# Final report on project under 1.3.301.2, Statistical Research Studies: Identification of Ability Distributions in IRT models for NAEP items

Prepared by:

Murray Aitkin and Irit Aitkin
School of Mathematics and Statistics
University of Newcastle-upon-Tyne

Prepared for:

US Department of Education
Office of Educational Research and Improvement
National Center for Education Statistics

This project was an activity of the Education Statistics Services Institute.

# Contents

# 1 Aim of the project

The current analysis of large-scale NAEP test data using item response models is based on the computational and statistical theory originally developed by Bock and Aitkin (1981) and subsequently developed by Mislevy (1986) and many others; a very detailed discussion of the theory and these developments is given in van der Linden and Hambleton (1997). Advances in theory and computing over the last 20 years have widened the possibilities for efficient and effective analyses of NAEP data. This project report examines the first of these possibilities:

- To examine the identifiability of the ability distribution underlying the item responses, and to assess the extent to which the current analysis could be made more effective, and/or simplified, as a result of this examination.

# 2 Summary

The results of this part of the study can be summarised as follows:

1. Modern theoretical and computational developments allow the *simultaneous* fitting of design strata parameters, item parameters, person (reporting group) parameters and latent ability distributions in a single multi-level (hierarchical) model structure. This model structure can be extended to additional levels to accommodate cluster sampling designs. The recognition of this general structure allows a unified computational approach to the analysis of NAEP data.

2. The 3PL model is in general unidentifiable. Bayesian methods using a prior distribution for the guessing parameter do not solve this problem, and the 3PL model is uninformative about ability unless the guessing parameter is known. In this case the model can be fitted by an extension of the method for the 2PL model.

3. For a unidimensional ability underlying all items, the true ability distribution is *essentially irrelevant* to the estimation of upper-level (reporting group) effects in the 2PL model. It *is* relevant to the estimation of the item parameters, but these are not important parameters for the reporting of NAEP results. Consequently,

4. Gaussian quadrature methods assuming a normal ability distribution can be used quite generally to estimate reporting group differences in a unidimensional ability over different true ability populations for the 2PL model – there is no need to use the more computationally intensive Bock-Aitkin semi-nonparametric method currently used to estimate the ability distribution. The semi-nonparametric method gives no improvement in *reporting group estimates*. The use of Gaussian quadrature would allow a substantially simplified analysis compared to the current NAEP analysis.

5. For a two-dimensional ability underlying all items in the 2PL model, upper-level (reporting group) estimates *are* affected, but not seriously, by ignoring the two-factor structure. For a general pattern of true loadings, the full two-factor model gives fully efficient upper-level parameter estimation, although the abilities themselves cannot be identified separately because of the rotational invariance of the two-factor model. However Gaussian quadrature for the 2PL model in two or more dimensions is computationally intensive, and alternative less intensive, though less efficient, analyses would be valuable. Setting the small loadings to zero to give independent blocks of items loading on each factor provides a nearly-efficient alternative.

# 3 Background

The two studies funded under 1.3.301.2 – Statistical Research Studies – examine features of the current methods for the analysis of NAEP data; these methods are based on the maximum likelihood and Bayesian methods for item response models originally developed for the 2PP model by Bock and Aitkin (1981) and extended by Mislevy (1986) and many others.

These methods were developed in the theoretical and computational climate of the mid-1980s, of slow processors and limited computer memory, and have remained relatively unchanged for some years. Developments in these areas in the last 20 years offer the possibility of improved analyses of NAEP, through faster and more general alternative computational methods.

The current approach to analysis through plausible values was developed to avoid the fitting of many large regression models incorporating the reporting group variables. This was replaced by the fitting of one extremely large model using 200 principal components of 1000 variables formed from all main effects and two-way interactions of the reporting variables, design variables and other relevant variables. This fitted model is used to simulate plausible values of individual ability from the individual posterior ability distributions, which are then used for any required reporting group comparisons, following the methods for multiple imputation developed by Rubin (1987). The stratification and clustering in the NAEP design are allowed for by weighting procedures in the current analysis.

The basis of the new methods is the recent recognition that item response models are a special case of generalized linear mixed models (GLMMs); they belong to a larger class of GLLAMMs (Skrondal and Rabe-Hesketh 2004) - *generalized linear latent and mixed models*. Thus the 2PL model can be recognized as a two-level Bernoulli variance component model with individuals at the upper level and items at the lower level, with conditionally independent logistic regressions of each item's binary response on a latent ability variable. In the 2PL model these regressions have different intercepts and slopes for each item; the Rasch model is the special case of identical

slopes for each item. A package capable of fitting GLMMs can fit both the Rasch and the 2PL model without any adaptation; we make use of this facility in our studies by using two packages which incorporate these models, STATA (which includes GLLAMM, a very powerful add-on to the STATA package) and GLIM, a generalized linear modeling package for which some macro facilities for GLMMs have been written (Aitkin and Francis 1995, Aitkin 1999).

A further important point is that these two-level models can be generalized to further levels, so that cluster designs, which themselves correspond to a two-level model, can be incorporated into the analysis directly, without weighting, through a three-level or four-level model of individuals – secondary sampling units – clustered in primary sampling units.

We do not discuss in this report the partial credit model of Masters (1982) for ordered polytomous response items which is also used in NAEP; this model is currently fitted by a further extension of the methods used in Bock and Aitkin (1981).

This report deals with the first study, of the extent to which the true ability distribution in the examinee population affects, or should affect, the analysis of NAEP data bearing on the major reporting groups.

# 4 Item response models

We consider three models in this report: the Rasch, 2PL and 3PL models. These models relate the binary (0,1 – incorrect or correct) response $Y$ of a respondent on an item to the latent ability $\theta$ of the respondent through a logistic regression. If an item has three or more *ordered response categories* the *partial credit model* is widely used (as for some NAEP items) but is not covered in this report.

## 4.1 The Rasch model

The response $Y_{ij}$ of respondent $i$ on item $j$ has probability $p_{ij}$ of being correct ($Y_{ij} = 1$), modelled by

$$\begin{aligned} \operatorname{logit} p_{ij} &= \alpha_j + \theta_i \\ p_{ij} &= \frac{1}{1 + \exp\{-(\alpha_j + \theta_i)\}}, \end{aligned}$$

where $\alpha_j$ is the "easiness" parameter of item $j$. This regression formulation will be used throughout this report, though the model is commonly re-

written in the psychometric literature as

$$\begin{aligned} \operatorname{logit} p_{ij} &= \theta_i - \delta_j \\ p_{ij} &= \frac{1}{1+\exp\{-(\theta_i - \delta_j)\}}, \end{aligned}$$

where $\delta_j = -\alpha_j$ is the *item difficulty* parameter.

### 4.2 The 2PL model

This model adds a *slope* parameter to the Rasch model for each item:

$$\begin{aligned} \operatorname{logit} p_{ij} &= \alpha_j + \beta_j \theta_i, \\ p_{ij} &= \frac{1}{1+\exp\{-(\alpha_j + \beta_j \theta_i)\}}, \end{aligned}$$

or in psychometric form

$$\begin{aligned} \operatorname{logit} p_{ij} &= a_j(\theta_i - b_j) \\ p_{ij} &= \frac{1}{1+\exp\{-a_j(\theta_i - b_j)\}}, \end{aligned}$$

where $a_j = \beta_j$ is the *item discrimination* parameter and $b_j$ is the difficulty parameter.

### 4.3 The 3PL model

The 2PL model does not allow for correct guessing in the absence of ability on the item. This is achieved in the 3PL model by adding a *guessing parameter* $c_j$ for item $j$ to the 2PL model:

$$p_{ij} = c_j + (1 - c_j)\frac{1}{1+\exp\{-a_j(\theta_i - b_j)\}}.$$

The lower asymptote of the logistic curve – the probability of a correct answer for a respondent with no ability – is then $c_j$ instead of zero.

## 5 Algorithms

Computational methods for fitting the Rasch, 2PL and 3PL models by maximum likelihood are of two kinds: those based directly on the EM algorithm, and those based on the second derivative (information) matrix using some form of Gauss-Newton method. These methods have different advantages and disadvantages.

EM methods are slow but relatively easy to program and guarantee a non-decreasing likelihood at each iteration. They do not provide standard errors for the model parameters without additional computation. In many models which are *unidentifiable* – the data do not provide enough information to estimate all the parameters – the EM algorithm nevertheless converges without difficulty (though slowly) to a *local* maximum of the likelihood surface. However in unidentifiable models there are other parameter values which have the same likelihood as the local maximum – in fact there is a whole *family* of parameter values all with the same likelihood. This phenomenon is described as a *ridge of maxima* in the parameter space – along this ridge all parameter values have the same maximized likelihood. The particular point to which the EM algorithm converges on the ridge is determined by the starting values used – changing the starting values will lead to convergence to another point with equal likelihood.

In unidentifiable models the information matrix is *singular* – it cannot be inverted to give the standard errors of the parameter estimates, because of the singularity (ridge) in the likelihood – there is no unique maximum of the likelihood. However if the EM algorithm is used without the additional computation of the information matrix, this singularity may not be observed, and it may be assumed that all the model parameters are identifiable.

Gauss-Newton methods using analytic second derivatives are much faster than EM, but may diverge – fail to converge – if starting values are not carefully chosen. In addition, if the model is unidentifiable then at some point in the Gauss-Newton iterations the information matrix will become singular and the algorithm will crash. Careful search methods have to be incorporated in GN algorithms to prevent this happening and to approximate the information matrix by one which can be inverted. Implementations of the GN algorithm generally report a *condition number*, which is a measure of closeness of the information matrix to singularity (it is the largest eigenvalue of the inverse of the information matrix). The larger this number, the closer is the information matrix to singularity. In extreme cases the condition number cannot be reported because the information matrix cannot be inverted.

The two progams we use for model fitting use both algorithms. The GLIM routines use EM and the GLLAMM routines use a version of GN in which the second derivatives are computed *numerically* rather than analytically. This makes the algorithm much slower, but allows very great generality in the class of models which can be fitted.

# 6 Unidentifiability of the 3PL model

Despite its appealing form in allowing for random guessing, and its common use in large-scale testing programs, the 3PL model suffers a serious identifi-

ability problem: without external information about the model parameters, the three parameters cannot be identified from test data – that is, the 3PL model is unidentifiable. This result follows from the general unidentifiability of compound or mixed Bernoulli models – they remain Bernoulli. The practical effect of this problem is to confound the $a$ and $c$ parameters, especially when $c$ is large, because an item with a large guessing parameter will be hard to distinguish from an easy item. Attempts to fit the model result in inconsistent estimates of the guessing parameters, with sampling variances which remain constant as the sample size increases.

We observed a related result in fitting the 3PL model to the LSAT7 data in GLLAMM with a *common* $c$ parameter for all items by constructing the *profile likelihood* in the $c$ parameter, by setting this parameter at different values on a grid and maximizing the likelihood over the other item parameters. The profile likelihood was flat, showing that the likelihood had a ridge of local maxima depending on the specification of the $c$ parameter.

The difficulty with this model is well-known: van der Linden and Hambleton (1997) comment (p. 16) "Unless huge samples ... or tight priors around the true parameter values are used, the [parameter] estimates are unstable." McDonald (1997) in the same reference notes (p. 265) "An attempt to estimate these parameters in an early version of the [NOHARM] program suffered the usual difficulties for models with such parameters."

Current NAEP practice with this model is to specify very tight prior distributions for the guessing parameters, which produce estimates for these parameters within the very narrow ranges of the priors. The resulting estimates depend completely on this prior specification, however, and convey a misleading impression of precision which is due only to the prior specification, which is in effect the specification of the value of the guessing parameter.

It seems to us bad statistical practice to routinely use unidentifiable models which can be made identifiable only by these very tight prior specifications. It would be better in our view to simply set the guessing parameters at the prior modal value and simplify the analysis correspondingly (and substantially).

We do not consider the 3PL model further in this report, but in a later study we will examine the effect of varying the true guessing parameter on the parameter estimates from the 2PL analysis.

# 7 Importance of the ability distribution

Individual ability estimates from NAEP subscales are fundamental to NAEP reporting. Although individual scores are not themselves reported, current indirect estimation methods for reporting group differences use individual ability estimates from plausible values generated from the posterior distri-

bution of ability for each examinee, and direct estimation methods use the ability distribution directly. Thus the distribution of true ability is central to both estimation methods.

If a normal distribution N(0,1) is assumed for the ability distribution (the usual approach), the likelihood which has to be maximized has no analytic form and has to be expressed as an integral. This integral is approximated by *Gaussian quadrature*, in which the continuous normal density is replaced by a discrete distribution at a finite set of *quadrature points*, with probabilities given by the normal density at these points, scaled to sum to one. The integral is thus replaced by a finite weighted sum over these quadrature points, with weights given by the scaled probabilities.

However this normal assumption may be false, and this might lead to biased conclusions about reporting group differences, from either the MLEs of these effects or from the plausible values generated from the posterior distribution assuming this "discrete normal" ability distribution. For this reason the current NAEP analyses are based on an ability distribution estimated on 40 equally spaced quadrature points - a forty-parameter multinomial distribution. This estimation is intended to allow for non-normal distributions, by allowing the probability ordinates to be unknown parameters estimated from the test scores. This estimate was proposed by Bock and Aitkin (1981) as a semi-parametric estimate, taking the quadrature point locations as fixed but estimating the probabilities.

A more recent alternative is to use fully nonparametric estimation, where both the quadrature point locations *and* the probabilities are estimated. This approach, a fully nonparametric estimation of the ability distribution, was discussed at length for variance component models by Aitkin (1999). This paper showed that important differences *can* occur between models assuming normal random effects and those in which the random effect distribution is estimated nonparametrically, and that therefore it is prudent to use nonparametric estimation at least to investigate the validity of a normal assumption.

If the ability distribution is itself estimated, three important theoretical questions arise:

- To what extent is this estimation possible?
- If it is possible, is fully nonparametric estimation superior to semi-parametric estimation, and/or Gaussian quadrature?
- What difference does the method of estimation of this distribution make to reporting group differences?

In this part of the report we give answers to these questions in the context of small tests with five or 10 items. This limitation to small tests allows large-scale simulations to run in reasonable time and cover a wide range of possible distributions and methods.

The basis of our answers, as noted above, is the two-level model of items within individual examinees. A standard feature of multi-level modeling is the routine use of both upper- (person) level and lower- (item) level explanatory variables in the model. In item response models, upper-level variables may be used to define the important reporting group membership of each examinee. Since *both* the item parameters *and* the reporting group parameters are estimated *simultaneously* from the data, it is not necessary to fix the item parameters in advance. This is a major benefit of the multi-level (or hierarchical) modeling approach.

In the second part of this report we compare this approach with the current NAEP analysis which does not directly use the group membership estimates from the multi-level model.

The effect of different true ability distributions and of different methods of analysis is assessed through simulation studies. We simulate data, generally 1000 samples of size 1000, from 2PL models with five or 10 binary items, and a single upper-level reporting group explanatory variable, called here sex, or (for the two-factor models) a main effect model of sex and ethnic group. We examine the effect, on estimation of the item and reporting group parameters, of varying both the true ability distribution (from normal to very skewed continuous, and discrete) and the estimation of this distribution.

# 8 Simulation studies

## 8.1 Study 1

In the first study we assessed the success of Gaussian quadrature in dealing with a normal, a binary and a skewed continuous ability distribution. The true model is the 2PL with five items, the item parameters being given by the MLEs of the LSAT7 data set used in Bock and Aitkin (1981), with a sex effect of 0.5 added to ability for half the sample of 1000 on the logit scale. The item parameters are given on the logit scale in Table 1.

Table 1: LSAT7 items

| Item | easiness | discrimination |
|---|---|---|
| 1 | 1.856 | 1.0 |
| 2 | 0.808 | 1.081 |
| 3 | 1.805 | 1.708 |
| 4 | 0.486 | 0.765 |
| 5 | 1.855 | 0.736 |

The item characteristic curves (ICCs) for these items are shown in Figure 1.

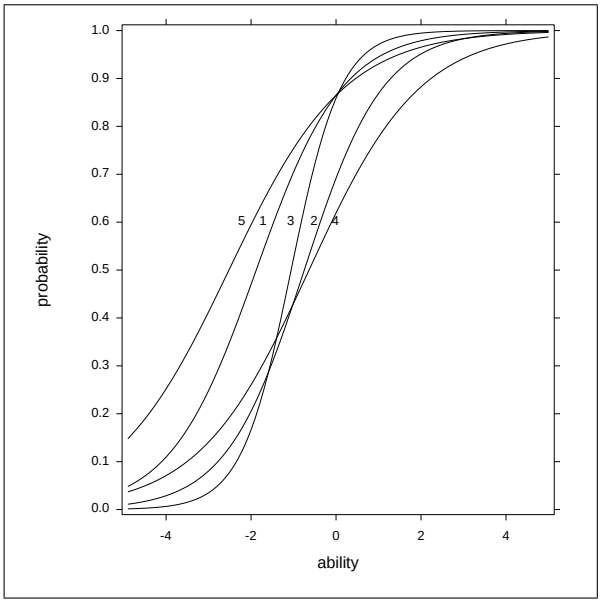


Figure 1: Item characteristic curves, LSAT7 items

Three ability distributions were considered; all three distributions have mean 0 and variance 1:

- N(0,1);
- a symmetric 2-point distribution, with masses of 0.5 at points -1 and +1 on the logit scale;
- a skewed continuous distribution, a mixture of N(.105, 0.75) and N(-2, 1.539), in proportions 0.95 and 0.05 respectively.

Only one fitting method – 8-point Gaussian quadrature (GQ) – was used in GLLAMM. Comparisons over 1000 samples are of the biases and mean square errors (MSEs) of the GQ parameter estimates under the different true ability distributions. In all analyses reported, the item 1 slope estimate was constrained to 1 for identifiability, so no biases or MSEs are reported for this parameter.

Results tables are in Appendix 1. The main points of the results are:

- For the sex effect, the normal and mixture distributions had 6-7% bias, the binary distribution had 1.8% bias. The biases were all small (around 10%) compared to the standard errors of the estimates.

  The MSEs for the sex effect were 20% larger for the normal and mixture compared to the binary.

- For the item intercepts, the normal and binary distributions were very similar; the mixture distribution had larger biases – up to 1 SE for a3 – and MSEs.

- For the item slopes, the binary distribution had the largest biases and MSEs.

Unexpectedly, both the non-normal distributions gave quite consistent estimates of all the parameters. This suggests that the assumption of a normal ability distribution may not be restrictive.

In the next study we looked in more detail at estimation for the 2-point ability distribution, by various methods.

### 8.2 Study 2

In this study the true model is the 2PL with parameters as in the previous study, the item parameters having the LSAT7 estimates as true values and the sex difference being 0.5 units on the logit scale.

The ability distribution is binary, a 2-point distribution with masses of 0.5 at points -1 and +1 on the logit scale; this distribution has mean zero and variance 1.

Four fitting methods for the 2PL model were used: 1) 8-point Gaussian quadrature, and 2) 2-point, 3) 3-point and 4) 4-point fully nonparametric estimation.

Because the fully nonparametric fitting is computationally intensive, only 40 replicate samples were generated from the true model, and the number of iterations of the fully nonparametric methods was limited to 20. The bias, variance and MSE were calculated for each parameter across the 40 samples. Results tables are in Appendix 2.

The main results are:

- The sex effect was unbiasedly estimated by all four methods, the worst relative bias being only 3.8% for method 1; method 2 gave the smallest MSE (smallest bias and variance), but the worst method (1) was only 8% worse in MSE.

- For the intercept estimates, method 1 performed uniformly best except on item 5 where it was second best. Methods 3 and 4 were consistently bad.

- For the slopes, method 1 performed best overall, mainly because none of the nonparametric methods could cope with item 3, with its large parameters. Methods 3 and 4 were not much worse than method 2.

Surprisingly, the correct 2-point analysis did not give the best estimates of the item parameters, though it gave the (just) best estimate of the sex effect

(though this was not much better than the Gaussian quadrature estimate). Gaussian quadrature performed consistently well though it corresponds to a mis-specified ability distribution. Overfitting the number of masspoints performed consistently poorly.

In the next study we extended the discrete ability distribution to 3 points.

### 8.3 Study 3

In this study the true model is the 2PL with parameters as in the previous study, the item parameters having the LSAT7 estimates as true values and the sex difference being 0.5 units on the logit scale.

The ability distribution is a 3-point one, with masses of 2/3 at zero and of 1/6 at points -1.732 and +1.732 on the logit scale. This distribution has mean 0 and variance 1, and is the discrete distribution for 3-point Gaussian quadrature.

Four fitting methods for the 2PL model were used, as in Study 2, in 40 replicate samples generated from the true model: 1) 8-point Gaussian quadrature, and 2) 2-point, 3) 3-point and 4) 4-point fully nonparametric estimation. The number of iterations of the fully nonparametric methods was limited to 20.

The bias, variance and MSE were calculated for each parameter across the 40 samples. Results tables are in Appendix 3.

The main results are:

- The sex effect had some negative bias (0.6-0.7 SEs) by all four methods, the relative biases ranging from 12% for method 3 to 14% for method 2; method 3 gave the smallest MSE (smallest bias and variance), but the worst method (2) was only 4.4% worse in MSE.

- For the intercept estimates, methods 1 and 2 performed equally well. Methods 3 and 4 were consistently bad, with MSEs 10 or more times those of methods 1 and 2.

- For the slopes, method 2 performed slightly better than method 1, much better (a factor of 5 in MSE) than methods 3 and 4 on items 2 and 3, and slightly better on items 4 and 5.

Somewhat surprisingly, the 3-point nonparametric analysis did not give the best estimates of the item parameters; it gave slightly the best estimate of the sex effect, though this was not much better than the Gaussian quadrature estimation. The 2-point estimation and Gaussian quadrature performed about equally well. The 3-point and 4-point estimation of the item parameters was far off for all the intercepts, and for several slopes, suggesting that the number of mass-points used was beyond the number identifiable from the 5 items. This result is supported by the available theoretical results, which

suggest that with $p$ items, at most $(p-1)/2$ masspoints are identifiable in the fully nonparametric approach.

In the next study we examined the performance of the (incorrect) Rasch model as well as the 2PL, with a normal ability distribution.

### 8.4 Study 4

In this study the true model is the 2PL with parameters as in the previous studies, the item parameters having the LSAT7 estimates as true values and the sex difference being 0.5 units on the logit scale. The ability distribution is N(0,1).

Four fitting methods were used, in 250 replicate samples generated from the true model: 1) the Rasch model (with only item difficulty and sex parameters) with 8-point Gaussian quadrature, 2) the 2PL model with 8-point Gaussian quadrature (the most nearly correct analysis), 3) the Rasch model with 2-point nonparametric estimation, and 4) the 2PL model with 2-point nonparametric estimation. The number of iterations of the 2-point nonparametric method was limited to 20. The bias, variance and MSE were calculated for each parameter (excluding the slopes for the Rasch analyses) across the 250 samples. Results tables are in Appendix 4.

The main results are:

- The sex effect had a small negative bias (less than 0.5 SEs) by all four methods, ranging from 1.1% for method 1 to 8.8% for method 4. Methods 1 and 3 had slightly smaller MSEs than methods 2 and 4.
- For the intercept estimates, method 1 performed best overall, though on item 5 it was worst. The other methods were similar.
- For the slopes, method 2 performed slightly better than method 4, and on item 3 method 4 was spectacularly worse, with a MSE twice that of method 2.
- The 2-point nonparametric analysis was worse than Gaussian quadrature for all parameters in the Rasch model except for intercept 5; for the 2PL model the 2-point analysis was better for intercepts 1, 2 and 5 but worse for the others, and better for slopes 2 and 5 but worse for 3 and 4.

Though it is not really of interest, the MSE can be calculated (as the squared bias, since the variance is zero) for the (constant) slope "estimates" of 1 in the Rasch model, and can be compared with the MSEs of the 2PL estimates. The Rasch MSEs are smaller for item 2, slightly larger for items 4 and 5, and much larger for item 3.

**Conclusions from the four studies**

For studies 1, 2 and 3 with 2-point, 3-point and asymmetric continuous ability distributions, Gaussian quadrature performed consistently well, and the "correct" number of masspoints in the discrete cases did not guarantee any better analysis.

The success of the Rasch model in Study 4 was unexpected, since it is definitely incorrect – the improvement in likelihood for the 2PL model over the Rasch model was uniformly large. In particular, its most precise estimation of the upper-level sex difference with an incorrectly specified model raises the important question of whether the effects of upper-level variables can be generally estimated correctly by an incorrect Rasch model.

If true, this would save a very large amount of computing time and effort, because the item parameters themselves are "nuisance" parameters (in the statistical sense) for the NAEP analysis – it is the upper-level demographic variables that are important, and these are estimated very well by Gaussian quadrature, even in the Rasch model.

In the fifth study we assessed the success of Gaussian quadrature with varying numbers of quadrature points in dealing with a true Gaussian ability distribution.

### 8.5 Study 5

The aim of this study was to assess the effect of varying the number of quadrature points in Gaussian quadrature.

In this study the same generating model was used as in Study 1, that is the 2PL model for the LSAT7 data. In all simulations the ability distribution is N(0,1) for males and N(0.5,1) for females, with 500 males and 500 females in the data set.

Models fitted were: 1) the 2PL model using Gaussian quadrature with 14, 8 and 2 points, and 2) the Rasch model using Gaussian quadrature with 14, 8 and 2 points.

As before, the bias, variance and mean square error for each of the intercept and slope parameters was calculated over 1000 independent samples generated from the model. All computations were done in GLIM4 using variance component macros developed for binary data; these use the EM algorithm for Gaussian quadrature, with a limit of 500 iterations. Results tables are in Appendix 5.

The main results are:

- For the 2PL analysis, 14-point and 8-point quadrature gave almost identical results on all criteria, with 8-point quadrature giving slightly smaller mean square errors than 14-point quadrature on intercepts, and comparable mean square errors on slopes and the sex effect. 2-point quadrature was substantially worse, with large biases and variances on item 3 for both parameters.

- The average deviance for 8-point quadrature was equal to that for 14-point quadrature over the 1000 simulations, while that for 2-point quadrature was larger by about 13.

- For the incorrect Rasch model, 14-point and 8-point quadrature again gave almost identical results, while 2-point quadrature gave much worse results. The sex effect was estimated slightly more precisely in this model than in the 2PL model, with a reduction of 3% in mean square error.

- The average deviance for 8-point quadrature was equal to that for 14-point quadrature over the 1000 simulations, while that for 2-point quadrature was larger by about 11. The 14- and 8-point Rasch deviances were about 17.5 larger than the corresponding 2PL deviances.

**Conclusions from study 5**

It is clear that Gaussian quadrature with a large number of mass-points is unnecessary for accurate estimation of model parameters in this data set. Eight points is as accurate as 14, but 2 points is inaccurate especially for the large slope and intercept parameters of item 3.

Estimation of the sex effect was slightly more accurate in the Rasch model, with a smaller bias, though this model is clearly incorrect, with a deviance difference compared to the 2PL model averaging 17.5 on 4 degrees of freedom.

### 8.6 Study 6

In this study we evaluated the semi-nonparametric approach suggested by Bock and Aitkin (1981), of estimating the ability distribution only through the probabilities associated with fixed mass-points – the latter were not themselves estimated as in the fully nonparametric approach evaluated in the earlier studies.

The generating model for true ability was an extremely skewed distribution with a very heavy tail of low abilities: a 14-point "triangular" distribution defined on the 14 mass-points used in 14-point Gaussian quadrature. The distribution had mass 14/105 at the extreme negative mass-point -4.73, decreasing linearly across the mass-points to a mass of 1/105 at the extreme positive mass +4.73. This was rescaled in the simulations to have mean 0 and variance 1.

The test-score model was again the 2PL LSAT7 model of Bock and Aitkin. Gaussian quadrature with 14 and 5 mass-points, and semi-NP analysis with 14 and 5 mass-points were used to fit the model. Results are given in Appendix 6 separately for 14-point and 5-point fitting.

For 14-point quadrature, the semi-NP analysis reduced the deviance by an average of 18.48 for the 12 extra parameters (13 estimated probabilities,

compared to the standard deviation in the GQ analysis), not a large reduction. The sex effect had almost zero bias by both methods, and its variance was reduced by 6% in the semi-NP analysis compared to GQ – the efficiency of the GQ estimate was 94%. The GQ item intercepts were heavily biased – as much as 1.9 standard errors for a3 – and the slope biases were worse, as much as 3 standard errors. The biases in the semi-NP intercept estimates were much smaller, not more than 0.2 standard errors; the slope biases were also smaller but large enough to be serious, as much as 2 standard errors. The variances of the GQ slope estimates were generally much smaller than those of the semi-NP estimates, and the MSEs were generally smaller.

The semi-NP estimation clearly improves the bias of the GQ item parameter estimation, although the slope estimates were still badly affected, but this gain is at the expense of much larger variances. It is impressive that the GQ upper-level sex effect estimate was unaffected by the gross misspecification of the ability distribution, and lost only 6% efficiency compared to the semi-NP estimate.

The Rasch estimates behaved badly in both analyses, with some very large biases and variances of the intercepts, and large variances for the sex effects, more than 40 times those for the 2PL model.

For 5-point quadrature, the GQ item parameter estimates were better than from 14-point GQ, with smaller biases and smaller variances. The semi-NP estimates were worse than those for 14-point semi-NP – larger biases and larger variances. The GQ estimates again had generally smaller MSEs than the semi-NP estimates. The sex effect again had almost zero bias, but had a slightly larger variance from GQ and a slightly smaller variance from semi-NP, and the efficiency of the GQ estimate was reduced to 80%.

The average semi-NP deviance was smaller by 40 than the average GQ deviance for the three extra parameters, showing a substantial improvement in fit, though the 5-point semi-NP deviance was almost the same as the semi-NP 14-point deviance, showing somewhat surprisingly that the additional estimated probability parameters improved the item parameter estimates but not the deviance.

The 14-point ability distribution was naturally not well estimated by the 5-point semi-NP masses, which averaged 0.0754, 0.2751, 0.2712, 0.3511 and 0.0272, on the mass-points -2.8570, -1.3556, 0, 1.3556 and 2.8570.

**Conclusions from Study 6**

The ability distribution used in this study was more extreme than any real ability distribution is likely to be. Despite this Gaussian quadrature with only 5 points performed very well in estimating the sex difference, nearly as well as 14-point GQ. For the item parameters the GQ estimates had large biases in both cases, but much smaller variances and MSEs, so there seems little benefit in the semi-NP estimation.

## 8.7 Study 7

In the following studies we examined the estimation of reporting group differences in a two-ability-factor model with 10 items and several patterns of item loadings on the two factors, with sex and ethnic group differences in a main-effect regression model.

The two-factor model has a rotational invariance as in the normal factor model: the factor loadings can be identified only up to an arbitrary rotation in the two-dimensional space of the factors. So in particular *we cannot identify the separate factors in the two-factor model*, only the *sum* of the "linear predictors" from both factors. Identification of the separate factors requires a strong prior specification of zero loadings on the factors; in general such a specification will imply correlated factors.

It may be clear on a particular test that items are *intended* to involve only one factor and thus such a prior specification may be reasonable. We have used this implicitly in the analysis in which we fit a 2-factor model with restricted loadings. However it is important to realise that the separate factors can be identified *only* by this specification. In our model of reporting group differences we have assigned these differences to the linear predictor, and not to the abilities on the separate factors. Different group differences on the two abilities would be confounded into a composite set of group differences on the linear predictor, so we do not attempt to model separate group differences on each factor.

In study 7 all 10 items loaded on both factors, with loadings decreasing with item number on the first factor and increasing with item number on the second. Factor loadings were

Table 2: Intercepts and loadings

| item | int. | slope1 | slope2 |
|---|---|---|---|
| 1 | 0.856 | 2.0 | 0.2 |
| 2 | -1.192 | 1.8 | 0.4 |
| 3 | -0.195 | 1.6 | 0.6 |
| 4 | -1.514 | 1.4 | 0.8 |
| 5 | -0.145 | 1.2 | 1.0 |
| 6 | -0.012 | 1.0 | 1.2 |
| 7 | -0.051 | 0.8 | 1.4 |
| 8 | 0.576 | 0.6 | 1.6 |
| 9 | -0.367 | 0.4 | 1.8 |
| 10 | -0.396 | 0.2 | 2.0 |

The two factors were independently N(0,1).

The reporting group regression structure was a main effect model of sex and ethnic group, with two sex groups M and F and four ethnic groups W, B, H and A with ability values on the logit scale shown below.

Table 3: Ability parameters

| eth/sex | M | F |
|---|---|---|
| W | 0.689 | 0.217 |
| B | -1.670 | -2.142 |
| H | -1.198 | -1.670 |
| A | 1.632 | 1.161 |

The sex difference (M-F) is 0.472 across all ethnic groups, and the ethnic group differences, consistent for both sexes, are (B-W) -2.359, (H-W) -1.887, and (A-W) 0.943. The sample of 1000 was generated as 500 M and 500 F, with proportions of 0.65 W, 0.25 B, 0.15 H and 0.05 A in each sex group.

Four models were fitted to the 1000 observations on the 10 items: a Rasch model, a single-factor 2PL model, a full 2-factor 2PL model and a model with orthogonal blocks of loadings with items 1-5 loading only on factor 1 and items 6-10 loading only on factor 2. The omitted items on each factor were those with the smallest true loadings on that factor.

Results tables are in Appendix 7, based on 125 samples because of the computational intensity of the 2-factor model. The main results are:

- The Rasch and 2PL models had much larger biases, but smaller variances, of the reporting group estimates than the two 2-factor models.
- The MSE efficiencies were similar, with the 2-factor models better, though the differences were smaller since the biases contributed less than 50% to the MSE.
- The model with constrained loadings was generally slightly better than the full 2-factor model.

### 8.8 Study 8

In this study the reporting group regression model was the same as in Study 7, but the item model structure was different. Items 1-5 loaded zero on the second factor, and items 6-10 loaded zero on the first factor. The non-zero loadings were the same as in Study 7.

Five models were fitted to the 1000 observations on the 10 items: a single-factor 2PL model, a full 2-factor 2PL model, a model with orthogonal blocks of loadings with items 1-5 loading only on factor 1 and items 6-10 loading only on factor 2, and two separate 1-factor 2PL models each using five items only. The omitted items on each factor were those with the smallest true loadings on that factor.

Results tables are given in Appendix 8, based on averages across 67 samples. The main results are:

- The separate 1-factor models using each set of items gave very badly biased estimates of all the reporting group parameters except the W-A difference. The variances of all the estimates were much smaller in the 1-factor models but the large biases outweighed the variances, giving large MSEs.

- The full 2-factor model had the smallest biases, but the largest variances, of the three models with all 10 items. The 1-factor 2PL model had the smallest variances but the largest biases, though these were not misleadingly large, ranging from 0.5 to 1 standard errors.

- The 5- and 5- item 2-factor model performed very well with MSE efficiency relative to the full 2-factor model of at least 86%; it was over 100% on two parameters.

- Relative to the variances of the 2-factor model with five items on each factor, the 1-factor model was quite efficient in MSE, ranging from 62% on the B-W difference to 105% on the A-W difference, with an average of 88%.

**Conclusions from Studies 7 and 8**

Since fitting the 2-factor 2PL model is computationally very intensive, any methods which avoid this and still give good estimates of the reporting group differences (assumed constant across the factors) are valuable. The 2-factor model with small loadings set to zero performed very well, and the 1-factor 2PL model performed quite well in these studies; it gave biased estimates of the parameters but these were within 0.5-1 standard errors of the true values, so confidence intervals for the true values would cover them, at least in samples of this size (1000). These two simpler models appear to be candidates for simpler analyses of the 2-factor model.

# 9 References


Aitkin, M. (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117-128.

Aitkin, M. and Francis, B. (1995) Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *The GLIM Newsletter* **25**, 37-45.

Bock, R.D. and **Aitkin, M.** (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika* **46**, 443-459.

McDonald, R.P. (1997) Normal-ogive multidimensional model. in *Handbook of Modern Item Response Theory*, eds. van der Linden, W.J and Hambleton, R.K. Springer-Verlag, New York.

Masters, G.N. (1982) A Rasch model for partial credit scoring. *Pyschometrika* **47**, 149-174.

Mislevy, R.J. (1986) Estimating latent distributions. *Psychometrika* **49**, 359-381.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall/ CRC, Boca Raton.

van der Linden, W.J and Hambleton, R.K. (eds) (1997) *Handbook of Modern Item Response Theory*. Springer-Verlag, New York.

# 10 Appendix

The results tables for all studies are collected below. The STATA output tables are reproduced verbatim. The item intercepts are denoted by a1,...,a5 and slopes by b1,...,b5. In the studies using GLLAMM the coefficient of b1 was fixed at 1 for identifiability.

## 10.1 Appendix 1

a1

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| 1.856 | 1.797127 | .0218824 | -.0588728 | .0253484 | binary |
| 1.856 | 1.968831 | .026762 | .1128308 | .0394928 | mixed |
| 1.856 | 1.891235 | .0248755 | .0352348 | .026117 | normal |

a2

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| .808 | .817211 | .0152092 | .009211 | .0152941 | binary |
| .808 | .8936521 | .013737 | .0856521 | .0210733 | mixed |
| .808 | .8537821 | .0145449 | .0457821 | .0166409 | normal |

a3

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| 1.805 | 1.596338 | .0329609 | -.2086621 | .0765008 | binary |
| 1.805 | 2.082867 | .0742924 | .2778671 | .1515025 | mixed |
| 1.805 | 1.92767 | .0592209 | .1226699 | .0742688 | normal |

a4

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| .486 | .4440137 | .0081615 | -.0419863 | .0099244 | binary |
| .486 | .4665201 | .0083809 | -.0194799 | .0087603 | mixed |
| .486 | .4505651 | .008371 | -.0354349 | .0096267 | normal |

a5

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| 1.855 | 1.783504 | .01741 | -.0714964 | .0225218 | binary |
| 1.855 | 1.876072 | .0187342 | .0210723 | .0191782 | mixed |
| 1.855 | 1.829017 | .0180708 | -.0259828 | .0187459 | normal |

b2

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| 1.081 | 1.36342 | .1683123 | .2824199 | .2480733 | binary |
| 1.081 | 1.02425 | .0727133 | -.0567497 | .0759339 | mixed |
| 1.081 | 1.130509 | .0990589 | .0495091 | .1015101 | normal |

b3

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| 1.708 | 1.561341 | .2278932 | -.1466592 | .2494022 | binary |
| 1.708 | 1.611721 | .2131111 | -.0962789 | .2223807 | mixed |
| 1.708 | 1.666401 | .2343722 | -.0415986 | .2361027 | normal |

b4

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| .765 | 1.007395 | .093239 | .2423948 | .1519942 | binary |
| .765 | .7269949 | .0344877 | -.0380051 | .035932 | mixed |
| .765 | .8088664 | .0511669 | .0438665 | .0530911 | normal |

b5

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| .736 | .8004031 | .0669211 | .0644031 | .0710688 | binary |
| .736 | .7694765 | .0416956 | .0334766 | .0428163 | mixed |
| .736 | .7579142 | .0546278 | .0219142 | .055108 | normal |

sex

| true | mean | var | bias | mse | true dist |
|---|---|---|---|---|---|
| .5 | .4908668 | .0103616 | -.0091332 | .010445 | binary |
| .5 | .4623752 | .0109866 | -.0376248 | .0124022 | mixed |

```
  .5     .4683005    .0118753  -.0316995    .0128802      normal
-------------------------------------------------------------------
```

## 10.2 Appendix 2

a1

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.856 | 1.838735 | .0105417 | -.0172652 | .0108398 | 1 |
| 1.856 | 1.91969 | .0250298 | .0636899 | .0290862 | 2 |
| 1.856 | 2.127784 | .2479332 | .2717838 | .3217996 | 3 |
| 1.856 | 2.2435 | .4767379 | .3875 | .6268942 | 4 |

a2

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .808 | .7853094 | .0103963 | -.0226906 | .0109111 | 1 |
| .808 | .795384 | .013752 | -.012616 | .0139112 | 2 |
| .808 | .9773363 | .2128067 | .1693363 | .2414815 | 3 |
| .808 | 1.076407 | .2959318 | .2684072 | .3679742 | 4 |

a3

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.805 | 1.659703 | .041686 | -.145297 | .0627976 | 1 |
| 1.805 | 3.14262 | 5.520508 | 1.33762 | 7.309736 | 2 |
| 1.805 | 3.439784 | 4.681503 | 1.634784 | 7.354023 | 3 |
| 1.805 | 3.597236 | 6.17927 | 1.792236 | 9.391379 | 4 |

a4

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .486 | .4235497 | .0089648 | -.0624504 | .0128648 | 1 |
| .486 | .4121228 | .009285 | -.0738772 | .0147428 | 2 |
| .486 | .4807703 | .0252084 | -.0052297 | .0252357 | 3 |
| .486 | .5448548 | .0776923 | .0588548 | .0811562 | 4 |

a5

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.855 | 1.764744 | .0155603 | -.0902559 | .0237064 | 1 |

```
 1.855   1.782802   .0152097   -.0721985   .0204223        2
 1.855   1.859954   .0658426    .0049536   .0658671        3
 1.855   1.877864   .0473187    .0228637   .0478415        4
------------------------------------------------------------
b2
------------------------------------------------------------
  true       mean        var        bias        mse   method
------------------------------------------------------------
 1.081   1.249242   .0826239    .1682419   .1109292        1
 1.081   1.078763   .0880706   -.0022368   .0880756        2
 1.081   1.087092   .1074391    .0060923   .1074762        3
 1.081   1.093707   .0964996    .0127068   .096661         4
------------------------------------------------------------
b3
------------------------------------------------------------
  true       mean        var        bias        mse   method
------------------------------------------------------------
 1.708   1.624893    .254648   -.0831066   .2615549        1
 1.708   2.925656   6.339429   1.217656   7.822116        2
 1.708   2.918664   5.532623   1.210664   6.998331        3
 1.708   2.699878   3.745051    .9918776  4.728872        4
------------------------------------------------------------
b4
------------------------------------------------------------
  true       mean        var       bias        mse   method
------------------------------------------------------------
  .765   .9555312   .0851766   .1905312   .1214787        1
  .765   .8070332   .0658319   .0420332   .0675987        2
  .765   .7741678   .0852941   .0091678   .0853782        3
  .765   .7708702   .0855312   .0058702   .0855657        4
------------------------------------------------------------
b5
------------------------------------------------------------
  true       mean        var        bias        mse   method
------------------------------------------------------------
  .736   .7169084   .0486882   -.0190916   .0490527        1
  .736   .6981637   .0555235   -.0378363   .0569551        2
  .736   .6775696   .0740213   -.0584304   .0774354        3
  .736   .6683916   .0786385   -.0676084   .0832094        4
------------------------------------------------------------
sex
------------------------------------------------------------
  true       mean        var        bias        mse   method
------------------------------------------------------------
```

| | | | | | |
|---|---|---|---|---|---|
| .5 | .4805603 | .0106573 | -.0194397 | .0110352 | 1 |
| .5 | .4998294 | .0102528 | -.0001705 | .0102528 | 2 |
| .5 | .5023953 | .0103461 | .0023954 | .0103518 | 3 |
| .5 | .5031638 | .010259 | .0031638 | .010269 | 4 |

### 10.3 Appendix 3

2PL

a1

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.856 | 1.897766 | .0357913 | .0417658 | .0375357 | 1 |
| 1.856 | 1.803208 | .0198668 | -.0527918 | .0226538 | 2 |
| 1.856 | 2.213829 | .2187294 | .3578288 | .3467709 | 3 |
| 1.856 | 2.237906 | .1969852 | .3819056 | .3428371 | 4 |

a2

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .808 | .8671438 | .0138701 | .0591437 | .0173681 | 1 |
| .808 | .7649817 | .010811 | -.0430182 | .0126616 | 2 |
| .808 | 1.408881 | .8860177 | .6008813 | 1.247076 | 3 |
| .808 | 1.419037 | .8519055 | .6110365 | 1.225271 | 4 |

a3

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.805 | 2.187113 | .1586735 | .3821126 | .3046836 | 1 |
| 1.805 | 1.884716 | .4577178 | .079716 | .4640725 | 2 |
| 1.805 | 2.987931 | 2.050764 | 1.182932 | 3.450091 | 3 |
| 1.805 | 3.071312 | 1.969582 | 1.266312 | 3.573128 | 4 |

a4

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .486 | .4721279 | .0070887 | -.0138721 | .0072812 | 1 |
| .486 | .4302703 | .006448 | -.0557297 | .0095538 | 2 |
| .486 | .7085304 | .1357044 | .2225304 | .1852242 | 3 |
| .486 | .7264104 | .1386981 | .2404104 | .1964953 | 4 |

a5

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|

```
-------------------------------------------------------------
 1.855   1.833092   .020532   - .0219083   .021012         1
 1.855   1.791673   .0165428   -.0633268   .020553         2
 1.855   2.078465   .181674     .2234651   .2316107        3
 1.855   2.093845   .1652103    .2388454   .2222574        4
-------------------------------------------------------------
b2
-------------------------------------------------------------
  true       mean        var       bias        mse   method
-------------------------------------------------------------
 1.081   1.156328   .0749296   .0753282    .080604         1
 1.081   1.154783   .0519436   .0737825    .0573874        2
 1.081   1.257599   .2055549   .1765995    .236742         3
 1.081   1.257975   .2250428   .1769753    .256363         4
-------------------------------------------------------------
b3
-------------------------------------------------------------
  true       mean        var       bias        mse   method
-------------------------------------------------------------
 1.708   2.069697   .5235744   .3616973    .6543993        1
 1.708   1.77273    .6185696   .0647297    .6227596        2
 1.708   2.338037  3.068166    .6300371   3.465113         3
 1.708   2.398904  2.890204    .6909037   3.367552         4
-------------------------------------------------------------
b4
-------------------------------------------------------------
  true       mean        var       bias        mse   method
-------------------------------------------------------------
  .765   .8177502   .0433205   .0527502    .046103         1
  .765   .8646688   .0376637   .0996688    .0475976        2
  .765   .8137444   .0483867   .0487444    .0507627        3
  .765   .8107855   .0551572   .0457855    .0572535        4
-------------------------------------------------------------
b5
-------------------------------------------------------------
  true       mean        var       bias        mse   method
-------------------------------------------------------------
  .736   .7493172   .0309768   .0133172   .0311541         1
  .736   .788505    .0248577   .052505    .0276145         2
  .736   .7499278   .0377445   .0139278   .0379385         3
  .736   .7476541   .0383285   .0116541   .0384643         4
-------------------------------------------------------------
sex
-------------------------------------------------------------
```

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .5 | .4356317 | .011277 | -.0643683 | .0154203 | 1 |
| .5 | .4289001 | .0107968 | -.0710999 | .015852 | 2 |
| .5 | .4397788 | .0115634 | -.0602212 | .01519 | 3 |
| .5 | .4391879 | .0115866 | -.060812 | .0152847 | 4 |

Rasch

a1

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.856 | 1.865856 | .0136965 | .0098563 | .0137937 | 1 |
| 1.856 | 1.891662 | .024972 | .0356622 | .0262438 | 2 |
| 1.856 | 1.806745 | .0131359 | -.0492554 | .015562 | 3 |
| 1.856 | 1.826948 | .0227875 | -.0290516 | .0236315 | 4 |

a2

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .808 | .805858 | .0104353 | .002142 | .0104399 | 1 |
| .808 | .8534808 | .0144759 | .0454808 | .0165444 | 2 |
| .808 | .7496139 | .0101061 | -.0583861 | .013515 | 3 |
| .808 | .775199 | .0119635 | .032801 | .0130394 | 4 |

a3

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.805 | 1.546992 | .0118292 | -.2580077 | .0783971 | 1 |
| 1.805 | 1.928696 | .0593698 | .1236963 | .0746706 | 2 |
| 1.805 | 1.489438 | .0115499 | -.3155617 | .1111291 | 3 |
| 1.805 | 1.812433 | .3179271 | .0074335 | .3179824 | 4 |

a4

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .486 | .4775429 | .008586 | -.0084571 | .0086575 | 1 |
| .486 | .4502009 | .0083502 | -.0357992 | .0096317 | 2 |

```
  .486    .4236207    .0083655    -.0623793    .0122567         3
  .486    .4179915    .0077666    -.0680085    .0123918         4
-------------------------------------------------------------------
a5
-------------------------------------------------------------------
  true       mean        var        bias        mse    method
-------------------------------------------------------------------
 1.855     1.95069    .014558      .0956899    .0237146         1
 1.855    1.829776    .0180521    -.0252236    .0186883         2
 1.855    1.890922    .0135707     .035922     .0148611         3
 1.855    1.796867    .0165975    -.0581327    .019977          4
-------------------------------------------------------------------
sex
-------------------------------------------------------------------
  true       mean        var        bias        mse    method
-------------------------------------------------------------------
  .5      .4957604    .0117366    -.0042396    .0117546         1
  .5      .4683312    .0118587    -.0316688    .0128616         2
  .5      .483654     .0116782    -.016346     .0119454         3
  .5      .4588941    .0116993    -.0411059    .013389          4
+-------------------------------------------------------------------
```

## 10.4 Appendix 4

a1

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.856 | 1.863672 | .014334 | .0076721 | .0143929 | 1 |
| 1.856 | 1.87568 | .0236711 | .01968 | .0240584 | 2 |
| 1.856 | 1.806553 | .0135525 | -.0494469 | .0159975 | 3 |
| 1.856 | 1.813333 | .0199039 | -.042667 | .0217244 | 4 |

a2

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .808 | .8069522 | .0095437 | -.0010478 | .0095448 | 1 |
| .808 | .8562235 | .0131396 | .0482235 | .0154651 | 2 |
| .808 | .7531188 | .0088907 | -.0548813 | .0119027 | 3 |
| .808 | .7802707 | .0107877 | -.0277293 | .0115566 | 4 |

a3

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.805 | 1.546377 | .0123112 | -.2586232 | .0791971 | 1 |
| 1.805 | 1.947467 | .064723 | .1424667 | .0850198 | 2 |
| 1.805 | 1.491041 | .0116334 | -.3139594 | .1102039 | 3 |
| 1.805 | 1.834272 | .3392618 | .0292724 | .3401187 | 4 |

a4

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| .486 | .474446 | .0081611 | -.011554 | .0082946 | 1 |
| .486 | .4476588 | .0077127 | -.0383412 | .0091827 | 2 |
| .486 | .4228434 | .0079871 | -.0631566 | .0119758 | 3 |
| .486 | .4164976 | .007285 | -.0695024 | .0121156 | 4 |

a5

| true | mean | var | bias | mse | method |
|---|---|---|---|---|---|
| 1.855 | 1.955112 | .0138027 | .1001116 | .023825 | 1 |

```
 1.855   1.832599   .0191107   -.022401    .0196125        2
 1.855   1.89728    .0120341    .0422799   .0138217        3
 1.855   1.799811   .0159599   -.055189    .0190058        4
------------------------------------------------------------
sex
------------------------------------------------------------
  true       mean        var        bias        mse   method
------------------------------------------------------------
  .5     .4942074   .0110663   -.0057926   .0110998        1
  .5     .4660435   .0106952   -.0339565   .0118482        2
  .5     .4811333   .0107297   -.0188667   .0110856        3
  .5     .455618    .0102575   -.044382    .0122273        4
------------------------------------------------------------
b2
------------------------------------------------------------
  true       mean        var       bias        mse   method
------------------------------------------------------------
 1.081   1.153314   .0803289   .0723145   .0855583        2
 1.081   1.128607   .0667528   .0476067   .0690192        4
------------------------------------------------------------
b3
------------------------------------------------------------
  true       mean        var        bias        mse   method
------------------------------------------------------------
 1.708   1.723483   .206219    .0154829   .2064587        2
 1.708   1.627197   .4659484  -.0808031   .4724775        4
------------------------------------------------------------
b4
------------------------------------------------------------
  true       mean        var       bias        mse   method
------------------------------------------------------------
  .765    .8230539  .0460811   .0580539   .0494514        2
  .765    .8511967  .045627    .0861967   .0530569        4
------------------------------------------------------------
b5
------------------------------------------------------------
  true       mean        var       bias        mse   method
------------------------------------------------------------
  .736   .7660906   .0578318   .0300905   .0587373        2
  .736   .7879062   .0549187   .0519062   .057613         4
------------------------------------------------------------
```

## 10.5 Appendix 5

```
2PL results

bias, 2pl 14, 8 and 2 mass-points

         14         8         2
a1  :  0.033414  0.033322  0.080446
a2  :  0.043530  0.043911  0.034828
a3  :  0.124807  0.119870  0.603601
a4  : -0.042018 -0.041902 -0.060533
a5  : -0.030776 -0.030695 -0.031645
sex : -0.021058 -0.021165 -0.024581

bias, 2pl 14, 8 and 2 mass-points

b2 :  0.032812  0.033957 -0.063162
b3 : -0.084749 -0.089597  0.320068
b4 :  0.023983  0.024500 -0.058353
b5 :  0.012678  0.012940  0.012572

variance, 2pl 14, 8 and 2 mass-points

a1  :  0.026335  0.026262  0.077207
a2  :  0.013923  0.013980  0.017864
a3  :  0.059088  0.056036  0.571158
a4  :  0.008910  0.008917  0.008755
a5  :  0.015727  0.015725  0.022104
sex :  0.010106  0.010097  0.010224

variance, 2pl 14, 8 and 2 mass-points

b2 :  0.037430  0.037805  0.033342
b3 :  0.089990  0.086639  0.620969
b4 :  0.023341  0.023358  0.018555
b5 :  0.027066  0.027071  0.044051

mse, 2pl 14, 8 and 2 mass-points

a1  :  0.027451  0.026262  0.077207
a2  :  0.015818  0.013980  0.017864
a3  :  0.074664  0.056036  0.571158
a4  :  0.010676  0.008917  0.008755
a5  :  0.016674  0.015725  0.022104
```

```
sex :  0.010549  0.010545  0.010828

mse, 2pl 14, 8 and 2 mass-points

b2 :  0.037430  0.037805  0.033342
b3 :  0.089990  0.086639  0.620969
b4 :  0.023341  0.023358  0.018555
b5 :  0.027066  0.027071  0.044051

 average deviance, 2pl 14, 8 and 2 mass-points

       4899.85   4899.85   4912.11

Rasch results

bias, 1pl 14, 8 and 2 mass-points

a1  :  0.002737  0.002863 -0.030659
a2  : -0.011192 -0.011078 -0.018205
a3  : -0.262870 -0.262746 -0.285270
a4  : -0.010838 -0.010728 -0.017587
a5  :  0.087273  0.087397  0.050652
sex :  0.009349  0.009275  0.003893
sig :  0.007636  0.007966 -0.055632

variance, 1pl 14, 8 and 2 mass-points

a1  :  0.013727  0.013736  0.013240
a2  :  0.009543  0.009547  0.009904
a3  :  0.011976  0.011984  0.011494
a4  :  0.008779  0.008784  0.009016
a5  :  0.015500  0.015506  0.014470
sex :  0.010153  0.010147  0.010659
sig :  0.004546  0.004553  0.003468

mse, 1pl 14, 8 and 2 mass-points

a1  :  0.013734  0.013736  0.013240
a2  :  0.009668  0.009547  0.009904
a3  :  0.081077  0.011984  0.011494
a4  :  0.008897  0.008784  0.009016
a5  :  0.023117  0.015506  0.014470
sex :  0.010240  0.010233  0.010674
sig :  0.004604  0.004617  0.006563
```

```
average deviance, 1pl 14, 8 and 2 mass-points
       4917.41   4917.40   4928.63
```

## 10.6 Appendix 6

We give first the 14-point analysis.

```
2PL results

 intercept bias, GQ and semiNP
 a1 : -0.262075 -0.065562
 a2 : -0.208361 -0.096605
 a3 : -0.468120 -0.012189
 a4 : -0.100347 -0.064001
 a5 : -0.148244 -0.045800
sex :  0.000619  0.007070

slope bias, GQ and semiNP
 b2 :   1.03658  0.428312
 b3 :   1.61056  0.825217
 b4 :  0.743416  0.310333
 b5 : -0.884245  0.296850

 intercept variance, GQ and semiNP
 a1 :  0.024555  0.074133
 a2 :  0.019166  0.062405
 a3 :  0.060144  0.450143
 a4 :  0.014411  0.025602
 a5 :  0.021912  0.039863
sex :  0.019034  0.017891

slope variance, GQ and semiNP
 b2 :  0.046767  0.046265
 b3 :  0.267254  0.467358
 b4 :  0.024156  0.015636
 b5 :  0.021840  0.022293

intercept mse, GQ and semiNP
 a1 :  0.093238  0.074133
 a2 :  0.062580  0.062405
 a3 :  0.279280  0.450143
 a4 :  0.024481  0.025602
 a5 :  0.043889  0.039863
sex :  0.019034  0.017941

slope mse, GQ and semiNP
 b2 :  0.046767  0.046265
 b3 :  0.267254  0.467358
```

```
  b4 :  0.024156  0.015636
  b5 :  0.021840  0.022293

 average deviance, GQ and semiNP
     5212.44   5193.96

 masses
0.00005  0.00125  0.01235  0.05710  0.12984  0.15278  0.11317
0.14491  0.21433  0.16057  0.01145  0.00131  0.00054  0.00035

  Rasch results

  intercept bias, GQ and semiNP
  a1 : -0.179093 -0.113185
  a2 : -0.264608 -0.190963
  a3 : -0.992532 -0.911083
  a4 : -0.011842  0.058847
  a5 :  0.238290  0.277580
 sex : -0.023618 -0.015210
sigma :  0.846994  0.336884

 intercept variance, GQ and semiNP
  a1 :   9.80449   10.6023
  a2 :   1.04369   1.35509
  a3 :   2.31338   2.81001
  a4 :  0.798384   1.06375
  a5 :   15.2705   15.8598
 sex :  0.810633  0.835670
sigma :  0.006762   6.23339

 intercept mse, GQ and semiNP
  a1 :   9.83657   10.6023
  a2 :   1.11371   1.35509
  a3 :   3.29850   2.81001
  a4 :  0.798524   1.06375
  a5 :   15.3272   15.8598
  sex :  0.811191  0.835901
  sigma :  0.724161   6.34688

 average deviance, GQ and semiNP
     5256.08   5230.06
  masses
0.00003  0.00086  0.00933  0.04970  0.13260  0.17709  0.12170
0.11245  0.19587  0.19197  0.00677  0.00083  0.00055  0.00025
```

The five-point analysis now follows, for the 2PL model only.

```
 2PL results

 intercept bias, GQ and semiNP
  a1 : -0.228522  0.276934
  a2 : -0.188995  0.261841
  a3 : -0.671274  0.426367
  a4 : -0.056198  0.214529
  a5 : -0.097007  0.224486
  sex : -0.011182 -0.002297

 slope bias, GQ and semiNP
  b2 :  0.131994  0.000991
  b3 : -0.040896 -0.001131
  b4 :  0.131747  0.019225
  b5 : -0.833007  0.016767

 intercept variance, GQ and semiNP
  a1 :  0.028916  0.095974
  a2 :  0.020948  0.067590
  a3 :  0.035852  0.439245
  a4 :  0.015751  0.037833
  a5 :  0.023433  0.053488
  sex :  0.021213  0.016772

 slope variance, GQ and semiNP
  b2 :  0.028568  0.027011
  b3 :  0.067968  0.162708
  b4 :  0.014771  0.014573
  b5 :  0.012373  0.017083

 intercept mse, GQ and semiNP
  a1 :  0.081138  0.095974
  a2 :  0.056667  0.067590
  a3 :  0.486461  0.439245
  a4 :  0.018909  0.037833
  a5 :  0.032843  0.053488
  sex :  0.021338  0.016777

slope mse, GQ and semiNP
  b2 :  0.028568  0.027011
  b3 :  0.067968  0.162708
```

```
b4 :  0.014771  0.014573
b5 :  0.012373  0.017083

average deviance, GQ and semiNP
   5233.88   5193.83
masses
0.07535  0.27514  0.27124  0.35111  0.02716
```

## 10.7 Appendix 7

sex

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| -.4717774 | -.4003795 | .1012061 | .0087869 | .0713979 | .0138846 | 1 |
| -.4717774 | -.3923196 | .1003296 | .0085764 | .0794578 | .0148899 | 2 |
| -.4717774 | -.4425962 | .1115618 | .0124346 | .0291811 | .0132861 | 3 |
| -.4717774 | -.424494 | .0872816 | .0098588 | .0472833 | .0120945 | 4 |

B-W

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| -2.358886 | -2.103144 | .1679237 | .0213791 | .2557421 | .0867832 | 1 |
| -2.358886 | -2.065129 | .1672854 | .0198332 | .2937571 | .1061264 | 2 |
| -2.358886 | -2.338026 | .1865937 | .0256824 | .0208606 | .0261176 | 3 |
| -2.358886 | -2.24388 | .1485837 | .0268711 | .1150061 | .0400975 | 4 |

H-W

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| -1.887109 | -1.722895 | .1648262 | .0307997 | .164214 | .0577659 | 1 |
| -1.887109 | -1.693074 | .1637478 | .0315966 | .1940346 | .069246 | 2 |
| -1.887109 | -1.932474 | .1806926 | .0351035 | -.0453653 | .0371615 | 3 |
| -1.887109 | -1.843214 | .1458032 | .0319288 | .0438946 | .0338555 | 4 |

A-W

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| .9435547 | .8625736 | .2120519 | .0470132 | -.080981 | .0535711 | 1 |
| .9435547 | .8414024 | .2115546 | .046944 | -.1021523 | .0573791 | 2 |
| .9435547 | .9524695 | .2418557 | .0518172 | .0089148 | .0518967 | 3 |
| .9435547 | .9192684 | .1887529 | .0499044 | -.0242863 | .0504942 | 4 |

## 10.8 Appendix 8

sex

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| -.4717774 | -.4225371 | .0971696 | .0112302 | .0492403 | .0136548 | 1 |
| -.4717774 | -.457999 | .1036119 | .0122095 | .0137783 | .0123994 | 2 |
| -.4717774 | -.4448327 | .0847814 | .0111525 | .0269446 | .0118785 | 3 |
| -.4717774 | .0341042 | .041424 | .0035886 | .5058815 | .2595047 | 4 |
| -.4717774 | -.1748484 | .0415408 | .0043103 | .2969289 | .0924771 | 5 |

B-W

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| -2.358886 | -2.184784 | .1643785 | .0238123 | .1741026 | .054124 | 1 |
| -2.358886 | -2.33144 | .1751553 | .0281738 | .0274459 | .0289271 | 2 |
| -2.358886 | -2.296649 | .1456723 | .0299205 | .0622377 | .0337941 | 3 |
| -2.358886 | -1.523054 | .0956687 | .0272868 | .8358321 | .725902 | 4 |
| -2.358886 | -1.633922 | .0986505 | .0209516 | .7249638 | .5465241 | 5 |

H-W

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| -1.887109 | -1.796543 | .1580272 | .0319451 | .090566 | .0401473 | 1 |
| -1.887109 | -1.929989 | .1669099 | .0439596 | -.0428798 | .0457983 | 2 |
| -1.887109 | -1.882489 | .1415066 | .0390862 | .00462 | .0391076 | 3 |
| -1.887109 | -1.188327 | .0881764 | .0263771 | .6987814 | .5146726 | 4 |
| -1.887109 | -1.287025 | .0905349 | .0250968 | .6000835 | .385197 | 5 |

A-W

| true | mean | se | var | bias | mse | method |
|---|---|---|---|---|---|---|
| .9435547 | .8488848 | .2049045 | .0407052 | -.0946699 | .0496676 | 1 |
| .9435547 | .9035459 | .2234571 | .0469223 | -.0400088 | .048523 | 2 |
| .9435547 | .8688834 | .1836426 | .0463979 | -.0746714 | .0519737 | 3 |
| .9435547 | .9096158 | .115642 | .032179 | -.033939 | .0333308 | 4 |
| .9435547 | .7390021 | .1118233 | .0252872 | -.2045526 | .0671289 | 5 |