Investigation of the identifiability of the 3PL model in the NAEP 1986 Math survey

November 7, 2006

Prepared by:

Murray Aitkin and Irit Aitkin
School of Behavioural Science
University of Melbourne

Prepared for:

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

This project was an activity of the Federal Statistics Program

Contents

1 Aim of the project

It is widely known, or believed, that the 3PL model is difficult, if not impossible, to identify in test item analysis. The aim of this project was to assess the identifiability of the 3PL model in simulations from known data structures, and in data analysis of the Numbers and Operations – Knowledge and Skills subscale for the NAEP 1986 Math test.

2 Summary

- In our simulations with 10 items taken by all 1000 examinees, with responses generated from five items with 2PL models and five with 3PL models, fitting the model with the *correct known guessing parameters* gave the highest maximized log-likelihood, and unbiased estimates of the student-level reporting group parameters.
- Ignoring the 3PL model structure by fitting the 2PL model to *all* items gave biased estimates of the student-level reporting group parameters, and a much lower maximized log-likelihood.
- Fitting the 3PL model to all items with all the guessing parameters set to *incorrect* known values, also resulted in biased student-level parameter estimates and a much lower maximized log-likelihood, close to that for the 2PL model for all items.
- Fitting the 3PL model with known guessing parameter values in a fine grid around the true values showed that these parameters should be identifiable.
- For the Numbers and Operations – Knowledge and Skills subscale of the 1986 NAEP Math survey, ML estimation in Stata with "known" guessing parameter values (from the ETS report) failed, probably because of the small number of items answered by most students.
- Successful maximum likelihood (ML) and standard error estimation of all item parameters, including the guessing parameters, was reported by Garcia-Perez (1999), in a series of simulations with 50 3PL items attempted by 500 examinees.

We concluded that 3PL models may be *identifiable*, and all parameters estimated with sufficiently rich data, but for the very sparse NAEP math data we examined, this estimation failed even with "known" guessing parameters. This raises a serious issue of bias in reporting group parameter estimation. We considered the current estimation methods for the guessing parameters and examined alternative guessing models, and suggest that these be further investigated.

3 Background

The 2PL model for the probability p_{ij} of examinee i with ability θ_i answering correctly item j is given by

$$\text{logit}\, p_{ij} = a_j(\theta_i - b_j),$$

where b_j is the difficulty of item j and a_j is its discrimination. We define $\phi_{ij} = a_j(\theta_i - b_j)$. The 3PL model (due to Birnbaum, in Lord and Novick 1968) extends the 2PL model to allow for random guessing by examinees. For the 3PL model the correct answer probability is given as

$$p_{ij} = c_j + (1 - c_j)e^{\phi_{ij}}/(1 + e^{\phi_{ij}}),$$

where c_j is the probability of a correct guess (formally, for an examinee with ability $-\infty$).

There are several descriptions of the basis for this model for a multiple-choice item (Birnbaum, Hutchinson 1991, San Martin, del Pino and De Boeck 2006). The most plausible is that the examinee, when presented with the item, first works on it, with success probability given by the 2PL model above. If the answer found is one of the response categories, that response is given. If it is not, the examinee makes a guess, with a success probability c_j.

This interpretation leads to the overall success probability

$$\begin{aligned} p_{ij} &= e^{\phi_{ij}}/(1 + e^{\phi_{ij}}) + [1 - e^{\phi_{ij}}/(1 + e^{\phi_{ij}})] \cdot c_j \\ &= e^{\phi_{ij}}/(1 + e^{\phi_{ij}}) + 1/(1 + e^{\phi_{ij}}) \cdot c_j \\ &= c_j + (1 - c_j)e^{\phi_{ij}}/(1 + e^{\phi_{ij}}). \end{aligned}$$

Birnbaum (pp. 404-5 in Lord and Novick 1968; we replace Birnbaum's probit by the logit) described this as a

"... highly schematized psychological hypothesis. .. if an examinee has ability θ, then the probability that he will *know* the correct answer is given by [the logit function]

$$G[a_j(\theta - b_j)] = e^{a_j(\theta - b_j)}/(1 + e^{a_j(\theta - b_j)})$$

" ... if he does not know it he will guess, and with probability c_j will guess correctly. It follows from these assumptions that the probability of an incorrect response is

$$\{1 - G[a_j(\theta - b_j)]\}(1 - c_j),$$

and that the probability of a correct response is the item characteristic curve

$$P_j(\theta) = c_j + (1 - c_j)G[a_j(\theta - b_j)]."$$

Whatever the strength of this justification, the introduction of the third parameter greatly weakens the information about examinee ability, since correct answers may be due to guessing, in which case they give no information about examinee ability except that it may be low. Since a guess on an item can only be poorly identified by an inconsistent response pattern across items, the difficulty of estimation of the item parameters is greatly increased.

The difficulties of maximum likelihood estimation in this model are well known and frequently mentioned in the literature. For example, van der Linden and Hambleton (1997 p. 16) point out that the estimates of the a_j and c_j are "sensitive to minor fluctuations in the responses used to produce these estimates. Unless huge samples of [examinees] or tight priors around the true parameter values are used, the estimates are unstable."

The simulation study of Garcia-Perez (1999) showed that for responses generated from 50 items with 3PL models and a sample of 500 examinees, there was no reported difficulty in estimating the 3PL model (for the range of parameters considered); evidently the examinee abilities are evaluated sufficiently well from this number of items to identify the guessing parameters.

In his study, fitting the 2PL model to all items gave slightly downwardly biased estimates of the item difficulties b_j, but the estimates of the item discriminations a_j were severely biased downwards, as would be expected – the slope must decrease to allow for the greater (0,1) range of the 2PL response probabilities.

Skrondal and Rabe-Hesketh (2004, pp. 292-298) reported a 3PL analysis of a small data set of four items (the Arithmetic reasoning data used by Mislevy 1985), with all 16 response patterns observed. They fitted the 3PL model to all four items by setting equal all four guessing parameters (the ML procedure did not converge with separate guessing parameters for each item), and running over a grid of known values to generate the *profile likelihood* in the single guessing parameter – the likelihood maximized over the other parameters for given values of the guessing parameter.

This process gave an ML estimate for the common guessing parameter, but the other parameter estimates were unstable with some very high parameter intercorrelations, suggesting that with this small number of items, even a single common guessing parameter was nearly unidentifiable. We comment further below on aspects of the model and ML analysis.

Garcia-Perez's simulations did not use explanatory variables; since we are concerned with the upper-level parameters we carried out simulations to assess the possible biases in these parameters.

4 Simulation structures

Data were simulated from a simple model structure similar to that used in previous simulations: a 10-item test with an explanatory variable structure of ethnic group (4 levels), sex(2) homework(3) and poverty (2). Five of the items had 2PL regressions on student ability, the other five had 3PL regressions. All the item parameters were similar to those of 10 of the items from the Numbers and Operations – Knowledge and Skills subscale of the 1986 NAEP math test. The parameter values are given in Table 1 of the Appendix. The non-zero guessing parameters ranged from 0.208 to 0.352.

4.1 First simulation series

Samples of size 1000 were drawn from the $N(0, 1)$ ability distribution and used to generate item responses from the model. For each sample three models were fitted to the item data:

- the correct model for all items, with the true guessing parameters for the 3PL items;
- the 2PL model for all items, setting the 3PL item guessing parameters to zero;
- the 3PL model for all items, with all guessing parameters set to 0.2.

Model fitting was done in Stata using Gllamm. The 3PL model was fitted using a composite link function, following the description in Skrondal and Rabe-Hesketh (2004, p. 294). This allowed different *known* values to be set for each guessing parameter.

Table 2 in Appendix 1 gives, for each of the three models fitted, the mean, bias, mean square error (mse), standard deviation across simulations (sdb) and the average across simulations of the reported standard error (se) for each of the reporting group parameters. These values are computed over 332 samples.

4.2 Conclusions from the first simulations

The correct 3PL model gave by far the best maximized log-likelihood over all samples. The 2PL model and the 3PL model with incorrect guessing parameters were almost equally bad, and significantly worse than the correct model in maximized log-likelihood.

For estimates, we present only those for the reporting group parameters, since biases in the item parameters would not be of concern if the reporting group parameter estimates were unbiased.

Standard errors were smallest for the 2PL model and largest for the incorrect 3PL model. Standard deviations of parameter estimates across

simulations agreed well with the average standard errors, showing that the latter were adequate representations of the sampling variability in the estimates.

Biases of the 2PL estimates were severe for the large parameters – as much as 8 SEs for the largest ethnic group parameter (ethnic2). For the small parameters the biases were small.

For the 3PL model with incorrect guessing parameters, the biases were substantially less, and of opposite sign to those from the 2PL model, but were still of concern for the largest parameters – more than 2 SEs for ethnic2.

Mean square errors for the 2PL estimates were very much larger than for the correct 3PL model, up to 17 and 14 times for the two largest ethnic group parameters. For the smallest homework parameter the 2PL estimates had smaller mean square error than those from the correct 3PL model, because their larger bias was outweighed by their smaller variance.

For the incorrect 3PL model the mean square errors for the largest parameters were 4 to 7 time those for the correct model.

Thus the effect of negelecting the 3PL structure of some of the items – whether by fitting the 2PL model or by mis-specifying the guessing parameter – was to produce serious biases in the largest model parameters, especially serious for the 2PL model fitting which ignored guessing.

This is a very serious consequence of the use of the 3PL model, and we conjecture that it is due to the inability to identify the discrimination parameter because of the range restriction mentioned above. We comment further below.

4.3 Second simulation series

To assess the sensitivity of the reporting group parameter estimates to mis-specification of the guessing parameters, we repeated the simulations above *for a single sample*, but over a 21-point grid of guessing parameter values, in which the five non-zero guessing parameters were all changed in steps of 0.01, above and below their true values. Thus the guessing parameters used varied from all 0.1 below their true values, to all 0.1 above their true values. The grid steps d, log-likelihood, parameter estimates and standard errors are given in Table 3, where d is the deviation from the true value.

The magnitudes of both parameter estimates and their standard errors increased steadily with increasing values of the guessing parameters. The log-likelihood was maximized at values 0.04 above the true guessing parameters, but the asymptotic 95% confidence interval (defined by a log-likelihood difference of 2.0 from the maximum) just included the true values of the guessing parameters ($d = 0$ in Table 3).

Thus it *should* be possible to maximize the likelihood in the guessing parameters, given a sufficiently careful algorithm.

5 3PL analysis of NAEP math data

We analyzed the 1986 NAEP math data on the Numbers and Operations – Knowledge and Skills subscale for Age 9/Grade 3 students, as described in previous reports. There were 12 items out of the 30 items on the subscale for which 3PL models had been used in the original reported NAEP analysis. Since we expected identification difficulties we did not attempt to estimate the guessing parameters individually for these items, but set them to known values equal to their estimates reported in the NAEP document for this survey.

The analysis was aimed at replicating the comparison in the simulations between the correct analysis, the incorrect guessing analysis and the analysis using only the 2PL model for all items. We restricted the model to three levels, ignoring the PSU sampling level; as shown in a previous report (Aitkin and Aitkin 2005), the variance component at this level was very small.

The Gllamm analysis was set up as described above in Skrondal and Rabe-Hesketh, and as used in the simulations, with a composite link function to handle the two different relations between item responses and ability.

Both the 3PL runs failed to converge, with a flat likelihood indicating unidentifiability of one or more model parameters. Since there were no guessing parameters to estimate in this model, we concluded that the problem was probably in the item discriminations: as described above the reduced range of response probabilities for the 3PL items increases the difficulty of estimating the item slopes, and provides less information in the response about examinee ability, specially with the very sparse item data from individual examinees.

Examination of the item assignment for the scale (see Appendix Tables 4 and 5) suggested one reason why the analysis might fail. The BIB spiralling assignment of items to booklets gives a fairly even spread of items across examinees: the number of examinees answering each item on the scale ranged from 1534 to 2734, with a mean of 2437. However the number of items answered by each examinee varied from 1 to 25, with median 7 and interquartile range 2 to 10. 31% of the 10,465 examinees who had *any* items on this scale had only 1 or 2 items in their booklets. So if the ability of each examinee on this subscale were *known*, the sample size for each item would be more than sufficient to estimate all its parameters, but there may not be enough *items for each examinee* to estimate their latent abilities sufficiently accurately for this purpose.

Another possibility is that the reported guessing parameters, which are determined by fitting a *null* model to the item data, are different from the ML estimates obtained by *jointly fitting* both items and reporting group variables. This may explain how the priors for the guessing parameters in the null model are able to determine the item parameters, while constraining the guessing parameters to their "true" values in the joint model (equivalent

to a terminally tight prior) does not.

A *single*-dimension ability model fitted to *all* the students on the full Math test might not have this problem as many more items would be available for analysis for each student.

We assessed Latent GOLD as an alternative package for the 3PL analysis, since it can handle three levels of nesting. However the program could not be set up to fit this model, as it could fit only a mixture of normals distribution to a Rasch model for the latent ability.

6 Conclusions from the data analysis and simulations

Despite the widespread assumption that the 3PL model is unidentifiable, or nearly so, it appears from Garcia-Perez's simulations and our results that the model *can* be identified given sufficient items answered by each examinee. A large number of examinees with a small number of items does not guarantee identification.

The results of the simulations and the NAEP data analysis are frustrating. The simulations reported by Garcia-Perez show that biases in the item difficulty and discrimination parameter estimates occurred, and our simulations showed that serious biases in the upper-level reporting group parameter estimates could occur, if the true 3PL form of the item response function was ignored and the 2PL model used instead. Biases were much smaller, though still significant for the largest parameters, if an incorrect value for all the guessing parameters was used. Of course this begs the question of how to know what values are correct or incorrect.

Our simulations were based on small but dense data sets, of 10 items with 1000 examinees. The NAEP data set is very sparse, with a median of 7 items per student out of the 30, and more than 10,000 students. The larger number of students does not help us in estimating individual abilities – it is the number of items which provides this information. Even with *known* guessing parameters, the small numbers of items per examinee from the 30 NAEP subscale items do not provide enough information to estimate the other item parameters given the fixed values of the guessing parameters (unless these are set to zero with the 2PL model).

Since many items on many NAEP surveys use the 3PL model (as guessing is expected to occur), this raises a serious difficulty for the analysis of these items and the reporting group effects. This difficulty requires a closer look at the 3PL model. We now consider some aspects of this model.

7 Further investigation and criticism of the 3PL model

7.1 Mixture interpretation of the 3PL model

We want to consider a more general test item model. Its properties are easier to understand if we exemplify with a biological model closely related to the 3PL model – the *natural mortality* model – which is used in biometrics to assess the effectiveness of insecticides.

A test group of insects is randomly divided into $K+1$ groups of approximately equal sizes $n_i, i = 0, \ldots, K$, and K of these groups are treated with an insecticide at increasing dose levels $x_i, i = 1, \ldots, K$. The remaining group is a control group and is not treated. The object of analysis is to estimate the dose-mortality relationship, in order to set appropriate dose levels for the spraying of insect populations.

In the i-th group r_i insects die out of the n_i. The probability of death $(Y_i = 1)$ in the i-th dose group is modelled as

$$\Pr[Y_i = 1 \mid x_i] \quad = \quad c + (1-c)e^{\alpha+\beta x_i}/(1+e^{\alpha+\beta x_i}), \tag{1}$$

where c is the probability of death from natural mortality. The logic of this model is clear: without any insecticide treatment a proportion c of insects die naturally, so this has to be allowed for in modelling the effect of the treatment – the death probability increases monotonically with dose from c to 1.

7.2 Biological model extension

We now consider a slight extension of this model which is relevant to the interpretation of the 3PL model.

The insect population now consists of two types: Type 0 dies with probability p regardless of dose, while the Type 1 death probability increases with dose. The two types of insects are indistiguishable. The proportion of Type 0 in the population (and therefore in each dose group) is c. The death probability for Type 1 is the logistic function of dose with parameters α and β. The death probability for any insect given in (1) now has a formal *mixture model* representation:

$$\begin{aligned} \Pr[Y_i = 1 \mid x_i] &= \Pr[Type\,0]\Pr[Y_i = 1 \mid x_i, Type\,0] \\ &+ \Pr[Type\,1]\Pr[Y_i = 1 \mid x_i, Type\,1] \\ &= c \cdot p + (1-c) \cdot e^{\alpha+\beta x_i}/(1+e^{\alpha+\beta x_i}). \end{aligned}$$

The case $p = 1$ gives the dose model in (1): in this model, if we accept the mixture interpretation, *all Type 0 insects die in all the treatment and control groups.*

7.3 Extended psychometric model

The analogous psychometric model (*for item j*) is an extension of the 3PL model. In this extension there are two indistinguishable types of examinees for each item j: Type 0_j examinees guess with success probability p_j regardless of ability, and their proportion in the population is c_j. For Type 1_j examinees the correct answer probability has the 2PL model with parameters a_j and b_j, so the model is

$$p_{ij} = c_j \cdot p_j + (1 - c_j) \cdot e^{\phi_{ij}}/(1 + e^{\phi_{ij}}),$$

where

$$\phi_{ij} = a_j(\theta_i - b_j)$$

as before. The usual 3PL model results from setting $p_j = 1$ – *the "guessors" guess perfectly*!

This result looks quite unreasonable. Guessors are much more likely to have p_j around 1/4 or 1/5, or the reciprocal of the number of response categories in a multiple-choice item. No-one can guess correctly 100% of the time. Yet that is the implication of the 3PL model, if we interpret it in terms of two groups of examinees following different strategies.

The fitting of the standard 3PL model (with a known guessing parameter) described in Skrondal and Rabe-Hesketh implicitly uses the extended model above: the model fitted with a known guessing parameter of 0.1 is expressed as (p. 294 – their notation is slightly different)

$$p_{ij} = 0.1g_1^{-1}(1) + 0.9g_2^{-1}(\phi_{ij}),$$

where $g_1^{-1}(1) = 1$ and $g_2^{-1}(x) = e^x/(1 + e^x)$. This is equivalent to the extended 3PL model above, with $p_j = 1$.

Further unattractive features of this aspect of the 3PL model are that guessing is a *random process* in which all examinees participate, and there is no relation between those who guess on item j and those who guess on any other item.

It seems more reasonable to suppose that guessing itself depends explicitly on ability, as well as the item difficulty, and that the probability of guessing *decreases* with increasing ability, rather than being constant. We now consider a model for this process.

8 Ability-based guessing

The basis for the ability-based guessing models (there are several versions – see San Martin, del Pino and de Boeck 2006) is that:

- the probability of guessing increases with item difficulty, and decreases with ability;

- when guessing occurs, the probability of a correct answer may be item-specific, and is generally small.

We extend the logistic function, to both the probability of a correct response to an item based on ability, and the probability of guessing. The most general form of the model uses 2PL functions for both these probabilities. Define $\psi_{ij} = \alpha_j(\theta_i - \beta_j)$, where α_j is negative, and define the dummy indicator $Z_{ij} = 1$ if examinee i guesses on item j, $Z_{ij} = 0$ if he or she does not guess. Then the model for guessing is

$$\Pr[Z_{ij} = 1 \mid \theta_i] = e^{\psi}_{ij}/(1 + e^{\psi}_{ij}),$$

and the probability of a correct answer by examinee i on item j is

$$\begin{aligned} \Pr[Y_{ij} = 1 \mid \theta_i] &= \Pr[Y_{ij} = 1 \mid \theta_i, Z_{ij} = 1] \Pr[Z_{ij} = 1 \mid \theta_i] \\ &+ \Pr[Y_{ij} = 1 \mid \theta_i, Z_{ij} = 0] \Pr[Z_{ij} = 0 \mid \theta_i] \\ &= p_j \cdot e^{\psi_{ij}}/(1 + e^{\psi_{ij}}) + e^{\phi_{ij}}/(1 + e^{\phi_{ij}}) \cdot 1/(1 + e^{\psi_{ij}}). \end{aligned}$$

Here the probability of a correct guess p_j is unspecified, other than depending on the item j. Compared with the 3 parameters for each item for the 3PL model, this model has 5 – the two parameters in each logistic regression and the correct guess probability p_j. So many parameters will certainly cause identifiability problems; there are several ways to reduce the number of parameters:

- specify the guessing parameters p_j to be $1/D_j$ where D_j is the number of response categories for the multiple choice items (this gives one less parameter per item);
- relate the parameters in the guessing probability (ψ) model to those in the correct answer probability (ϕ) model – for example set $\beta_j = b_j$ (this gives one less parameter per item);
- specify equal item discriminations in one or both models, as in the Rasch model.

8.1 An example

San Martin et el (2006) introduce the *1PL-AG* model, in which the ability model is Rasch and the guessing model is 2PL, but with a common discrimination parameter across items. This has 2 parameters for each item, plus one additional discrimination parameter; if this common discrimination parameter is 1, the guessing model is also Rasch; if it is zero, the guessing model is random, independent of ability.

They report analyses of both simulated and real test data using this model. Model fitting was done with SAS NLMIXED using non-adaptive

Gaussian quadrature with 15 nodes and Newton-Raphson for optimization. The parameter estimates were recovered well in simulations, though over only four generated data sets because of heavy computing time.

The model was fitted to several replicate sub-samples of 2,000 examinees from the Chilean SIMCE tests in mathematics and language. Conclusions from the replicate samples were consistent, and showed that the model fitted well, with guessing occurring in the language test but not in the mathematics test.

9 Conclusions and recommendations

It appears that with the NAEP BIB spiralling item design, for scales with small numbers of items there may be insufficient items per examinee for reliable estimation of the guessing parameters for the 3PL items. With current software and the BIB design, an intensive search of the guessing parameter space appears to be needed to obtain maximum likelihood estimates, assuming this is possible.

Since fitting mis-specified 2PL models leads to biases in reporting group estimates, this is a serious matter for NAEP analysis. However, the 3PL model is not the only possible model for guessing, and is in any case not a *gold standard* for NAEP analysis. The difficulty is finding a suitable model which can incorporate guessing and is identifiable in NAEP scales with small numbers of items. It appears that the ability-based guessing model can be fitted to the NAEP item data in Gllamm, using the same composite link function approach as used for the 3PL model. With a sufficiently restricted parameter structure in this model, it may be possible to estimate a guessing model based on ability and item difficulty. Guessing parameters based on the number of response categories, or plausible distractors, could be used to simplify the model.

If this turns out not to be possible, it appears that the 3PL model should be abandoned, at least for scales with small numbers of items like the Knowledge and Skills subscale. If the individual scale results are not reported, it would be simplest to abandon the analyses of these scales, and reduce the overall computational load of the NAEP analysis by analyzing and reporting only a single math ability scale *for all items*; this might be able to support the estimation of guessing parameters.

We recommend

- that further investigation of the class of ability-based guessing models be made, with the object of assessing their suitability for NAEP item analysis;
- that the current estimation of guessing parameters be reviewed, for the role of the tight priors used and the methods by which estimates

of the guessing parameters are obtained;

- that consideration be given to the use of guessing parameters equal to the reciprocal of the number of items, or the number of plausible distractors;

- that consideration be given to abandoning analyses of scales using the 3PL model with small numbers of items per examinee, and analyzing only a single scale, or scales identified by large numbers of items.

10 References

Aitkin, M. and Aitkin, I. (2005) *Multi-level model analysis of the Knowledge and Skills scale of the NAEP 1986 math data (final report).* NCES report.

Garcia-Perez, M. (1999) Fitting logistic IRT models: small wonder. *The Spanish Journal of Psychology* **2**, 74-94.

Hutchinson, T.P. (1991) *Ability, partial information and guessing: statistical modelling applied to multiple-choice tests.* Rumsby Scientific Publishing, Rundle Mall, South Australia.

Mislevy, R.J. (1985) Estimation of latent group effects. *Journal of the American Statistical Association* **80**, 993-997.

San Martin, E., del Pino, G. and De Boeck, P. (2006) IRT models for ability-based guessing. *Applied Psychological Measurement* **30**, 183-203.

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

11 Appendix

Table 1 - true parameter values

ethnic2	ethnic3	ethnic4	hw2	hw3	poverty	sex
-2.359	-1.887	0.944	0.100	0.300	-0.800	-0.472

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
1.738	1.202	0.841	1.090	0.855	1.150	1.162	0.894	0.898	0.620

b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
-0.217	-0.359	0.540	0.789	0.735	0.339	0.985	0.858	-1.643	-1.159

c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
0	0	0	0.238	0	0.208	0	0.280	0.352	0.225

Table 2 - simulation results, first series

true	mean	bias	mse	sdb	se	param	method
-2.359	-1.526	0.833	0.708	0.119	0.114	ethnic2	2pl
-2.359	-2.464	-0.106	0.040	0.169	0.160	ethnic2	correct_
-2.359	-2.841	-0.482	0.293	0.247	0.224	ethnic2	wrong_gu
-1.887	-1.253	0.634	0.415	0.113	0.110	ethnic3	2pl
-1.887	-1.972	-0.085	0.032	0.158	0.150	ethnic3	correct_
-1.887	-2.181	-0.294	0.132	0.213	0.192	ethnic3	wrong_gu
0.944	0.715	-0.229	0.081	0.169	0.153	ethnic4	2pl
0.944	0.967	0.023	0.042	0.204	0.194	ethnic4	correct_
0.944	0.917	-0.027	0.044	0.207	0.197	ethnic4	wrong_gu
0.100	0.069	-0.031	0.007	0.076	0.078	hw2	2pl
0.100	0.103	0.003	0.010	0.099	0.103	hw2	correct_
0.100	0.103	0.003	0.011	0.106	0.112	hw2	wrong_gu
0.300	0.208	-0.092	0.020	0.109	0.098	hw3	2pl
0.300	0.311	0.011	0.019	0.137	0.129	hw3	correct_
0.300	0.313	0.013	0.023	0.151	0.139	hw3	wrong_gu
-0.800	-0.518	0.282	0.090	0.103	0.101	poverty	2pl
-0.800	-0.834	-0.034	0.020	0.137	0.138	poverty	correct_
-0.800	-0.892	-0.092	0.036	0.166	0.162	poverty	wrong_gu
-0.472	-0.323	0.149	0.027	0.066	0.067	sex	2pl
-0.472	-0.485	-0.013	0.008	0.089	0.089	sex	correct_
-0.472	-0.492	-0.020	0.010	0.098	0.098	sex	wrong_gu

Table 3 - simulation results, second series

d	log L	sex	se	poverty	se	hw2	se	hw3	se	
-0.10	-5651.00	-.419	.081	-.700	.128	.015	.094	.301	.117	
-0.09	-5647.99	-.423	.082	-.708	.130	.016	.096	.305	.118	
-0.08	-5645.16	-.427	.082	-.716	.131	.017	.097	.310	.119	
-0.07	-5642.52	-.431	.083	-.724	.132	.018	.098	.314	.120	
-0.06	-5640.08	-.435	.084	-.731	.134	.019	.098	.319	.121	
-0.05	-5637.85	-.439	.085	-.738	.135	.021	.099	.323	.123	
-0.04	-5635.83	-.443	.086	-.745	.136	.022	.100	.328	.124	
-0.03	-5634.03	-.447	.086	-.752	.137	.023	.101	.332	.125	
-0.02	-5632.46	-.451	.087	-.758	.138	.024	.102	.336	.127	
-0.01	-5631.11	-.454	.088	-.763	.139	.025	.103	.341	.127	
0	-5630.00	-.458	.089	-.769	.141	.027	.104	.345	.129	
+0.01	-5629.12	-.461	.090	-.775	.142	.028	.105	.349	.130	
+0.02	-5628.49	-.464	.091	-.780	.143	.029	.106	.352	.131	
+0.03	-5628.11	-.467	.091	-.784	.144	.030	.107	.356	.132	
+0.04	-5627.99	-.470	.092	-.789	.145	.031	.108	.359	.134	MLEs
+0.05	-5628.13	-.472	.093	-.793	.146	.033	.108	.361	.135	
+0.06	-5628.53	-.474	.094	-.796	.146	.034	.109	.363	.136	
+0.07	-5629.22	-.476	.094	-.800	.147	.035	.110	.364	.137	
+0.08	-5630.19	-.478	.095	-.803	.148	.036	.111	.365	.138	
+0.09	-5631.46	-.479	.096	-.806	.149	.038	.111	.365	.139	
+0.10	-5633.05	-.481	.096	-.809	.150	.039	.112	.363	.140	

Table 3 - simulation results, second series

d	ethnic2	se	ethnic3	se	ethnic4	se	
-0.10	-2.305	.151	-1.923	.143	.935	.174	
-0.09	-2.338	.153	-1.952	.144	.947	.175	
-0.08	-2.370	.154	-1.979	.145	.960	.177	
-0.07	-2.398	.155	-2.004	.146	.973	.179	
-0.06	-2.426	.156	-2.029	.147	.986	.180	
-0.05	-2.452	.158	-2.052	.149	1.001	.182	
-0.04	-2.476	.159	-2.073	.150	1.016	.184	
-0.03	-2.499	.160	-2.094	.151	1.031	.186	
-0.02	-2.520	.161	-2.113	.152	1.047	.189	
-0.01	-2.540	.162	-2.131	.153	1.065	.191	
0	-2.559	.163	-2.148	.154	1.083	.193	
+0.01	-2.576	.164	-2.164	.155	1.101	.196	
+0.02	-2.592	.165	-2.179	.156	1.121	.198	
+0.03	-2.607	.166	-2.193	.156	1.141	.201	
+0.04	-2.621	.167	-2.206	.157	1.162	.204	MLEs
+0.05	-2.634	.168	-2.218	.158	1.184	.206	
+0.06	-2.646	.169	-2.229	.159	1.206	.209	
+0.07	-2.657	.170	-2.239	.160	1.228	.212	
+0.08	-2.668	.171	-2.248	.161	1.249	.214	
+0.09	-2.677	.172	-2.257	.162	1.270	.216	
+0.10	-2.686	.173	-2.265	.162	1.290	.218	

Table 4 - Number of items per examinee

items	frequency	%	cumulative %
1	1,276	12.19	12.19
2	1,990	19.02	31.21
3	88	0.84	32.05
4	162	1.55	33.60
5	212	2.03	35.62
6	171	1.63	37.26
7	1,580	15.10	52.36
8	1,662	15.88	68.24
9	331	3.16	71.40
10	1,576	15.06	86.46
11	486	4.64	91.10
12	243	2.32	93.43
13	238	2.27	95.70
14	4	0.04	95.74
15	16	0.15	95.89
16	10	0.10	95.99
17	14	0.13	96.12
18	6	0.06	96.18
19	10	0.10	96.27
20	14	0.13	96.41
21	13	0.12	96.53
22	54	0.52	97.05
23	18	0.17	97.22
24	16	0.15	97.37
25	275	2.63	100.00
Total	10,465	100.00	

Table 5 - Number of examinees per item

item	frequency	%	cumulative %
1	2,697	3.69	3.69
2	2,694	3.68	7.37
3	2,688	3.68	11.05
4	2,599	3.55	14.61
5	2,527	3.46	18.06
6	2,446	3.35	21.41
7	2,372	3.24	24.65
8	2,318	3.17	27.82
9	2,734	3.74	31.56
10	2,718	3.72	35.28
11	2,690	3.68	38.96
12	2,671	3.65	42.61
13	2,656	3.63	46.24
14	2,640	3.61	49.86
15	2,613	3.57	53.43
16	2,027	2.77	56.20
17	1,945	2.66	58.86
18	1,890	2.59	61.45
19	1,764	2.41	63.86
20	1,534	2.10	65.96
21	2,699	3.69	69.65
22	2,695	3.69	73.34
23	2,692	3.68	77.02
24	2,688	3.68	80.70
25	2,537	3.47	84.17
26	2,344	3.21	87.37
27	2,266	3.10	90.47
28	1,905	2.61	93.08
29	2,684	3.67	96.75
30	2,378	3.25	100.00
Total	73,111	100.00	