Comparison of Direct Estimation with the Conditioning Model and Plausible Value Imputation

Prepared by:

Murray Aitkin and Irit Aitkin
School of Mathematics and Statistics
University of Newcastle-upon-Tyne
September 30, 2004

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 developed further by Mislevy (1985) 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 report examines the second of two of these possibilities:

- To examine the current method for assessing ability differences across important reporting groups through the generation of plausible values from a large conditioning model, and compare this with the direct assessment of these differences through an appropriate multi-level regression model.

2 Summary

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

1. The direct maximum likelihood estimates of the reporting group difference parameters from the data generating model, obtained by fitting jointly the reporting group variables and the items, were consistently superior to all the other estimates, with smaller biases and mean square errors than those based on any of the conditioning models.

2. The current NAEP analysis method using plausible values gave seriously biased estimates of some of the reporting group difference parameters.

3. These biases were reduced, but not eliminated, by using the data generating model as the conditioning model, and joint estimation of the item and reporting group difference parameters. In particular, one-way tabulations of plausible values for correlated reporting group variables gave biased parameter estimates for these variables.

4. The additional computational effort required to generate and analyze the plausible values is no longer warranted with current computational power; multi-level model analysis provides better estimates and also allows the sample design to be incorporated efficiently into the analysis.

5. If multi-level model analysis is to be adopted as the standard for reporting survey results, the form of presentation of these results needs consideration.

3 Model structures used in this study

For the purpose of this relatively small-scale simulation, the NAEP test will consist of a set of 10 binary items, which assess the respondent's ability through a set of two-parameter logit (2PL) models for each item. The binary responses Y_{ij} of the i-th respondent to the j-th item are conditionally independent given the ability θ_i of the respondent, with probability p_{ij} that the i-th respondent gives a correct answer $(y_{ij} = 1)$ on item j.

We write the 2PL model as

$$\begin{aligned} \Pr[Y_{ij} = 1 \mid \theta_i] &= p_{ij} \\ \operatorname{logit} p_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \alpha_j + \gamma_j \theta_i, \\ \theta_i &\sim N(\boldsymbol{\beta}'\mathbf{x}_i, 1), \end{aligned} \tag{1}$$

with respondent ability θ_i normally distributed with variance 1 about a mean $\boldsymbol{\beta}'\mathbf{x}_i$ determined by the values for the i-th respondent of a set of explanatory variables $\mathbf{x}$. The slopes γ_j and intercepts α_j of the logistic regressions are used as the *item parameters*; the *discrimination parameters* are the slopes γ_j and the *difficulty parameters* are $-\alpha_j/\gamma_j$. The aim of the analysis is to describe group differences in ability for relevant *reporting groups*. These differences are represented by parameters $\boldsymbol{\beta}_1$ in a regression model $\boldsymbol{\beta}_1'\mathbf{x}_{1i}$, where $\mathbf{x}_1$ is the subset of reporting group variables of the full set of variables $\mathbf{x}$. The item parameters are not themselves of direct interest but their unknown values need to be estimated to obtain information about the regression model parameters $\boldsymbol{\beta}_1$.

(The NAEP model is not the only possible model relating ability to the item responses. A detailed discussion of this point is given in Appendix 2, where we point out that the NAEP model has the inherent property of *differential item functioning*).

Any analysis of the data relates the item and regression model parameters to the item responses y_{ij} and reporting group variable values $\mathbf{x}_{1i}$ through the *likelihood function* – the probability of the observed data as a function of the model parameters. Here the likelihood function has to be expressed as an *integral* over the unobserved ability θ_i:

$$L(\alpha_1, \gamma_1, ..., \alpha_{10}, \gamma_{10}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} \prod_{j=1}^{10} p_{ij}^{y_{ij}} (1-p_{ij})^{1-y_{ij}} \phi(\theta_i - \boldsymbol{\beta}'\mathbf{x}_i) \mathrm{d}\theta_i,$$

where $\phi(z)$ is the standard normal density function. Both the *direct method* and the *current NAEP method* discussed in this report make use of the *maximum likelihood estimates* (MLEs) of the model parameters in the likelihood, but do so in quite different ways. Formally, the MLEs of the item and regression model parameters are the solutions to the *score equations* which equate to zero the derivatives of the log-likelihood with respect to these parameters. The standard errors of the MLEs are obtained by inverting the *information matrix*, the matrix of negative second derivatives of the log-likelihood evaluated at the MLEs.

With a large number of respondents, items and reporting group variables, the iterative solution of the score equations and inversion of the information matrix are time-consuming computations, though modern processing speeds have greatly reduced this burden. We used the GLLAMM routines in STATA for all the analyses reported here. These routines can handle arbitrary statistical models; this generality comes at the expense of slow processing since the first and second derivatives of the log-likelihood required for the MLEs have to be computed numerically. These derivatives can be computed analytically instead and this would reduce computation times considerably. However we used the package facilities for generality, and simplicity of programming.

4 The current NAEP analysis

The current NAEP analysis uses the model components as described above, with a very large "conditioning" regression model vector $\mathbf{x}$, one with all the two-way interactions of the reporting group variables, and a large number of other variables, some involving the survey design. This total number of variables is so large, and the correlations between them so high, that the total set (of up to 1000) variables is not used directly but is first reduced to a set of uncorrelated principal components, and a subset (around 200 or so) of these principal variables is used instead.

We do not follow this part of the current analysis because we use at most 24 variables in our analysis (the full two-way interactions of the four reporting group variables we define) and so no reduction in the number of variables is needed computationally; this also allows us to compare the direct method with the current NAEP method without any loss of information in the reduction to principal variables.

NAEP model fitting

An important aspect of the current NAEP analysis is that the NAEP method *does not fit simultaneously the item parameters and the regression model parameters* by maximum likelihood. Instead, the NAEP method fits *first* the item parameters *without any regression model parameters.* The item parameter estimates from this analysis are then taken as fixed known values of the item parameters, and the full regression "conditioning model" is fitted to ability.

Plausible value generation

The fitted conditioning model is then used to generate $M = 5$ plausible values of ability for each individual from the posterior distribution of ability given the item responses. This is done in four stages:

1. The posterior distribution of the conditioning model regression parameter vector is assumed to be normal, with mean the MLE of the parameter vector and covariance matrix the estimated covariance matrix of the MLE. A random value of the conditioning regression vector is then generated from this normal posterior distribution, and combined with the reporting group variable values for each individual to give the conditioning model value for this individual.

2. The posterior distribution of ability θ_i for each individual i is then constructed on 41 equally spaced quadrature points $\theta \in -5(0.25)+5$ by evaluating the likelihood for this individual, multiplying by the discrete normal prior distribution on these quadrature points, and scaling the sum to 1.0.

3. A random value of each individual's ability is then generated from the posterior distribution, by first drawing at random a quadrature point with probability equal to the quadrature mass, and then drawing uniformly a plausible (imputed) ability value between the upper and lower end points of the interval at which the quadrature point was centered.

4. The three steps above are repeated $M = 5$ times to give M plausible values of ability for each individual.

The M plausible values are then used in M analyses, which involve (one-way or two-way) *tabulations* of the ability values by each of the reporting group variables, to give reporting group means and standard errors.

Finally the M sets of group estimates and standard errors are combined using the Rubin rules for multiple imputation to give a single set of reporting group means and standard errors. These are converted in our analysis to reporting group mean differences and standard errors, to be comparable with the direct regression parameter estimates and standard errors.

5 The direct estimation method

The direct method of estimation is much simpler. This method uses the reporting group variables $\mathbf{x}_1$ as the variables in the regression and performs maximum likelihood computations for the item and regression model parameters $\boldsymbol{\beta}_1$ *fitted jointly* or *simultaneously* in the analysis, and reports the regression model parameter MLEs $\hat{\boldsymbol{\beta}}_1$ and their standard errors (and covariances if these are needed) as the descriptions of group mean differences and their standard errors.

6 The data

We used 10 items, with item slopes γ_j and intercepts α_j chosen to cover a wide range of item difficulties and discriminations. We used 1000 subjects, to allow reasonable computing times for around 250 samples. We used four reporting group variables, representing real NAEP practice as far as possible. We defined a four-group variable *ethnicity*, with groups
(W)hite, (B)lack, H(ispanic), and A(sian);
a two-group *gender* variable (M,F), coded to a dummy variable *sex* $= 1$ for F, 0 for M;
a *poverty* variable ($0 =$ no,$1 =$ yes)
and a three-group *homework* variable defined by two dummy variables:
hw2 $= 1$ if one hour of homework daily, and 0 otherwise,
hw3 $= 1$ if more than one hour of homework daily, and 0 otherwise.

The ethnicity classification was converted to three dummy variables *ethnic2, ethnic3* and *ethnic4*, representing the contrasts (B-W), (H-W) and (A-W):

Group	ethnic2	ethnic3	ethnic4
White	0	0	0
Black	1	0	0
Hispanic	0	1	0
Asian	0	0	1

Thus the four reporting variables are defined by seven dummy variables: three for ethnicity, one for sex, one for poverty and two for homework. A main effect model with these variables has seven model parameters (apart from the intercept) which represent the reporting group mean ability differences.

Data generating models

We generated the item response data from two different data generating models. In the first, model *M1*, the regression model in equation (1) had only the main effects of the reporting group variables. In the second, model *M2*, the regression model had these main effects and one two-way interaction term, the interaction between sex and homework. The values of the regression coefficients, and the structure of the population model, are described in detail in Appendix 1.

7 Analyses of the simulated data

The project was carried out in stages, determined by the results from each stage. Over the whole study we fitted a wide range of models, which led to a large number of methods for estimating the reporting group parameters and their standard errors. These are reported on the simulation model scale, not the NAEP reporting scale, though this does not affect any interpretation since the NAEP scale is a linear transformation of the model scale we use in the simulations.

It is helpful to describe and list the estimation methods here, before discussing specific cases.

- **Method 1** : MLEs from the one-interaction regression model, based on the current NAEP method of separate fitting of the item parameters and the main effects and one interaction regression model.

- **Method 2** : MLEs from the main effect regression model, based on the current NAEP method of separate fitting of the item parameters and the main effect regression model.

- **Method 3** : MLEs from the one-interaction regression model, based on *joint* fitting of the item parameters and the main effect and one interaction regression model.

- **Method 4** : MLEs from the main effect regression model, based on joint fitting of the item parameters and the main effect regression model.

- **Method 5** : Reporting group mean differences based on tabulation of plausible values generated using the current NAEP method, based on separate fitting of the item parameters and the full two-way interaction regression model.

- **Method 6** : Reporting group mean differences based on tabulation of plausible values generated using the current NAEP method, based on separate fitting of the item parameters and the main effect and one interaction regression model.

- **Method 7** : Reporting group mean differences based on tabulation of plausible values generated using the current NAEP method, based on separate fitting of the item parameters and the main effect regression model.

- **Method 8** : Reporting group mean differences based on tabulation of plausible values generated using the current NAEP method, but based on *joint* fitting of the item parameters and the main effect regression model.

- **Method 9** : Reporting group mean differences based on tabulation of plausible values generated using the current NAEP method, but based on *joint* fitting of the item parameters and the full two-way interaction regression model.

These methods group into three categories:

- Methods 1 and 2 use the current NAEP separate estimation method to obtain MLEs of item parameters, fix them, obtain the regression model parameter estimates, and then treat these as direct estimates.
- Methods 3 and 4 use joint estimation of the item and model parameters and use the MLEs of the regression model parameters as direct estimates.
- Methods 5-7 use the current NAEP separate estimation method (first items, then regression) to generate plausible value-based estimates.
- Methods 8 and 9 also generate plausible value-based estimates, but using *joint* estimation of the item and regression model parameters rather than separate two-step estimation.

8 Stage 1

We generated $N = 241$ random samples of size $n = 1000$ from the model *M1*, and fitted a sequence of models to each sample.

These models were:

- A: a *constrained full interaction model*, in which the items were first fitted without the reporting group variables; the item parameter estimates from this model were then held fixed in a second model including the items and the regression of ability on the main effects and all two-way interactions of the reporting group variables. The fitted regression model was then used to generate five plausible values, which were used to obtain the **Method 5** estimates and standard errors for the model parameters defined by the reporting group variables, as described above. These are the estimates reported in current NAEP analyses (apart from the effect of the principal component reduction mentioned above).
- B: a *constrained main effect model*, in which the items were first fitted without the reporting group variables; the item parameter estimates from this model were then held fixed in a second model including the items and the regression of ability on *only* the main effects (the correct data generating model) of the reporting group variables. The main effect model parameter estimates from this analysis are the **Method**

2 estimates. The fitted main effect regression model was then used to generate five plausible values, and reporting group difference estimates and standard errors as in A. The estimates resulting from the analysis of the plausible values are the **Method 7** estimates.

- C: a *joint main effect model*, in which the items *and* the ability main effect regression model were fitted *jointly*, by full maximum likelihood. The main effect model parameter estimates from this analysis are the **Method 4** estimates.

In presenting the comparative results in Table 1, we summarize estimates for the seven main effect parameters - the sex difference (F-M), ethnic group differences relative to White (B-W, H-W, A-W), poverty (Yes-No), and homework (1 hr-0, >1 hr-0). The summary values reported for each estimation method (averaged over 241 samples) are mean, bias, MSE – mean square error across samples (squared bias plus variance), standard deviation SD – the square root of the variance across samples, and SE – the average across samples of the standard error reported by the method. We also assessed the estimated item intercepts and slopes in the same way. Since these parameters are not presented in NAEP reports and are not fundamental to the analysis, we do not report them here, though we refer to the results in the text.

TABLE 1

TRUE	MEAN	BIAS	MSE	SD	SE	PARAM	METHOD
-0.472	-0.035	0.436	0.194	0.057	0.062	sex	2
	-0.469	0.002	0.006	0.078	0.081		4
	-0.320	0.152	0.027	0.065	0.078		5
	-0.276	0.195	0.041	0.057	0.076		7
-2.359	-1.637	0.722	0.534	0.117	0.128	ethnic2	2
	-2.394	-0.035	0.031	0.171	0.165		4
	-1.932	0.427	0.196	0.117	0.099		5
	-1.882	0.477	0.240	0.109	0.095		7
-1.887	-1.278	0.610	0.384	0.113	0.119	ethnic3	2
	-1.928	-0.041	0.025	0.153	0.148		4
	-1.561	0.326	0.117	0.105	0.098		5
	-1.526	0.361	0.143	0.111	0.095		7
0.944	0.945	0.002	0.021	0.143	0.149	ethnic4	2
	0.929	-0.014	0.031	0.174	0.173		4
	0.766	-0.178	0.050	0.136	0.131		5
	0.771	-0.172	0.049	0.138	0.126		7
0.100	0.365	0.265	0.076	0.074	0.078	hw2	2
	0.099	-0.001	0.008	0.088	0.092		4
	0.109	0.009	0.008	0.091	0.091		5
	0.149	0.049	0.010	0.088	0.088		7
0.300	0.522	0.222	0.057	0.089	0.100	hw3	2
	0.305	0.005	0.014	0.117	0.117		4
	0.273	-0.027	0.017	0.126	0.115		5
	0.312	0.012	0.014	0.118	0.110		7
-0.800	-0.484	0.316	0.112	0.107	0.113	poverty	2
	-0.823	-0.023	0.016	0.126	0.128		4
	-1.011	-0.211	0.055	0.101	0.120		5
	-0.966	-0.166	0.037	0.098	0.115		7

Mean Method 2 maximized log-likelihood : -5655.229 (7 parameters)
Mean Method 4 maximized log-likelihood : -5582.049 (7 parameters)

9 Conclusions from Stage 1

It is very clear that the current NAEP method of analysis, Method 5, when applied to our relatively small model and data sets, does not give satisfactory estimates of the reporting group difference parameters because of the serious biases in the estimates of the larger parameters – these are up to 3 or 4 times their sampling standard errors.

The source of this unsatisfactory performance is also clear from the analyses. The Method 4 *joint* ML estimation of both item parameters *and* reporting group difference parameters gave uniformly good performance, in terms of negligible bias and small mean square error. This is to be expected from standard results from statistical theory on consistency and asymptotic efficiency of ML estimates; the sample size of 1000 is large enough for both small bias of the estimates and good agreement between the reported standard error (averaged across samples) and the actual sampling variation of the estimates.

The Method 2 analysis fits the correct generating model, but by the same two-stage process as in Method 5. The "ML" regression parameter estimates produced by this method (Method 2) have even larger biases than the Method 5 estimates – they are the worst of all the methods. When plausible value are generated from the fitted model estimated by Method 2, and then tabulated by the reporting group variables to give the Method 7 estimates, the biases are substantially reduced, and are then very similar to those of the Method 5 estimates.

Method 2 is based on a *constrained* two-stage version of ML estimation – the regression model parameter estimates obtained by this two-stage version will not in general be full ML estimates, so they lack the properties of consistency and efficiency of the full ML estimates. The only condition under which they *are* equivalent to full ML estimates is that of *independence* of the sampling distributions of the item parameter estimates and the regression parameter estimates, in which case the constrained estimates of the item parameters and the regression model parameters will be very close to their full ML estimates. It is clear from the very large change in maximized likelihood between the two-stage estimates Method 2 estimates and the full ML Method 4 estimates (an average difference in log-likelihood of 73.18), and the serious biases of the Method 2 estimates, that the two sets of item parameter and regression parameter estimates are far from equal.

It might be thought that the smaller sampling variance of the Method 5 estimates shows some degreee of superiority of these estimates relative to the Method 4 estimates. However this improvement is quite outweighed by the large biases of the Method 5 estimates, giving them consistently larger mean square errors than the Method 4 ML estimates.

Further, at least part of this apparent improvement in variance comes about for the same reason as the serious bias – the two stage modeling of the item parameters, treating them as known in the second stage, *incorrectly reduces* the sampling variance of the constrained "MLE"s. This is clear by comparing the variances of the Method 2 and Method 5 estimates – they are very similar to each other, and consistently smaller than the variances of the Method 4 ML estimates.

It might be argued that the use of a main effect data generating model might have contributed to the differences described above, since the current NAEP analysis method allows for two-way interactions but the data generating model has only main effects. We addressed this point in Stage 2, where we extended the data generating model by including an interaction. The simulation times required for fitting the general two-way interaction model increased considerably, so we restricted the complexity of the model to a single interaction, of sex by homework.

10 Stage 2

The main effect data generating model *M1* was extended to an interaction model *M2* with a sex by homework interaction added to $M1$, defined by the additional term $0.4 * sex * hw3$. No term $sex * hw2$ was added (this is equivalent to adding $0.0 * sex * hw2$). We generated $N = 250$ samples of size $n = 1000$ from *M2*, and fitted a larger sequence of models to each sample. Models A, B and C were fitted as in Stage 1, leading to corresponding estimates by methods 5, 2, 7 and 4 as before. In addition, we fitted the following two models:

- D: a *constrained single-interaction model*, which was the model in B with the addition of the two interaction terms $sex * hw2$ and $sex * hw3$. The regression parameter estimates obtained from fitting this model are the **Method 1** estimates. The fitted interaction model was then used to generate five plausible values as in A and B. The estimates resulting from the analysis of the plausible values are the **Method 6** estimates.

- E: a *joint single-interaction model*, in which the items and the ability main effects and single interaction regression model were fitted jointly by full maximum likelihood. The model parameter estimates from this analysis are the **Method 3** estimates.

Table 2 gives the mean, average bias, mean square error, sampling standard deviation and average standard error, for each of the regression model parameters, as in Table 1.

TABLE 2

TRUE	MEAN	BIAS	MSE	SD	SE	PARAM	METHOD
-0.472	0.107	0.578	0.340	0.071	0.073	sex	1
	-0.002	0.469	0.224	0.061	0.061		2
	-0.464	0.007	0.012	0.110	0.107		3
	-0.405	0.067	0.011	0.081	0.080		4
	-0.273	0.198	0.043	0.062	0.078		5
	-0.246	0.226	0.055	0.063	0.077		6
-2.359	-1.685	0.673	0.467	0.114	0.128	ethnic2	1
	-1.653	0.706	0.512	0.116	0.129		2
	-2.399	-0.040	0.029	0.164	0.166		3
	-2.389	-0.030	0.027	0.163	0.165		4
	-1.939	0.420	0.189	0.111	0.099		5
	-1.906	0.453	0.217	0.106	0.096		6
-1.887	-1.318	0.569	0.336	0.108	0.118	ethnic3	1
	-1.287	0.600	0.372	0.107	0.117		2
	-1.910	-0.023	0.024	0.153	0.147		3
	-1.903	-0.015	0.023	0.152	0.147		4
	-1.556	0.331	0.121	0.108	0.097		5
	-1.526	0.361	0.141	0.103	0.094		6
0.944	0.920	-0.023	0.023	0.151	0.153	ethnic4	1
	0.956	0.012	0.022	0.148	0.151		2
	0.970	0.026	0.031	0.173	0.174		3
	0.966	0.023	0.030	0.172	0.174		4
	0.804	-0.140	0.038	0.137	0.131		5
	0.793	-0.151	0.041	0.137	0.127		6
0.100	0.547	0.447	0.210	0.102	0.102	hw2	1
	0.354	0.254	0.071	0.078	0.078		2
	0.106	0.006	0.019	0.139	0.130		3
	0.105	0.005	0.009	0.094	0.092		4
	0.113	0.013	0.009	0.092	0.091		5
	0.137	0.037	0.010	0.092	0.089		6
0.300	0.552	0.252	0.082	0.136	0.134	hw3	1
	0.510	0.210	0.055	0.105	0.101		2
	0.304	0.004	0.015	0.124	0.116		3

```
          0.110   -0.190   0.064   0.167   0.164                    4
          0.270   -0.030   0.017   0.126   0.115                    5
          0.294   -0.006   0.015   0.124   0.111                    6

-0.800   -0.519    0.281   0.091   0.108   0.112   poverty          1
         -0.494    0.306   0.106   0.111   0.111                    2
         -0.819   -0.019   0.015   0.123   0.128                    3
         -0.815   -0.015   0.015   0.123   0.128                    4
         -1.018   -0.218   0.057   0.100   0.121                    5
         -0.972   -0.172   0.038   0.090   0.115                    6

 0.000   -0.472   -0.472   0.247   0.156   0.159   sex*hw2          1
         -0.000   -0.000   0.036   0.191   0.184                    3
          0        0       0                                        4
          0.115    0.115   0.051   0.195   0.202                    5
          0.132    0.132   0.054   0.191   0.200                    6

 0.400   -0.169   -0.569   0.364   0.202   0.201   sex*hw3          1
          0.391   -0.009   0.059   0.243   0.233                    3
          0       -0.400   0.160   0       0                        4
         -0.198   -0.598   0.424   0.258   0.259                    5
         -0.167   -0.567   0.387   0.256   0.251                    6

Mean Method 5 maximized log-likelihood : -5632.759 (24 parameters)
Mean Method 1 maximized log-likelihood : -5644.464 ( 9 parameters)
Mean Method 2 maximized log-likelihood : -5649.766 ( 7 parameters)
Mean Method 4 maximized log-likelihood : -5580.971 ( 7 parameters)
Mean Method 3 maximized log-likelihood : -5578.372 ( 9 parameters)
```

11 Conclusions from Stage 2

Results from this analysis are very similar to those from Stage 1. The full ML Methods 3 and 4 based on the main effect and single interaction model *or* the main effect model only performed equally well, except for Method 4 with the $sex * hw3$ interaction (which is not fitted – so set to zero) and one SE biases in the main effects of sex and $hw3$. Methods 5 and 6 again performed poorly for the large parameters, and for the $sex.hw3$ interaction. Methods 1 and 2 were worst.

The addition of a small interaction term did not change the relative performance of the methods. The large biases of the plausible value methods led us, in the third stage of the study, to examine separately the effect on these estimates of the model fitted and the plausible value generation process.

12 Stage 3

In this stage we generated data from the main effect Model *M1* as in Stage 1, and fitted two models : model C – the items and main effect regression model fitted by *full joint maximum likelihood* as in Stage 1, and an additional model F – the items and full two-way interaction model, again fitted by full joint maximum likelihood. The model C analysis provided both **Method 4** ML estimates and **Method 8** estimates from the plausible value generation from the Method 4 estimates. The Model F analysis provided plausible values and corresponding **Method 9** estimates; these were analogous to the Method 7 estimates in Stage 1 but were based on full joint ML estimation of the model parameters.

The comparison of the Method 4 and Method 8 estimates shows the effect of the plausible value generation process from a *correctly* specified and estimated model; the comparison of the Method 8 and Method 9 estimates shows the effect of an *over-complex* model (with many redundant terms in the regression) fitted by full maximum likelihood on the precision of the plausible value estimates.

Table 3 gives the mean, average bias, mean square error, sampling standard deviation and average standard error, for each of the regression model parameters by Methods 4, 8 and 9, as in Tables 1 and 2.

TABLE 3

TRUE	MEAN	BIAS	MSE	SD	SE	PARAM	METHOD
-0.472	-0.469	0.003	0.008	0.087	0.081	sex	4
	-0.462	0.009	0.008	0.087	0.092		8
	-0.325	0.147	0.026	0.066	0.077		9
-2.359	-2.396	-0.037	0.033	0.178	0.165	ethnic2	4
	-2.459	-0.100	0.037	0.163	0.102		8
	-1.930	0.429	0.199	0.121	0.099		9
-1.887	-1.903	-0.016	0.019	0.135	0.147	ethnic3	4
	-1.991	-0.104	0.030	0.139	0.105		8
	-1.545	0.342	0.127	0.099	0.098		9
0.944	0.945	0.002	0.034	0.186	0.174	ethnic4	4
	0.948	0.005	0.036	0.191	0.144		8
	0.781	-0.163	0.050	0.152	0.131		9
0.100	0.102	0.002	0.009	0.097	0.092	hw2	4
	0.101	0.001	0.015	0.124	0.107		8
	0.109	0.009	0.010	0.101	0.091		9
0.300	0.296	-0.004	0.015	0.124	0.116	hw3	4
	0.291	-0.009	0.030	0.172	0.135		8
	0.260	-0.040	0.020	0.135	0.115		9
-0.800	-0.812	-0.012	0.015	0.123	0.128	poverty	4
	-1.269	-0.469	0.234	0.118	0.135		8
	-1.004	-0.204	0.052	0.103	0.119		9

13 Conclusions from Stage 3

Method 4 again performed uniformly well, with the smallest biases and mean square errors in all cases except sex, where Method 8 had the same mean square error, and homework2, where the Method 8 bias was even smaller. Method 9 was generally worse than Method 8, with larger biases and mean square errors except for poverty, where Method 8 had a severe bias – nearly four standard errors, while Method 9 had a bias of two standard errors – and for both homework parameters, where Method 9 had a slightly larger bias but slightly smaller mean square error.

Method 8 had appreciable biases for the ethnic2 and ethnic3 parameters – about one standard error, and a very severe bias for poverty. For the other parameters it performed quite well. The biases in the poverty and ethnic parameters result from the one-way tabulations used to construct the plausible value estimates; since the proportions of the population in poverty vary across the ethnic groups, these variables are correlated in the design matrix, and so one-way tabulations (equivalent to one-factor regression models) with these variables separately lead to biased estimates of the model parameters. The Model 4 estimates of these parameters are unbiased because these variables (and all the others) are fitted jointly.

14 Overall conclusions from the study

14.1 The performance of plausible value-based estimates

This study shows that *the direct simultaneous maximum likelihood estimation of both item and regression model parameters provided unbiased estimates (in the sample size used) of the reporting group difference model parameters, and these estimates had almost uniformly smallest mean square errors.* The estimates produced by the current NAEP analysis based on plausible values generated from a full two-way interaction fitted conditioning model, with item parameters held at their estimates from a null regression model, were poor, with serious biases and large mean square errors.

These current method estimates could be improved in two ways, by jointly estimating the item parameters and regression model parameters, and by using a conditioning model which corresponds to the data generating model. Since the generation of the plausible values requires as starting values the maximum likelihood estimates of the parameters of the model being fitted, *there seems no good reason to continue the current process of generating plausible values from a model whose likelihood is not fully maximized.*

Even with both these improvements, the plausible value estimates had some biases for correlated parameters and generally larger mean square errors than the direct ML estimates. As a consequence, *there seems no point* *in performing the additional heavy computation needed to generate plausible values for the NAEP model considered here*; the full ML estimates based on joint estimation had better properties than any of those based on plausible values, from any fitted model.

14.2 The generality of plausible value-based estimates

In defense of plausible values, it can be argued that the aim of the plausible value generation process is to allow any appropriate reporting group analysis, and other analyses not specified in advance, to be carried out from the plausible values without the need for repeated full analyses of the original survey data. Since we do not know, in real NAEP data, what is the "true" regression model (or operationally, the simplest model which provides an accurate representation of the population data within sampling error), the plausible value approach appears to provide a "fallback" analysis which should provide good estimates whatever the "true" model.

However this study shows that *when the conditioning model used in the current NAEP approach is over-complex relative to the true data generating model, the plausible value approach does not give good estimates.* There is also some evidence, based on the Stage 2 study, that an over-simple model (*M1*) may give good Method 4 estimates of the corresponding parameters in a more complex true model (*M2*) if the discrepancy between the models is not great.

14.3 Correspondence between the fitted and true models

An important point for both approaches is therefore that *the model being fitted and interpreted should correspond as far as possible to the "true" model generating the data.* Achieving this correspondence is a standard part of regression and multi-level modeling, in which a complex model is reduced by backward elimination or other stepwise model reduction procedure to a parsimonious model containing the important effects, which is then interpreted.

Further, if there are complex interactions or correlated reporting group variables in the real data, *the presentation of one-way tabulations of the variables will itself be potentially misleading unless the existence of such interactions is investigated.* In the simulated main effect model data with no interactions but correlated ethnic group and poverty variables, the one-way tabulations of plausible values for these two variables gave biased estimates of the effects of these variables, while those for the other two variables sex and homework did not. The investigation process therefore *requires* a model examination and simplification process for the full extraction of the information about reporting group differences.

Such simplification procedures will be assessed in a later stage of this study.

14.4 Survey design

An important point in the maximum likelihood analysis given here is that the data generating model used does not allow for a stratified or clustered survey design, as is typically used in NAEP. However *the maximum likelihood analysis above is easily extended to incorporate both these design aspects*:

- *Stratification* is allowed for by incorporating the strata as a factor in the model, together with its interactions with the explanatory variables if different regressions are expected in different strata. Disproportionate sampling within strata is simply handled by reweighting across strata when aggregate population results are reported.

- A *multi-stage cluster design* is represented by a multi-level model incorporating as many levels as necessary for the sample design.

For example, a two-stage cluster sample of regions $\ell = 1, \ldots, L$, schools $k = 1, \ldots, n_\ell$ within regions, and students $i = 1, \ldots, n_{k\ell}$ within schools would use the extended model

$$\begin{aligned} \log\left(\frac{p_{ijk\ell}}{1-p_{ijk\ell}}\right) &= \alpha_j + \gamma_j\theta_i + \phi_\ell + \psi_{k\ell}, \\ \theta_i &\sim N(\boldsymbol{\beta}'\mathbf{x}_i, 1), \end{aligned}$$

where $\phi_\ell \sim N(0, \sigma_\ell^2)$, $\psi_{k\ell} \sim N(0, \sigma_{k\ell}^2)$, ϕ_ℓ and $\psi_{k\ell}$ are the region and school random effects, and σ_ℓ^2 and $\sigma_{k\ell}^2$ are the region and school variance components. This model can be fitted straightforwardly in GLLAMM. The assumption of normally distributed random effects is very weak, as shown in the first part of this study under this contract.

This extension will be assessed in a later stage of this study.

14.5 Reporting the multi-level model analysis

Current presentation of the reporting group differences is through one-way or two-way tabulations of the plausible values. The presence of correlations between reporting group variables complicates the presentation of results, since one-way tabulations by correlated variables are potentially misleading. The model parameter ML estimates and standard errors are the basic summary of the analysis, but these are not easily interpreted as they stand. From these parameters we may compute fitted values and their standard errors for any combination of the reporting group variables, so in the simulation model we could compute and print a four-way table of "predicted means" and standard deviations of ability, transformed to the NAEP reporting scale. Such tables are complex, extensive and difficult to read.

A simple alternative is to provide a Web tool to allow the user to access and compute his or her own tables, or other forms of presentation. The AM software provides an example, of direct computation of tables from the original survey data. The modeling approach suggested here uses only the model parameter estimates and covariance matrix, so is much faster. A simple version of a model-based tool using this apporoach was developed by Adnan Khan at ESSI under Murray Aitkin's direction as a small project. Such approaches could be considered as alternatives, or supplements, to the publication of extensive tables.

15 References

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

Mislevy, R.J. (1985) Estimation of latent group effects. *Journal of the American Statistical Association*, **80**, 177-196.

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

16 Appendix 1 – the data generating models

Main effect model *M1*

Table 1: Model parameters

Item	Intercept	Slope	Difficulty	Variable	Coefficient
1	0.856	2.0	-0.428	sex (F-M)	-0.472
2	-1.192	1.8	0.662	ethnic2(B-W)	-2.359
3	-0.195	1.6	0.122	ethnic3(H-W)	-1.887
4	-1.514	1.4	1.081	ethnic4(A-W)	0.943
5	-0.145	1.2	0.121	poverty(Y-N)	-0.800
6	-0.012	1.0	0.012	hw1(1-0)	0.100
7	-0.051	0.8	0.064	hw2(>1-0)	0.300
8	0.576	0.6	-0.960		
9	-0.367	0.4	0.918		
10	-0.396	0.2	1.980		

Blacks are 2.359 units below Whites, Hispanics are 1.887 units below Whites, and Asians are 0.943 units above whites. Girls are 0.472 units below boys.

The main effect regression *M1* model structure (the value of the regression model for the given combination of variables) is shown below crossed by sex and ethnic group, for the no poverty, homework 0 category combination.

Ethnicity		W	B	H	A
Sex	M	0.689	-1.670	-1.198	1.632
	F	0.217	-2.142	-1.670	1.160

Poverty reduces all values by 0.8, while 1 hour homework increases them by 0.1, and more than 1 hour homework increases them by 0.3.

For ease of interpretation, this model structure is given below on the NAEP reporting scale, with a mean of 250 and range 1-500, tabulated by sex, ethnic group and poverty.

Poverty	Ethnicity		W	B	H	A
No	Sex	M	260	225	232	274
		F	253	218	225	267
Yes	Sex	M	248	213	220	262
		F	241	206	213	255

One hour homework increases these values by 1.5, and more than one hour by 4.5, NAEP scale units.

Sample sizes for the subgroups are 700 W, 120 B, 120 H, 60 A, equally split into males and females. Proportions in poverty are fixed: 10% W and A, 25% B and 23% H, for both males and females. *Thus ethnicity and poverty are correlated in the population: poverty is more common in Black and Hispanic groups than in White and Asian groups.* Homework is assigned randomly, with a probability 0.57 for no homework, 0.28 for one hour, and 0.15 for more than one hour.

The full model is made up of the 2PL item model of responses given normally distributed ability, and the regression model of ability on the explanatory variables.

Interaction model *M2*

For the interaction model *M2*, the additional term $0.4 * sex * hw3$ reduces the sex difference between boys and girls from 0.472 to 0.072 for those with more than one hour homework; the difference remains at 0.472 for those with one hour or less homework. On the NAEP reporting scale, the interaction is 6 scale units.

17 Appendix 2 – the two possible item/regression models

The analysis carried out in this report used the NAEP model, in which the regression conditioning model acts directly on the ability distribution. However there is another possible form of this model, in which the regression model acts directly, on the logit scale, on the correct response probabilities for each item. The two possible models have different meanings and different implications for item functioning, and can be discriminated from test data.

We use the notation p_{ij} for the probability of a correct response on item j by person i, and $\operatorname{logit} p_{ij}$ for the logit transformation of p_{ij}:

$$\operatorname{logit} p_{ij} = \log \frac{p_{ij}}{1 - p_{ij}}.$$

The ability on the test of individual i is θ_i, and the 2PL model for the probability p_{ij} is

$$\operatorname{logit} p_{ij} = \alpha_j + \gamma_j \theta_i.$$

The reporting group variables for the i-th individual, and any interactions between them, are denoted by the vector $\mathbf{x}_i$, and the regression model affecting the responses of the i-th individual is $\boldsymbol{\beta}'\mathbf{x}_i$.

The two ways in which these models can be linked are:

1. The regression model *directly affects* the response probability on the logit scale:

$$\begin{array}{rcl} \operatorname{logit} p_{ij} & = & \alpha_j + \gamma_j \theta_i + \boldsymbol{\beta}'\mathbf{x}_i, \\ \theta_i & \sim & N(0,1). \end{array}$$

2. The regression model *indirectly affects* the response probability through the *ability* inside the logit model:

$$\begin{array}{rcl} \operatorname{logit} p_{ij} & = & \alpha_j + \gamma_j \theta_i, \\ \theta_i & \sim & N(\boldsymbol{\beta}'\mathbf{x}_i, 1). \end{array}$$

The analysis reported above used the second formulation, the standard NAEP model. This formulation is equivalent to

$$\begin{array}{rcl} \operatorname{logit} p_{ij} & = & \alpha_j + \gamma_j \theta_i^* + \gamma_j \boldsymbol{\beta}'\mathbf{x}_i, \\ \theta_i^* & \sim & N(0,1), \end{array}$$

where $\theta^* = \theta - \boldsymbol{\beta}'\mathbf{x}$.

In the second formulation the regression variables $\mathbf{x}$ *interact* on the logit scale with the items – gender and ethnic group differences depend on the item. This is a particular form of *differential item functioning* – the item discrimination parameter is different for different gender and ethnic groups.

If $\gamma_j = \gamma$ for all j (that is, the item model is a Rasch model), the two formulations are equivalent, but for the 2PL model they are *not* equivalent.

The models are specified differently in GLLAMM. The first formulation is fitted directly as a two-level GLMM with a logistic link and regressions of each item on ability, with explanatory variables in the logit linear predictor. The second formulation is fitted as a MIMIC (Multiple Indicators, Multiple Causes) model, with logistic regressions of items on ability, and a normal regression of ability on the explanatory variables.

Interpretation of the two formulations

The two models have the same number of parameters but these have different meanings, both mathematically and psychometrically. In the first model, ethnic and gender differences *do not directly affect ability*, but they *do* affect directly the response probabilities on the items. This can be interpreted as *uniform cultural and gender differences* in the test - all the test items are affected in the same way (on the logit scale).

In the second model, ethnic and gender differences *directly affect ability* – the model specifies different normal distributions of ability in different reporting groups. However the effect of these ability differences on the test items is *indirect*, since it is *modulated* by the item discrimination parameters. The consequence of this is that in this model *all* items show differential item functioning – those with high discrimination parameters will show large group differences, while those with small discrimination parameters will show small group differences.

An important feature of the model-based approach is that *it is possible to discriminate between these models from the available test data*, provided both models can be fitted by maximum likelihood. However this difference in interpretation occurs only for the 2PL model – for the Rasch model, these generating models are identical and cannot be discriminated.

To illustrate this, we give below the mean (over samples) of the log-likelihoods from the two fitted models when the data are generated from the usual NAEP model.

Mean log-likelihood for the logistic model : -5722.36

Mean log-likelihood for the NAEP model : -5577.24

The difference in favour of the correct NAEP model is very substantial – there is no doubt that in samples of this size the two models can be clearly discriminated.

It would be of considerable interest to compare the two models on actual

NAEP data. This comparison will be reported in a later analysis.

18 Appendix 3 – STATA program listing