and the short test				
Normal	0.488	1.668	0.671	0.092	0.677
Bimodal	0.472	0.875	0.676	0.062	0.679
Uniform	0.511	0.887	0.682	0.060	0.685
Chi-square	0.404	0.818	0.613	0.065	0.617
	JML and the long test				
Normal	0.036	0.274	0.148	0.038	0.153
Bimodal	0.064	0.255	0.147	0.031	0.150
Uniform	0.068	0.243	0.148	0.028	0.151
Chi-square	-0.052	0.274	0.134	0.053	0.144

	MML-Discrete and the short test				
Normal	0.490	0.881	0.674	0.065	0.678
Bimodal	0.477	0.882	0.683	0.063	0.686
Uniform	0.513	0.907	0.689	0.061	0.691
Chi-square	0.416	0.807	0.631	0.067	0.635
	MML-Discrete and the long test				
Normal	0.034	0.269	0.145	0.038	0.149
Bimodal	0.059	0.250	0.142	0.032	0.146
Uniform	0.069	0.244	0.150	0.028	0.153
Chi-square	-0.026	0.313	0.138	0.051	0.147

Figure Captions

Figure 1. Ability distributions used to generate simulated data

Figure 2. Bias of item difficulty estimates. (a) MML-Normal and the short test; (b) MML-Normal and the long test; (c) JML and the short test; (d) JML and the long test; (e) MML-Discrete and the short test; (f) MML-Discrete and the long test.

Figure 3. CDF graphs for four distributions

Figure 4. RMSE of item difficulty estimates. (a) MML-Normal and the short test; (b) MML-Normal and the long test; (c) JML and the short test; (d) JML and the long test; (e) MML-Discrete and the short test; (f) MML-Discrete and the long test.

Figure 5. Ratio of SE square over sampling variance of item difficulty estimates. (a) MML-Normal and the short test; (b) MML-Normal and the long test; (c) JML and the short test; (d) JML and the long test.

Figure 6. Bias of ability variance estimate

Figure 7. RMSE of ability variance estimate

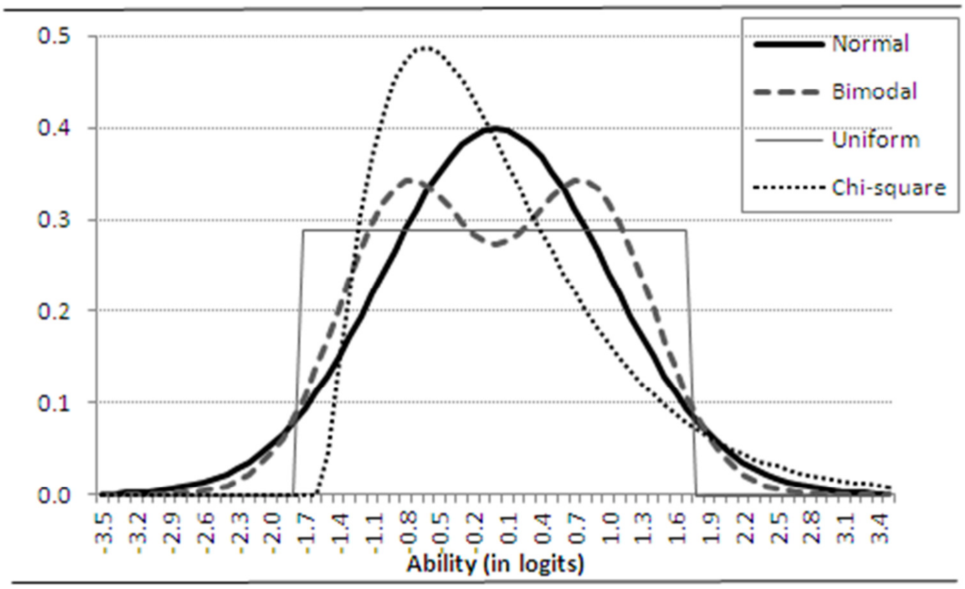
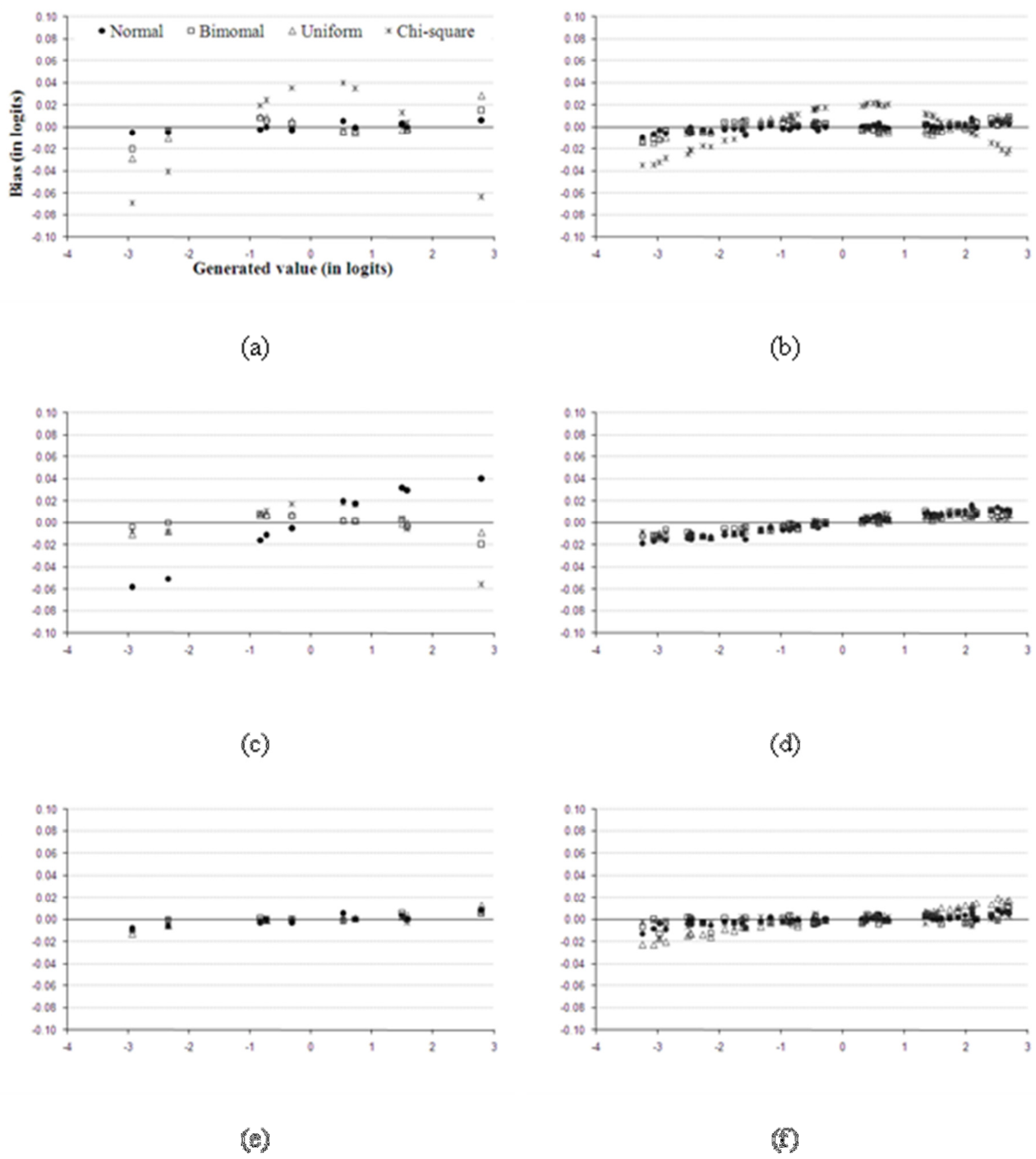


Figure 1. Ability distributions used to generate simulated data



(a) (b)

(c) (d)

(e) (f)

Figure 2. Bias of item difficulty estimates. (a) MML-Normal and the short test; (b) MML-Normal and the long test; (c) JML and the short test; (d) JML and the long test; (e) MML-Discrete and the short test; (f) MML-Discrete and the long test.

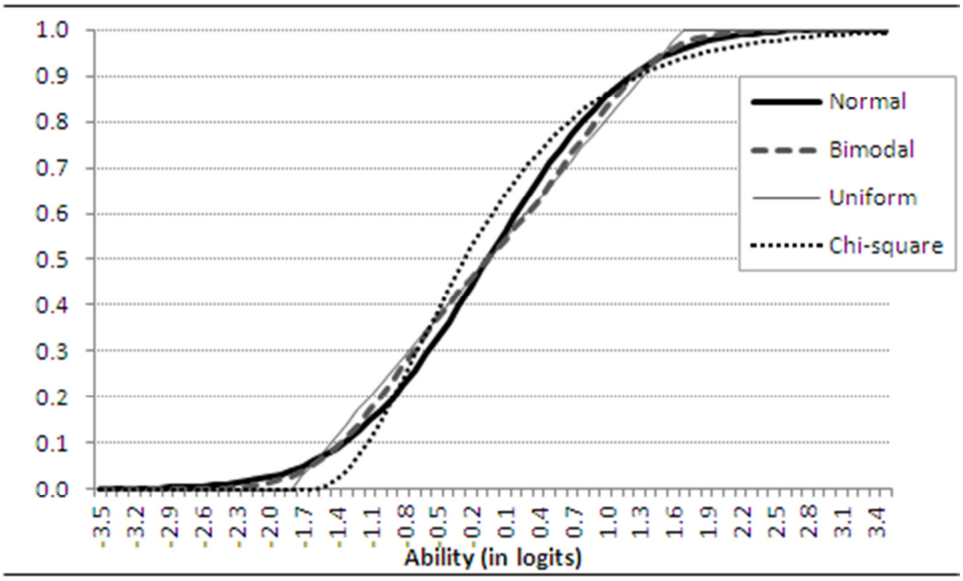
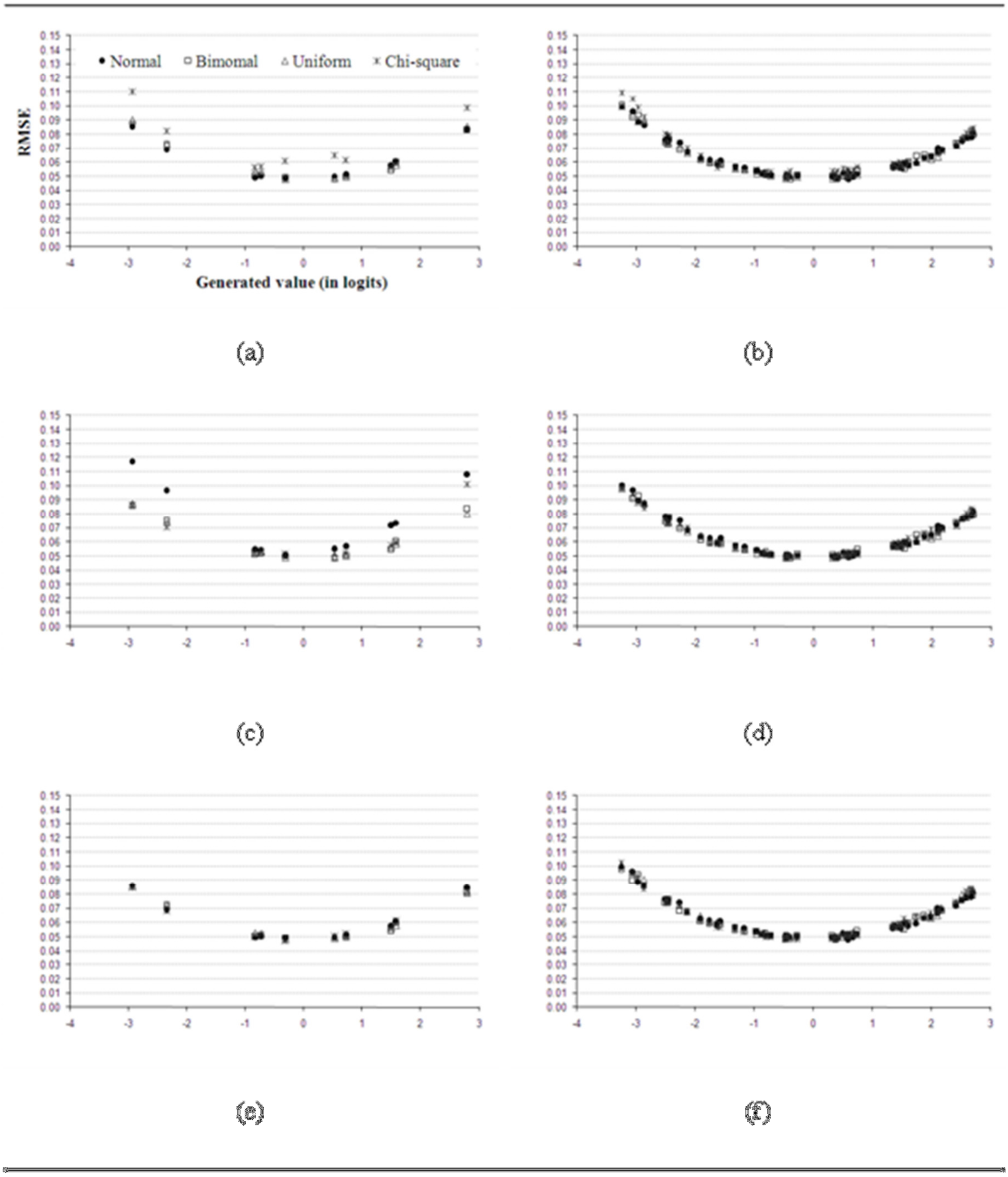


Figure 3. CDF graphs for four distributions



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4. RMSE of item difficulty estimates. (a) MML-Normal and the short test; (b) MML-Normal and the long test; (c) JML and the short test; (d) JML and the long test; (e) MML-Discrete and the short test; (f) MML-Discrete and the long test.

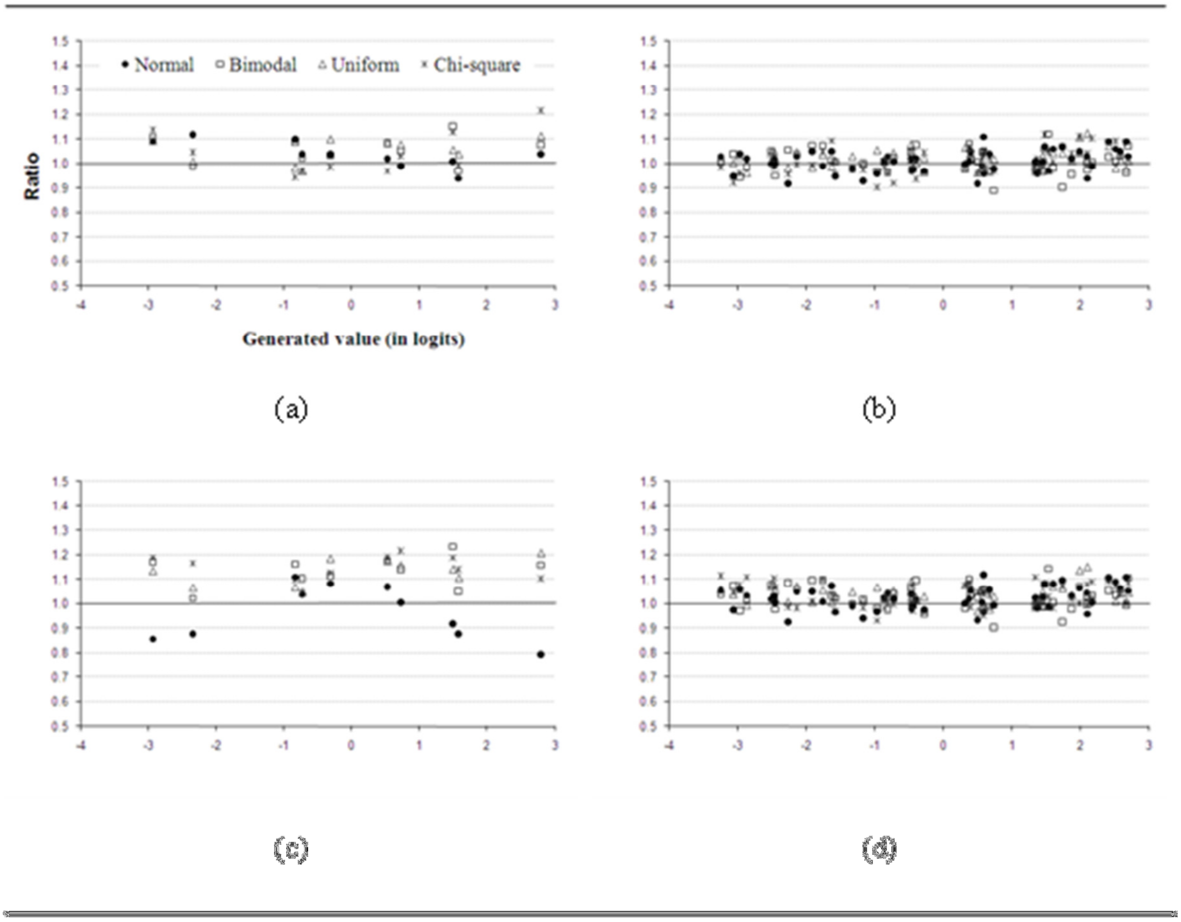


Figure 5. Ratio of SE square over sampling variance of item difficulty estimates. (a) MML-Normal and the short test; (b) MML-Normal and the long test; (c) JML and the short test; (d) JML and the long test.

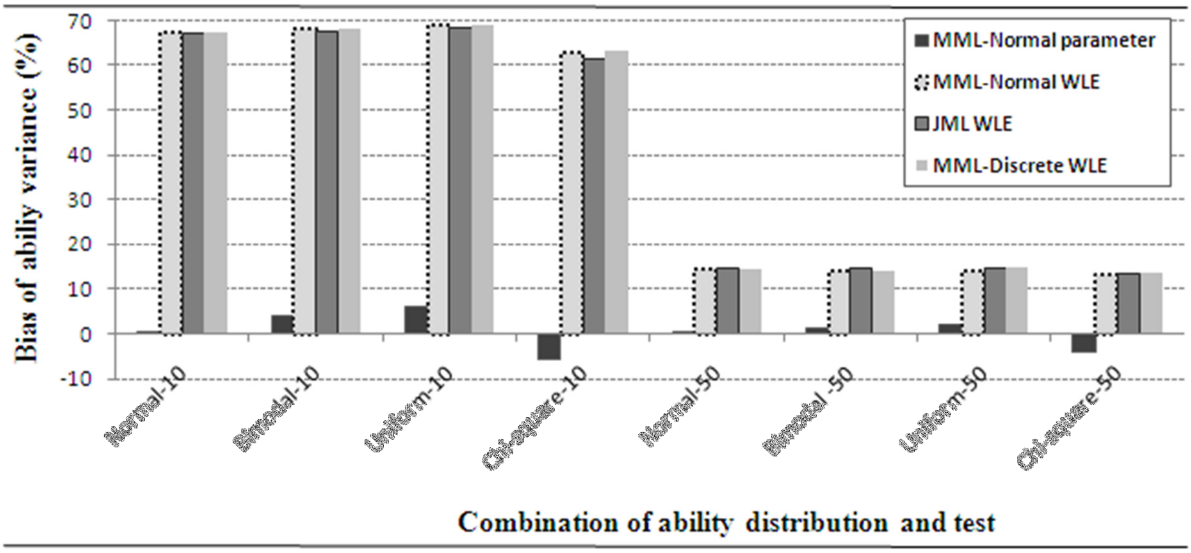


Figure 6. Bias of ability variance estimate

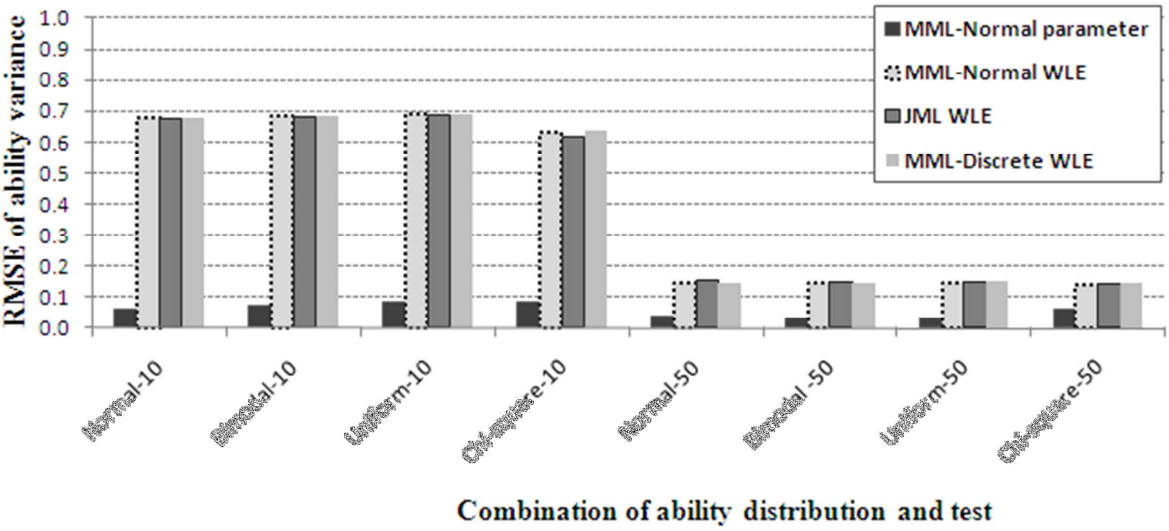


Figure 7. RMSE of ability variance estimate