Statistics for Social and Behavioral Sciences

Statistical Modeling of the National Assessment of Educational Progress

Springer

Statistical Modeling of the National Assessment of Educational Progress

Statistics for Social and Behavioral Sciences

Series Editors:
S. E. Fienberg
W. J. van der Linden

For other titles published in this series, go to
http://www.springer.com/series/3463

Murray Aitkin • Irit Aitkin

Statistical Modeling of the National Assessment of Educational Progress

Springer

Murray Aitkin
Department of Mathematics and Statistics
University of Melbourne
Melbourne Victoria 3010
Australia
murray.aitkin@unimelb.edu.au

Irit Aitkin
Department of Mathematics and Statistics
University of Melbourne
Melbourne Victoria 3010
Australia
irit.aitkin@unimelb.edu.au

ISBN 978-1-4419-9936-8 e-ISBN 978-1-4419-9937-5
DOI 10.1007/978-1-4419-9937-5
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011927940

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book is a description of research and data analysis carried out by the authors with substantial funding from the National Center for Education Statistics (NCES) and the Institute of Education Sciences (IES), divisions of the U.S. Department of Education, in partnership with the American Institutes for Research (AIR). The purpose of this work was to evaluate a new approach to the analysis and reporting of the large-scale surveys for the National Assessment of Educational Progress (NAEP) carried out for the NCES.

The new approach was based on a full statistical and psychometric model for students' responses to the test items, taking into account the design of the survey, the backgrounds of the students, and the classes, schools and communities in which the students were located.

The need for a new approach was driven by two unrelated issues: the demands for *secondary analysis* of the survey data by educational and other researchers who needed analyses more detailed than those published by NCES, and the need to accelerate the processing and publication of results from the surveys.

The modeling approach is complex and computationally intensive, but less so than the existing methods used for these surveys, and it has the twin advantages of *efficiency* in the statistical sense – making full use of the information in the data – and *optimality*: given the validity of the statistical model, this form of analysis is superior to any other non-Bayesian analysis in terms of precision of the estimates of group differences and regression coefficients of important variables.

The use of a full statistical model avoids the ad hoc methods that are otherwise necessary for the analysis of the data. It is dependent, for successful adoption, on efficient computational implementations in generally available software. Developments in this area have been rapid in the last ten years: we began our analyses in 2003 using Gllamm in Stata (Rabe-Hesketh and Skrondal 2005); see the Website.

```
http://www.gllamm.org
```

By 2008 we were able to use the very fast Latent Gold program which makes large-scale model fitting straightforward for NAEP data sets; see the Website.

```
http://www.statisticalinnovations.com/products
/latentgold_v4.html
```

The following chapters, apart from the first, are set out in a sequence representing the main aspects of the NAEP surveys and the development of methods for fitting the increasingly complex models resulting from the incorporation of these aspects. The content of the book is drawn mostly from our NCES research reports, which are described briefly in Chapter 4. We generally do not give references to specific reports in the text, as the chapters draw from many of the reports. The full reports themselves are available on our Website, as described in Chapter 4.

The models and analysis approach are illustrated with detailed results from two NAEP surveys. The first is from the 1986 national NAEP mathematics test and includes results on the set of 30 items from the Numbers and Operations: Knowledge and Skills subscale, for age 9/grade 3 children. The "explanatory" regression model fitted is quite small and was chosen to nearly replicate the tables of "reporting group" variables published by NCES for this survey. We extended this analysis to all 79 test items on three scales.

The second survey is from the 2005 national NAEP mathematics test and includes results on the set of 70 items from the Numbers and Operations scale for age 10/grade 4 children. We fitted a much larger regression model with variables from the student, teacher and school questionnaires. We analysed the California and Texas state subsamples with more complex item response models.

Chapter 1 is an introduction to the current theories of data analysis used for large-scale surveys. It may surprise non-statistician readers to find that there are major disputes within the statistics profession about the role of statistical models in official (national government) survey analysis. We describe the critical theoretical issues that divide the several theories, and give an indication of the extent to which each theory is used in current official practice.

Chapter 2 describes the current method of analysis of NAEP surveys. This has changed several times; we give the analysis that was used for the 1986 survey, which we use as an illustration in later chapters, and note the changes that have occurred since then. The design and analysis of the 1986 survey were very complicated, and we have omitted aspects of the design that are not critical to the analysis. Some complex sections (for example, jackknifing) have been described at length because these are critical for the comparison with our approach.

Chapter 3 sets out the psychometric models used in the NAEP analysis, gives some extensions of them using mixture distributions for student ability, discusses the survey designs used in the surveys, and gives the multilevel model representation of the designs.

Chapter 4 summarises the main conclusions from our extensive simulation studies, which showed the improvement in precision and the reduction in bias resulting from the fully model-based analysis of small-scale models compared with the current approach. References to the full reports on this work are given there.

Chapter 5 sets out the series of analyses we used with the range of models from Chapter 3 for the 30-item scale from the 1986 math test for age 9/grade 3 children. Chapter 6 extends these analyses to the full set of 79 items on the test. Chapter 7

applies more complex analyses to the 2005 national NAEP subsample for Texas for age 10/grade 4 children. Chapter 8 applies the same analyses to the 2005 subsample for California for age 10/grade 4 children.

Chapter 9 discusses the results of the analyses and draws conclusions about the benefits and limitations of fully model-based large-scale survey analysis.

Acknowledgements

Murray's interest in psychometric modeling began with his post-doctoral position with Lyle Jones at the Psychometric Laboratory, University of North Carolina at Chapel Hill in 1966–67, and developed substantially from his visiting year as a Fulbright Senior Fellow in Frederic Lord's psychometric research group at the Educational Testing Service (ETS), Princeton in 1971–72.

In his large-scale research programme on EM algorithm applications to incomplete data problems at the University of Lancaster 1979–85, Murray developed with Darrell Bock (Bock and Aitkin 1981) an EM algorithm for the 2PP model. This algorithm has been very widely extended to other psychometric models.

Murray returned to ETS as a Visiting Scholar in 1987–88. Here he reviewed (Aitkin 1988) the extent to which hierarchical variance component modeling, incorporating the survey design in additional levels of the model, could be used for the analysis of NAEP data, and the possible information that it could provide. He noted that fitting the 3PL model in a full hierarchical model was beyond the capabilities of available programs, and suggested variations to the E step of the EM algorithm that could give an approximate analysis.

Our joint interest in NAEP developed with Murray's appointment as Chief Statistician at the Education Statistics Services Institute (ESSI), American Institutes for Research, in Washington, D.C. in 2000–2002, as a senior consultant to NCES. It continued through a series of research contracts with NCES through AIR in subsequent years.

We have benefited greatly from discussions and interactions with many staff members at NCES, particularly Andrew Kolstad, Steve Gorman, and Alex Sedlacek, and are grateful for the ongoing support of Peggy Carr, NCES Associate Commissioner for Assessment. The outline of statistical theories in Chapter 1 has greatly benefited from two "brown-bag lunch" seminar series that Murray gave to NCES and AIR staff in Washington; we appreciate the comments and feedback from participants in these seminars. We much appreciate the many discussions with current and former senior staff at AIR, particularly Gary Phillips (a former Deputy and Acting Commissioner of NCES), Jon Cohen, Eugene Johnson, Ramsay Selden, and Laura Salganik. We are particularly grateful for administrative support from Laura Salganik and from Natalia Pane, Janet Baldwin-Anderson, and Linda Schafer at ESSI and its successor NAEP ESSI. We much appreciate the many other welcoming and helpful people at these institutes supporting the work of NCES.

We thank John Mazzeo and Andreas Oranje at the Educational Testing Service for help with data access and interpretation, Kentaro Yamamoto at ETS and Charles Lewis at Fordham University for discussions on psychometrics, and Chan Dayton, Bob Lissitz, and Bob Mislevy at the University of Maryland for helpful discussions. We thank Sophia Rabe-Hesketh and Jeroen Vermunt for many technical discussions that have helped greatly to clarify aspects of the analyses in Gllamm and Latent Gold, and Karl Keesman for much help with Stata.

The very large-scale simulations needed for the evaluation of competing methods were run in Melbourne on the cluster computers of the Victorian Partnership for Advanced Computing (VPAC). We thank Chris Samuel and Brett Pemberton of the VPAC staff for much help. We particularly thank Michael Beaty of the School of Mathematics and Statistics of the University of Newcastle, UK, for detailed and continuing help with many computing aspects of our work, and our son Yuval Marom for a great deal of help with programming.

At the University of Melbourne, we are grateful for the support and help of Pip Pattison, Henry Jackson, and Bruce Ferabend in the Department of Psychology, Peter Hall and Richard Huggins in the Department of Mathematics and Statistics, and Dirk van der Knijff in the High Performance Computing unit.

We extend our special thanks to Sue Wilson for constant encouragement and support.

In preparing the final version of this book, we were greatly helped by suggestions and advice from Steve Fienberg on the first draft, which was changed substantially. Any remaining errors or obscurities are entirely our responsibility.

Murray Aitkin — Melbourne
Irit Aitkin — December 2010

This work has been funded by federal funds from the U.S. Department of Education, National Center for Education Statistics, and Institute for Education Sciences under various contracts and grants. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education, National Center for Education Statistics, or AIR, nor does mention of trade names, commercial products, or organisations imply endorsement by the U.S. Government or AIR.

Contents

Chapter 1
Theories of Data Analysis and Statistical Inference

1.1 Introduction

Every survey, in any field, begins conceptually with a *population list*, a *sampling plan* or *sample design* by which an appropriate sample is to be drawn from the population, a *measurement instrument* specifying the information – *response* variables and *covariates* or *explanatory variables* – to be obtained from the sampled population members, and an *analysis plan* by which the response variables, and their relation to the covariates or explanatory variables, are to be analysed.

In the NAEP surveys that we describe and analyse, the population list is of school students of several ages and grades, the sampling plan is a complex clustered and stratified design, and the measurement instrument is a set of test items measuring achievement in mathematics or another subject (the response variables in many analyses, including ours) and a set of questionnaire items describing students, their home background, and teacher and school characteristics that we use as covariates for achievement, though many may be response variables in other analyses.

We use from now on the term *covariates*, rather than explanatory variables, as the NAEP surveys we discuss are *observational studies* in which the issue of *causality* – of whether variation in the covariates *causes* or *explains* variations in the outcome variables – cannot be assessed from the surveys, as these are not experimental studies involving *randomisation* of students to classes or to educational and family contexts.

The analysis plan is the subject of this book, which discusses in this chapter the different philosophies in the statistics profession about how data from such studies should be analysed. Our view of analysis is *model-based*: defined by a full statistical model – in contrast to the current analysis of these surveys, which is a mixture of model-based and *design-based*: defined by hypothetical replications of the survey design.

In describing and discussing the important differences in these approaches, we adapt the discussion in Chapter 1 of Aitkin (2010) and use several very simple examples that, however, make clear the importance of the philosophical differences.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 1
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_1, © Murray Aitkin and Irit Aitkin 2011

1.2 Example

We have a simple random sample of size 40 from a finite population of 648 families and for each family record the family income for the previous tax year. From this sample, we wish to draw an inference about the *population mean* family income for that tax year. How is this to be done? The sample of incomes, reported to the nearest thousand dollars, is given below.

```
        Family income, in units of 1000 dollars
-------------------------------------------------
26 35 38 39 42 46 47 47 47 52 53 55 55 56 58 60 60
60 60 60 65 65 67 67 69 70 71 72 75 77 80 81 85 93
96 104 104 107 119 120
-------------------------------------------------
```

Theories of data analysis and inference can be divided into two classes: those that use the *likelihood function* (defined below) as an important, or the sole, basis for the theory and those that do not give the likelihood any special status.

Within the first class, there is a division between theories that regard the likelihood as the *sole* function of the data that provides evidence about the model parameters and those that interpret the likelihood using other factors.

Within the second class, there is a division between theories that take some account of a *statistical model* for the data and those based exclusively on the properties of estimates of the parameters of interest in repeated sampling of the population. Comprehensive discussions of the main theories can be found in Welsh (1996) and Lindsey (1996), to which we refer frequently. We illustrate these theories with reference to the income problem above.

1.3 Statistical models

Theories that use the likelihood require a *statistical model* for the population from which the sample is taken, or more generally for the *process* that generates the data. Inspection of the sample income values shows that (in terms of the measurement unit of \$1000) they are *integers*, as are the other unsampled values in the population. So the population of size N can be expressed in terms of the *population counts* N_J at the possible distinct integer values of income Y_J or, equivalently by the *population proportions* $p_J = N_J/N$ at these values.

A (simplifying) statistical model is an *approximate representation* of the proportions p_J by a *smooth probability distribution* depending on a small number of *model parameters*. The form of the probability function is chosen (in this case of a large number of distinct values of Y) by matching the cumulative distribution function (cdf) of the probability distribution to the empirical cdf of the observed values. Figure 1.1 shows the empirical cdf of the sample values. A detailed discussion of this process is given in Aitkin et al. (2005) and Aitkin et al. (2009). We do not give

details here, but the matching process leads to the choice of an approximating continuous cdf model $F(y\,|\,\lambda)$ and corresponding density function $f(y\,|\,\lambda) = F'(y\,|\,\lambda)$; the probability p_J of Y_J is approximated by $F(Y_J + \delta/2\,|\,\lambda) - F(Y_J - \delta/2\,|\,\lambda)$, where δ is the measurement precision (which equals 1 in the units of measurement). When the variable Y is inherently discrete on a small number of values, as with count data, the values p_J are approximated directly by a discrete probability distribution model.

Figure 1.2 shows the cdf of a normal distribution with the same mean (67.1) and standard deviation (22.4) as the sample income data, superimposed on the empirical cdf. Figure 1.3 shows the same cdfs, but on the vertical probit scale of $\Phi^{-1}(p)$. On this scale it is clearer that the income sample has some degree of *skew*, with a longer right-hand tail of large values, so an approximating model with right skew might be appropriate. The gamma, lognormal, and Weibull distributions are possible choices. However, to establish which of several possible models is most appropriate for sample data requires advanced model comparison methods, which we discuss in later chapters.

Here we will assume that the normal distribution with parameters μ (the mean) and σ (the standard deviation) is a reasonable model, where μ is the *parameter of interest* and σ is a *nuisance parameter* – we want to draw conclusions about the parameter of interest, μ, but the model depends as well on the nuisance parameter σ:

$$f(y\,|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

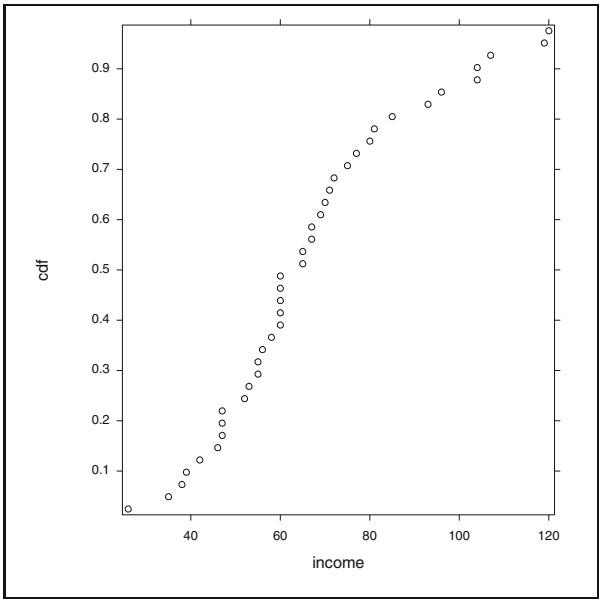

Fig. 1.1 cdf of sample income data

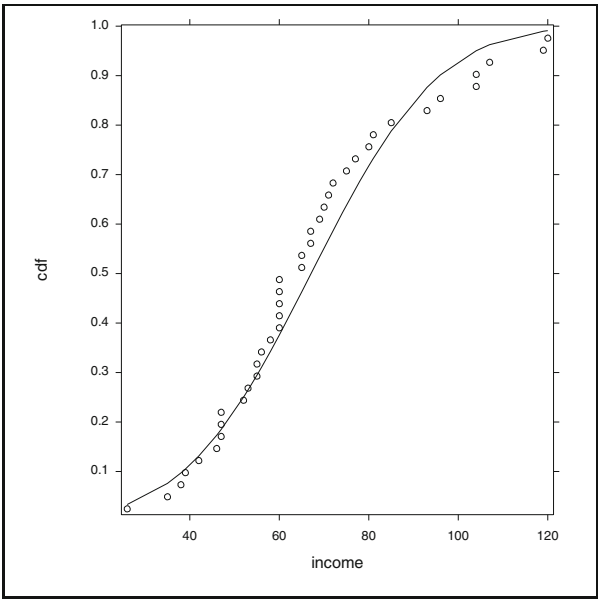

Fig. 1.2 cdfs of sample income data and normal distribution

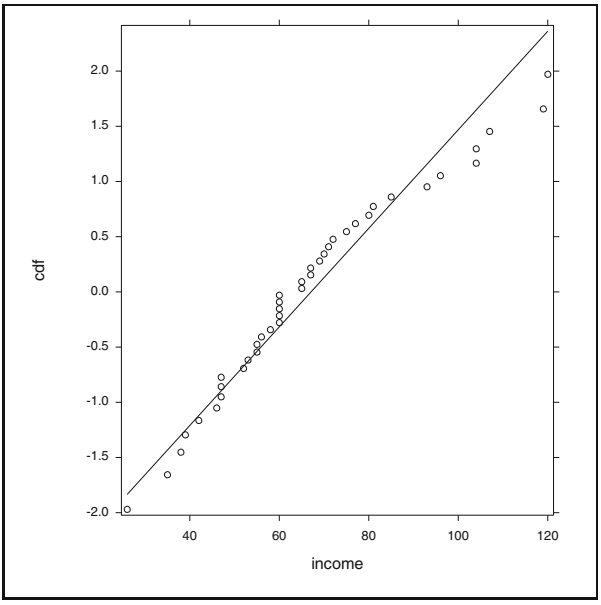

Fig. 1.3 cdfs, probit scale

1.4 The likelihood function

Given a simple random sample $\mathbf{y} = (y_1, \ldots, y_n)$ of size n, drawn from the population (assumed for the moment to be large compared with the sample) and an approximating statistical model $F(y \mid \lambda)$, the likelihood function $L(\lambda \mid \mathbf{y})$ (of the model parameters λ) is the probability of the observed data as a function of these parameters[1],

$$\begin{aligned} L(\lambda \mid \mathbf{y}) &= \Pr[y_1, \ldots, y_n \mid \lambda] \\ &= \prod_{i=1}^{n} [F(y_i + \delta/2 \mid \lambda) - F(y_i - \delta/2 \mid \lambda)] \\ &\doteq \left[\prod_{i=1}^{n} f(y_i \mid \lambda) \right] \cdot \delta^n, \end{aligned}$$

if the measurement precision is high relative to the variability in the data. In general, the parameter vector λ can be partitioned into a subvector θ of parameters of interest and a subvector ϕ of nuisance parameters.

Then, for high measurement precision δ in the normal model above, the *normal likelihood function*, given the sample data $\mathbf{y}$, can be written as

$$\begin{aligned} L(\mu, \sigma \mid \mathbf{y}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right\} \delta \right] \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{ -\frac{1}{2\sigma^2} \left[n(\bar{y} - \mu)^2 + RSS \right] \right\} \cdot \delta^n, \end{aligned}$$

where $\bar{y} = \sum_i y_i / n$, $RSS = \sum_i (y_i - \bar{y})^2$. An important theoretical point is that the likelihood function here depends on the data through only the two data functions $\bar{y}$ and RSS (and the sample size n, assumed fixed). These *sufficient statistics* are all that is needed to compute the likelihood function – we do not need the data values themselves.

We will generally drop the $\mathbf{y}$ from the notation for the likelihood function, but it is always implicit in its definition that the data have been observed. We now describe briefly the theories and how they deal with inference about the population mean income.

[1] The likelihood is frequently defined to be *any constant multiple* of the probability of the observed data, but in our approach likelihoods *are* probabilities.

1.5 Theories

1.5.1 Likelihood-based repeated sampling theory

This theory was dominant from the 1930s to the 1990s and is still prominent. The theory uses the likelihood function $L(\lambda \mid \mathbf{y})$ to provide both *maximum likelihood estimates* (MLEs) $\hat{\lambda}$ of the parameters and *likelihood-based confidence intervals* or *regions*. These are obtained in general from the first and second derivatives of the log-likelihood function with respect to the parameters λ.

The estimates and confidence intervals or regions are interpreted through their behavior in (hypothetical) *repeated sampling from the same population* (whence comes the Bayesian term *frequentist* to describe the theory). In large samples, MLEs have optimality properties as *point estimates*, though they may be biased in small samples.

For the income example, the MLEs of μ and σ are

$$\hat{\mu} = \bar{y},$$
$$\hat{\sigma}^2 = RSS/n.$$

The sampling distribution of $\bar{y}$ is $N(\mu, \sigma^2/n)$, so $\mathrm{E}[\hat{\mu}] = \mu$ – the MLE of μ is unbiased. The sampling distribution of RSS (independent of that of $\hat{\mu}$) is $\sigma^2\chi^2_{n-1}$, so $\mathrm{E}[RSS] = (n-1)\sigma^2$, and $\mathrm{E}[\hat{\sigma}^2] = (n-1)\sigma^2/n$ – the MLE of σ^2 is slightly biased.

To make an inferential statement about μ that does not depend on σ, we need the t-distribution

$$t = \frac{\sqrt{n}(\bar{y} - \mu)}{s} \sim t_{n-1},$$

the Student's t-distribution with $n-1$ degrees of freedom, where $s = \sqrt{RSS/(n-1)}$. A (central) 95% confidence interval for μ is then $\bar{y} \pm t^{0.975}_{n-1} s/\sqrt{n}$. For our income example, $\bar{y} = 67.1$, $s^2 = 500.87$, and the 95% confidence interval for μ from the t-distribution is [59.9,74.3].

1.5.2 Bayes theory

Bayes theory was dominant (indeed, it was the *only* theory) from the 1800s to the 1920s and began a major resurgence in the 1990s. It is *fully conditional* on the observed data, and conclusions about the population from which it was drawn are based on the likelihood function $L(\lambda \mid \mathbf{y})$ (representing the data information) and the *prior* (probability) *distribution* $\pi(\lambda)$ of the model parameters, representing the information we have about these parameters external to, and in advance of, the sample data.

Inference is expressed through the *posterior* distribution $\pi(\lambda \mid \mathbf{y})$ of the model parameters, *updated* from the prior by the likelihood through Bayes's Theorem:

$$\pi(\lambda \mid \mathbf{y}) = \frac{L(\lambda)\pi(\lambda)}{\int L(\lambda)\pi(\lambda)\mathrm{d}\lambda}.$$

If λ can take just one of the two values λ_1 and λ_2, with prior probabilities π_1 and π_2, the ratio of posterior probabilities (the *posterior odds* for λ_1 to λ_2) is

$$\begin{aligned}\frac{\pi(\lambda_1 \mid \mathbf{y})}{\pi(\lambda_2 \mid \mathbf{y})} &= \frac{L(\lambda_1)\pi_1}{L(\lambda_2)\pi_2} \\ &= \frac{L(\lambda_1)}{L(\lambda_2)} \cdot \frac{\pi_1}{\pi_2},\end{aligned}$$

so that *the posterior odds is equal to the likelihood ratio multiplied by the prior odds*. So the likelihood ratio provides the *data evidence* for one parameter value over another, but this is complemented in Bayes theory by the *prior information* about these values – their prior probabilities.

The theory requires that we express prior information as a probability distribution. In many cases we may not have well-developed information or views that are easily expressed as a probability distribution, and much use is made, by many Bayesians, of *weak* or *noninformative* priors, which are "uniformative" relative to the information in the data: the data were presumably collected to obtain information about parameters for which we had little prior information, and so the prior should reflect this lack of information. A non-informative prior for the case above of two parameter values would be one with equal prior probabilities, leading to the posterior odds being equal to the likelihood ratio.

A subdivision of Bayes theorists regard noninformative priors as at best undesirable (especially when they are improper) and at worst denying the whole point and advantage of the Bayesian approach, which is to accommodate *both* sample data *and* external information in the same unified probabilistic framework. It argues that *all* prior distributions should reflect the actual information available to the analyst; this may mean that different analysts using different prior distributions come to different conclusions. Analysts who have difficulty formulating priors need to be trained in prior *elicitation* (Garthwaite et al. 2005).

To draw conclusions about the parameters of interest θ, we need to eliminate the nuisance parameters ϕ in some way. This is achieved in Bayes theory by a standard probability procedure: we integrate the *joint posterior distribution* $\pi(\theta, \phi \mid \mathbf{y})$ over ϕ to give the *marginal posterior distribution* $\pi(\theta \mid \mathbf{y})$:

$$\pi(\theta \mid \mathbf{y}) = \int \pi(\theta, \phi \mid \mathbf{y})\mathrm{d}\phi.$$

For the income example, we need to specify the prior distribution for (μ, σ). A simple approach is to specify independent flat priors $\pi(\mu) = c$ for μ and for $\log(\sigma)$; the latter is equivalent to the prior on σ of $\pi(\sigma) = 1/\sigma$. For this choice of prior, the joint posterior distribution of μ and σ^2 can be expressed as

$$\pi(\mu, \sigma \mid \mathbf{y}) = c \cdot L(\mu, \sigma \mid \mathbf{y})\pi(\mu, \sigma)$$

$$= c \cdot \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\bar{y}-\mu)^2 + RSS\right]\right\}$$
$$= c' \cdot \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right\}$$
$$\cdot \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \exp\left\{-\frac{RSS}{2\sigma^2}\right\}\left(\frac{RSS}{\sigma^2}\right)^{\nu/2-1},$$

where $\nu = n-1$ and c and c' are integrating constants, not dependent on μ or σ.

So, given σ, μ has a (conditional) posterior $N(\bar{y}, \sigma^2/n)$ distribution, while RSS/σ^2 has a marginal χ^2_ν posterior distribution. Integrating out σ, we obtain the standard result that $t = \sqrt{n}(\bar{y}-\mu)/s$ has a marginal posterior t-distribution with ν degrees of freedom or, equivalently, the mean μ has a shifted and scaled $\bar{y} + t_\nu s/\sqrt{n}$ marginal posterior distribution. So the posterior mean (and median) of μ is $\bar{y} = 67.1$, and a central 95% credible interval for μ is identical to the central 95% confidence interval: $[\bar{y} \pm 2.045 s/\sqrt{n}]$. For the income sample, the 95% credible interval is [59.9,74.3].

The interpretation of the likelihood function for the income example is *reversed* compared with the frequentist interpretation: $\bar{y}$ and *RSS* are now fixed constants given the data, while the parameters μ and σ have *noninformative* prior distributions and consequent posterior distributions. The Bayesian posterior distributions depend on the particular diffuse priors for μ and $\log\sigma$: if conjugate *informative* priors are used, the posterior distributions are more informative than the frequentist results.

Our work on the NAEP surveys does not use Bayesian methods, though we raise, in the final discussion section of the book, serious difficulties with frequentist methods for these surveys, which require fully Bayesian methods for satisfactory analyses.

1.5.3 "Model assisted" survey sampling theory

The term *model assisted* (as used in Särndal et al. 1992) is relatively new in survey sampling theory, which was extensively developed in the 1950s. It refers to the usefulness of model-based *estimators* of model parameters, but without reliance on the *correctness* of the model. Without a formal model, inference is based on the repeated sampling distribution of the *sample selection indicators*, not of the population values themselves. The *survey design* determines the inference, hence the term *design-based* (as opposed to *model-based*) inference. The design-based approach, and several aspects of the model-based approach, are clearly set out in the book by Lohr (1999), which we follow in our discussion below.

The income mean example gives a simple example of the approach. We introduce notation for the finite population. Y is the variable of interest, in the finite population of size N. The population values of Y are $Y_1, Y_2, ..., Y_N$. The population mean μ is

$$\mu = \sum_{I=1}^{N} Y_I / N$$

and the population variance is

$$\sigma^2 = \sum_{I=1}^{N} (Y_I - \mu)^2 / N;$$

both the mean and variance must be finite if the values Y_I are finite. (In the survey literature, the variance denominator is usually $N-1$.)

We draw a simple random sample without replacement of fixed predetermined size n and obtain observed values $y_1, ..., y_n$ with sample mean $\bar{y}$ and sample variance

$$s^2 = \sum_i (y_i - \bar{y})^2 / \nu.$$

Define *indicator variables* $Z_1, Z_2, ..., Z_I, ..., Z_N$. Let

$$\begin{aligned} Z_I &= 1 \text{ if population member } I \text{ is selected} \\ &= 0 \text{ if population member } I \text{ is not selected.} \end{aligned}$$

Then

$$\begin{aligned} \bar{y} &= \sum_{i=1}^{n} y_i / n \\ &= \sum_{I=1}^{N} Z_I Y_I / \sum_{I=1}^{N} Z_I \\ &= \sum_{I=1}^{N} Z_I Y_I / n. \end{aligned}$$

Inference about μ is based on the *repeated sampling properties of the random variable* $\bar{y}$ *as an estimator of* μ. The fundamental inferential principles are that the Y_I are *fixed constants* and the Z_I are Bernoulli *random variables* with

$$\begin{aligned} \Pr[Z_I = 1] &= \frac{\text{no. of samples containing unit } I}{\text{no. of samples of size } n} \\ &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} = \pi, \end{aligned}$$

the *sampling fraction*. The properties of $\bar{y}$ are easily established.

$$\mathrm{E}[Z_I] = \mathrm{E}[Z_I^2] = \pi, \mathrm{Var}[Z_I] = \pi(1-\pi) = \frac{n}{N}\left(1 - \frac{n}{N}\right).$$

Hence

$$\begin{aligned}\mathrm{E}[\bar{y}] &= \frac{1}{n}\sum_{I=1}^{N}\mathrm{E}[Z_I]Y_I \\ &= \frac{1}{N}\sum_{I=1}^{N}Y_I \\ &= \mu.\end{aligned}$$

So, as a random variable, $\bar{y}$ is *unbiased* for μ. For the variance of $\bar{y}$, we need the joint distribution of pairs of the Z_I. These are not independent,

$$\begin{aligned}\Pr[Z_I = 1, Z_J = 1] &= \Pr[Z_I = 1]\Pr[Z_J = 1 \mid Z_I = 1] \\ &= \frac{n}{N}\cdot\frac{n-1}{N-1},\end{aligned}$$

and so

$$\begin{aligned}\mathrm{Cov}[Z_I, Z_J] &= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{1}{N-1}\frac{n}{N}\left(1-\frac{n}{N}\right) \\ &= -\pi(1-\pi)/(N-1), \\ \mathrm{Var}[\bar{y}] &= \sum_I Y_I^2\mathrm{Var}[Z_I]/n^2 + \sum_I\sum_{\neq J} Y_I Y_J \mathrm{Cov}[Z_I, Z_J]/n^2 \\ &= \frac{1-n/N}{nN(N-1)}\left[(N-1)\sum_I Y_I^2 - \sum_I\sum_{\neq J} Y_I Y_J\right] \\ &= \frac{1-n/N}{n(N-1)}\sum_I (Y_I-\mu)^2 \\ &= (1-n/N)\sigma^2/n \\ &= (1-\pi)\sigma^2/n,\end{aligned}$$

if the population variance is defined by the $N-1$ denominator.

The first term, $(1-n/N) = (1-\pi)$, is a *finite population correction* (FPC): for a small sample fraction, $\mathrm{Var}[\bar{y}] \simeq \sigma^2/n$, but as $\pi \to 1$, $n \to N$, and $\mathrm{Var}[\bar{y}] \to 0$, since the sample exhausts the population. For our income example, the FPC is $(1-40/648) = 0.938$ – the variance of the sample mean is reduced by 6.2% relative to the usual frequentist variance.

For the sample variance,

$$\begin{aligned}\mathrm{E}[s^2] &= \mathrm{E}\left[\sum_i y_i^2 - n\bar{y}^2\right]/(n-1) \\ &= \mathrm{E}\left[\sum_I Z_I Y_I^2 - n\bar{y}^2\right]/(n-1)\end{aligned}$$

$$
\begin{aligned}
&= \left\{ n \sum_I Y_I^2/N - n(\mathrm{Var}[\bar{y}] + \mathrm{E}\bar{y}]^2) \right\} / (n-1) \\
&= \left\{ n \sum_I Y_I^2/N - n([1-1/N]\sigma^2/n + \mu^2) \right\} / (n-1) \\
&= (1-1/N)\sigma^2.
\end{aligned}
$$

So s^2 is an almost unbiased estimator of σ^2, regardless of any distribution model for Y, and under the Bernoulli model $\bar{y}$ is the *minimum variance linear unbiased estimator of* μ.

For confidence interval statements about μ, the theory uses the Central Limit Theorem in its general form. The sample mean has expectation μ and variance $(1-1/N)\sigma^2$ in repeated samping, and since it is a (weighted) linear combination of (correlated) random variables Z_I, as $n \to \infty$ (and $N \to \infty$), the sampling distributions of

$$
z = \frac{\sqrt{n}(\bar{y}-\mu)}{\sigma} \text{ and } t = \frac{\sqrt{n}(\bar{y}-\mu)}{s} \to N(0,1),
$$

giving the usual large-sample confidence interval

$$
\bar{y} - z_{1-\alpha/2} s/\sqrt{n} < \mu < \bar{y} + z_{1-\alpha/2} s/\sqrt{n}.
$$

The accuracy of the confidence interval coverage depends on the sample size n – it may be quite inaccurate for small n – and may depend on other properties of the Y population. Without other information about this population, we cannot say more.

For the example, we have $\bar{y} = 67.1$, $s = 22.4$, and the (approximate) 95% confidence interval for the population mean is [60.2, 74.0]. Remarkably, we seem to be able to make (asymptotically) the same inferential statement about μ without *any* model for the population values Y, or invoking the Central Limit Theorem for the sampling distribution of $\bar{y}$ as a function of $(y_1, \ldots, y_n)$!

From the viewpoint of model-based likelihood theory, however, this approach is unsatisfactory. The argument is clear if we construct the likelihood as the probability of *all* observed data. The data are *both* the sample selection indicators Z_I *and* the observed response variables y_i for the selected population members, so we need a population model for the Y_I as well.

The fundamental probability relation we use is

$$
\begin{aligned}
\Pr[Y_I, Z_I] &= \Pr[Z_I \mid Y_I] \Pr[Y_I] \\
&= \Pr[Y_I \mid Z_I] \Pr[Z_I].
\end{aligned}
$$

where $\Pr[Z_I \mid Y_I]$ is the *sample selection model* for Z – it specifies how the selection probability of population member I depends on the value of the response Y_I for that member – and $\Pr[Y_I]$ is the *population model* for Y. For simple random sampling,

$$
\Pr[Z_I \mid Y_I] = \Pr[Z_I] = \pi^{Z_I}(1-\pi)^{1-Z_I},
$$

the Bernoulli model above. Correspondingly,

$$\Pr[Y_I \mid Z_I] = \Pr[Y_I]$$

– the model for the selected population values is the same as that for the unselected values. So

$$\Pr[Y_I, Z_I] = \Pr[Y_I] \Pr[Z_I],$$

and the likelihood is

$$\begin{aligned} L &= \Pr[y_1, \ldots, y_n] \cdot \Pr[Z_1, ..., Z_N] \\ &= \Pr[y_1, \ldots, y_n] \cdot \frac{1}{\binom{N}{n}} \\ &= \Pr[y_1, \ldots, y_n] \cdot \frac{n}{N} \frac{n-1}{N-1} \cdots \frac{1}{N-n+1}. \end{aligned}$$

The last term in the selection probabilities is completely known from the design – it is just a constant. For inferential statements about the parameters based on likelihoods and their *ratios*, as with known constants in any probability function for the Y_I, these constant terms are *irrelevant* (they *cancel* in likelihood ratios). Thus, regardless of the kind of model we might have for the Y_I, any inference through likelihood ratios does not depend on the sample design if this is *noninformative* in the sense described above – that $\Pr[Z_I \mid Y_I] = \Pr[Z_I]$ – membership of the I-th population member in the sample does not depend on the value of the response Y_I.

In survey sampling theory, this difficulty is countered by the difficulty of the dependence of model-based inference on the correctness of the model – if this is incorrect, the conclusions from the analysis could be wrong. Since every model is by definition wrong (as it is a simplification), the risk of wrong conclusions from the model-based approach is inherent in the approach. Also, if the sample design is *informative*, likelihood-based inference becomes much more difficult because the form of dependence in $\Pr[Z_I \mid Y_I]$ needs to be specified and included in the likelihood.

In Aitkin (2008) and Chapter 4 of Aitkin (2010), this argument is addressed by using a general multinomial model for the population values Y_I. This allows full likelihood and Bayes analyses but makes no restrictive smooth assumptions – the "model" is *always* correct!

In the NAEP analyses following this chapter, our response variables are *binary* test items with a correct answer probability that is modeled using a logistic regression function. So there is no formal "model" assumption (like a normal distribution) for the test item responses – it is only the form of the "link" function, relating the probability to covariates, which might be questioned as appropriate. We discuss this further in the following chapters.

1.6 Weighting

The importance of the difference in model-based and design-based philosophies is greatly increased in dealing with *weighting*. The use of sample weighting in stratified designs has a long tradition in survey sampling, dating at least from Klein and Morgan (1951). DuMouchel and Duncan (1983) gave a summary of textbook recommendations for weighting and asked (p. 535):

> Which estimator should be used? Controversy has raged since Klein and Morgan (1951).

Recent formalisations of weighting through *weighted likelihoods* clarify this issue. We first discuss stratified sampling.

1.6.1 Stratified random sampling

Stratified sampling is used extensively in NAEP survey designs. It is designed to reduce variability in estimation due to known population heterogeneity – the population is made up of homogeneous subpopulations with substantial differences among them. A simple random sample may miss the small subpopulations completely or give only small subsamples from them.

So the population is stratified into H subpopulations or *strata* with subpopulation sizes $N_h, h = 1,...,H$, and random samples of sizes n_h are drawn separately from each stratum. Extending the earlier notation, we write the responses of the population members in stratum h as $Y_{Ih}, I = 1,...,N_h$, and those of the sample members in stratum h as $y_{ih}, i = 1,...,n_h$. The rare strata are usually *oversampled* to obtain reliable sample sizes. In estimating the full population mean, this oversampling has to be taken into account. We call $\pi_h = n_h/N_h$ the *sampling fraction* in stratum h and $p_h = N_h/N$ the *population proportion* in stratum h. The stratum means and variances are defined correspondingly:

$$\mu_h = \sum_{I=1}^{N_h} Y_{Ih}/N_h,$$

$$\sigma_h^2 = \sum_{I=1}^{N_h} (Y_{Ih} - \mu_h)^2/N_h,$$

$$\bar{y_h} = \sum_{i=1}^{n_h} y_{ih}/n_h,$$

$$s_h^2 = \sum_{i=1}^{n_h} (y_{ih} - \bar{y_h})^2/(n_h - 1).$$

These are summarised in Table 1 below.

The full population mean is

$$\mu = \sum_{h=1}^{H} \sum_{I=1}^{N_h} Y_{Ih}/N = \sum_{h=1}^{H} N_h \mu_h / N = \sum_{h=1}^{H} p_h \mu_h,$$

and the full population variance is

$$\begin{aligned}
\sigma^2 &= \sum_{h=1}^{H} \sum_{I=1}^{N_h} (Y_{Ih} - \mu)^2 / N \\
&= \sum_{h=1}^{H} \left[\sum_{I=1}^{N_h} (Y_{Ih} - \mu_h)^2 + N_h (\mu_h - \mu)^2 \right] / N \\
&= \sum_{h=1}^{H} N_h [\sigma_h^2 + (\mu_h - \mu)^2] / N \\
&= \sum_{h=1}^{H} p_h [\sigma_h^2 + (\mu_h - \mu)^2].
\end{aligned}$$

The overall sample mean is

$$\bar{y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} y_{ih}/n = \sum_{h=1}^{H} n_h \bar{y}_h / n,$$

and the overall sample variance is

$$\begin{aligned}
s^2 &= \sum_{h=1}^{H} \sum_{i=1}^{n_h} (y_{ih} - \bar{y})^2 / (n-1) \\
&= \sum_{h=1}^{H} [(n_h - 1) s_h^2 + n_h (\bar{y}_h - \bar{y})^2] / (n-1).
\end{aligned}$$

1.6.2 Design-based analysis

From §1.5.3, each $\bar{y}_h$ is an unbiased estimator of μ_h with variance σ_h^2/n_h (omitting the FPC), so an unbiased estimator of μ is $\tilde{\mu} = \sum_h p_h \bar{y}_h$, with variance (ignoring the FPC) estimated by

$$\tilde{V}(\tilde{\mu}) = \sum_h p_h^2 s_h^2 / n_h.$$

Suppose we used the "simple-minded" estimator $\bar{y}$ to estimate μ. Then

Table 1.1 Stratified sampling

	Stratum	Proportion	Fraction	Mean	Variance
Population	h	$p_h = N_h/N$		μ_h	σ_h^2
Sample		n_h/n	$\pi_h = n_h/N_h$	$\bar{y}_h$	s_h^2

$$\mathrm{E}[\bar{y}] = \sum_h n_h \mu_h / n,$$

and this is *not* equivalent to μ unless $n_h/n = p_h = N_h/N$, which is equivalent to $n_h/N_h = \pi_h = n/N$, that is, the stratum sampling fraction is *constant*, as in probability proportional to size (PPS) sampling. Since a constant sampling fraction would give small sample sizes for small strata, it would not generally be used for stratified sampling, so $\bar{y}$ is generally *biased*. However, if $\mu_h \equiv \mu$, then $\bar{y}$ is *unbiased*, regardless of the sampling fractions – we are back to a homogeneous (in the means) population.

1.6.3 Model-based analysis

The advantage of the fully model-based approach is that it requires us to specify and include in the model all relevant features of the data process. What "simple-minded" model would lead to the unreasonable analysis above? For the homogeneous population and simple random sampling, the model $Y_i \sim N(\mu, \sigma^2)$ would lead to $\hat{\mu} = \bar{y}$. But we are given a heterogeneous population with different strata means μ_h, so this cannot be the correct model. We need the *one-way classification model* (with stratum-specific variances)

$$Y_{ih} \sim N(\mu_h, \sigma_h^2),\ i = 1, ..., n_h,\ h = 1, ..., H,$$

for which

$$\hat{\mu}_h = \bar{y}_h,\ \hat{\sigma}_h^2 = \frac{n_h - 1}{n_h} s_h^2.$$

To make a model-based inference about the linear function $\mu = \sum_h p_h \mu_h$, we use linear random variable theory:

$$\hat{\mu} = \sum_h p_h \bar{y}_h \sim N\left(\sum_h p_h \mu_h, \sum_h p_h^2 \sigma_h^2 / n_h\right).$$

So, by including the strata as a *factor* in the model – *modeling the stratifying factor* – we obtain the same model-based estimate and sampling variance as those from the design-based estimate: since the stratified design is used because of the population heterogeneity, we use the stratifying factor in any model and estimate the heterogeneous population mean from the appropriate weighting of the strata means.

This approach resolves a standard criticism of model-based analysis from the design-based viewpoint: that it ignores, for estimation of the population mean in stratified sampling, the need to weight the strata sample means. Instead of weighting

the observations, we model the strata and then simply weight the strata means for inference about the population mean.[2]

The criticism above may have had some validity in the 1970s, before the development of multilevel modeling in the early 1980s (Aitkin, Anderson, and Hinde 1981; Aitkin, Bennett, and Hesketh 1981), but the fully model-based analysis incorporating *both* stratification and clustering has been available, and has been widely used, since the 1990s.

1.6.4 Weighted likelihoods

In the design-based framework, the use of weights in analysis is generally recommended far beyond simple population mean estimation, extending to multiple regression and other model-based procedures. The extension is motivated by its expression in a model-assisted analysis through a *weighted likelihood* (Skinner 1989, pp. 80–84). The idea is that, since each sampled individual i in stratum h "represents" w_h individuals in that population stratum, the contribution of individual i to the analysis should be represented through a formal weight attached to the individual i's response y_{ih}. This is formalised in a weighted likelihood: the weighted or *pseudo-likelihood* is defined by

$$L^*(\theta) = PSL(\theta) = \prod_h \prod_i f^{w_h}(y_{ih} \mid \theta),$$

where θ are the model parameters. Any model-based analysis could then maximise the log-pseudo-likelihood

$$\ell^* = \log L^*(\theta) = \sum_h \sum_i w_h \log f(y_{ih} \mid \theta),$$

in which w_h appears as an explicit weight for observations in stratum h in the likelihood equations for the maximum likelihood estimates. So, for example, for the simple one-population mean model $Y_{ih} \sim N(\mu, \sigma^2)$ but under stratified sampling, the weighted log-likelihood (omitting constants) would be

$$\log \ell^* = \sum_h \sum_i w_h \left[-\log \sigma - \frac{1}{2\sigma^2}(y_{ih} - \mu)^2) \right].$$

If we treat this as a regular log-likelihood, the first and second derivatives would give the maximum pseudo-likelihood (PML) estimate $\widetilde{\mu}$ and its large-sample variance as

$$\widetilde{\mu} = \sum_h p_h \bar{y}_h,$$

[2] Our NAEP analysis is focused on *regression parameters*; the overall population achievement mean on the NAEP scale is not directly identifiable but is set by a scaling procedure – see Chapter 2, §5.6.

$$\mathrm{Var}[\widetilde{\mu}] = \sigma^2/N.$$

So the PML estimate of μ from the one-population mean likelihood is correct. However, the sampling variance expression for $\widetilde{\mu}$ from weighting does *not* correspond to the survey sampling (and the weighted one-way classification model-based) variance estimate $\sum_h p_h^2 s_h^2/n_h$ – it is generally much smaller. This disconcerting result follows from the obvious fact that each stratum sample mean is being treated as though it were based on a very much larger sample size, the stratum *population* size, instead of the *sample* size, summing to the full population size. It is clear that this approach does not provide correct standard errors, even if it provides an unbiased estimate of the population mean. Weighting the likelihood for the one-way classification model has the same effect; we omit the details.

In more complex regression models in which continuous covariates are used, with common slopes across strata for each variable, both the standard errors of the estimated regression coefficients *and the estimates themselves* are biased away from those of the correct likelihood. For example, for a single continuous variable in the stratified design, with the common-slope model

$$Y_{ih} \mid x_{ih} \sim N(\beta_{0h} + \beta_1 x_{ih}, \sigma^2),$$

the MLE of β_1 is

$$\hat{\beta}_1 = \frac{\sum_h \sum_i (y_{ih} - \bar{y}_h)(x_{ih} - \bar{x}_h)}{\sum_h \sum_i (x_{ih} - \bar{x}_h)^2},$$

which has sampling variance $\sigma^2 / \sum_h \sum_i (x_{ih} - \bar{x}_h)^2$, while the PML estimate

$$\widetilde{\beta}_1 = \frac{\sum_h \sum_i w_{ih}(y_{ih} - \bar{y}_h)(x_{ih} - \bar{x}_h)}{\sum_h \sum_i w_{ih}(x_{ih} - \bar{x}_h)^2}$$

has sampling variance $\sigma^2 / \sum_h \sum_i w_{ih}(x_{ih} - \bar{x}_h)^2$. While the PML estimate is unbiased, the contributions to its numerator and denominator will be more strongly determined by the undersampled strata and less strongly by the oversampled strata than those for the MLE. This can lead to substantial differences from the MLE. The variances will be affected in the same way and will be far too small if the sampling weights are used directly as observation weights.

Weighted likelihoods were extended to two-level models in Pfefferman et al. (1998) in a complex analysis. Sampling rates varied at both levels of the two-level model, requiring double weighting of each observation, though the weights advocated had to be estimated in a complex process and were not simply the product of the weights at the two levels.

It is sometimes proposed that the survey weights be scaled to give "pseudo sample sizes" (our term) summing to the overall sample size, that is, the weights would be defined by $w_h^* = w_h n_h / \sum_h w_h n_h = N_h/N$, with pseudo sample sizes $n_h^* = n \cdot w_h^*$. This would reduce the incorrect scaling of variances, but would still lead to incorrect and unreasonable variances of model parameters in a pseudo-likelihood analysis. The scaled weighted analysis corresponds to the sample sizes that we *would*

have obtained (in expectation) had the sample been a PPS sample with a common sampling rate across strata. But we do *not* have this sample design, and the pseudo-likelihood *is not the likelihood of the observed data* – it refers to a design that was not used, and is irrelevant to the analysis of the design that *was* used.

In his discussion of Pfefferman et al., Chambers (1998) raised the same question that also determines our approach to stratification and weighting – his comments apply equally well to single-level survey designs. If weighting is only needed for informative designs (which are rare), and results in a loss of efficiency in any case relative to an unweighted model-based analysis that includes the stratification factor in the model, why is weighting recommended for *all* designs?

Sample survey textbooks emphasise that weights should be used *only for point estimation of parameters and must not be used to compute standard errors*. Other non-model-based methods have to be used for this (see, for example, Lohr 1999, pp. 104, 226, 234, 367). So the pseudo-likelihood approach, which appears to be model-based, cannot use the usual model-based variance estimation procedure but has to rely on non-model-based methods for variances. It is therefore difficult to understand why the method of pseudo- or weighted likelihoods is recommended so frequently, or at all.

1.7 Missing data and non-response

The discussion in this section is based on the authoritative books by Rubin (1987) and Little and Rubin (1987). Suppose that after we draw the sample of size n we do not get a response – the family income in our first example – from m sampled individuals. Define another response indicator $R_i = 1$ if selected person i responds and $R_i = 0$ if there is no response, for $i = 1, ..., n$. This indicator is defined only for the sampled population members.

Consider the joint distribution

$$\Pr[Y_i, R_i] = \Pr[R_i \mid Y_i] \Pr[Y_i].$$

This is needed to construct the full likelihood of *all* random quantities, as in the sample selection process. Here $\Pr[R_i \mid Y_i]$ is the *response model* and $\Pr[Y_i]$ is, as before, the population model.

If $\Pr[R_i = 1 \mid Y_i] = \Pr[R_i = 1] = p_i$ independently of Y_i and its model parameters, we have (response) *missingness at random (MAR)*. The missingness probability p_i may *vary across individuals i*, but it is *uninformative* about the response. If $p_i = p \,\forall\, i$, a constant, then we have *missingness completely at random (MCAR)*.

The observed data are then

$$(y_1, r_1), ..., (y_m, r_m), r_{m+1}, ..., r_n,$$

with

$$r_1 = ... = r_m = 1, r_{m+1} = ... = r_n = 0,$$

and likelihood

$$L(\theta) = \Pr[y_1 \mid \theta] \ldots \Pr[y_m \mid \theta] \cdot p_1 \ldots p_m (1 - p_{m+1}) \ldots (1 - p_n).$$

It is immediately clear again that for inference based on the likelihood and likelihood ratios, the missingness response model is *irrelevant* since it is independent of the responses y_i that might have been observed and the response model parameters θ. Then the observed data $y_1, \ldots, y_m$ are a *random subsample* of $y_1, \ldots, y_n$ (but are not a simple random sample unless $p_i = p \forall i$) and can be analysed as a complete sample for inference about θ.

However, if $\Pr[R_i = 1 \mid Y_i]$ depends on Y_i, we have a *biased sampling mechanism* – the missing data are *missing nonrandomly* (MNR). For example, the probability of response might decrease with increasing income. Then high incomes would be under-represented in the sample and the sample mean would be biased downward. If the biased sampling mechanism (that is, the probability model for missingness) is *known*, then it can be included in the full likelihood and allowance made for it. In general, however, we cannot *estimate* such a nonresponse model *from the observed data* – we lack just the data that would establish it! *Sensitivity analysis* is the only way of assessing the possible effect of nonrandom missingness on the inferential conclusions.

1.7.1 Weighting adjustments for nonresponse

It is common practice in survey sampling analysis to make *weighting adjustments* to population or other estimates to "correct" for nonresponse. Suppose, in our income example, that a random sample of 45 families was originally drawn but the incomes for five families were not recorded. If this was an administrative error, not related to the family incomes, then for these five families income is MCAR. If missingness is related to the father's age, but not income, then income is MAR. What effect does this have on the analysis?

In the model-based framework, with either MAR or MCAR, the *achieved* sample of 40 families is treated as if it had been the *target* sample. The MAR or MCAR missingness for the five families simply reduces the sample size.

In the survey sampling framework, a second set of "response" weights is usually defined, the inverse of the probabilities of response, and the overall weight attached to an individual is the product of the selection weight and the response weight. As described above, one should *not* use these as formal weights in any regression but use the stratification variable as a model factor. No such factor needs to be defined for MAR or MCAR nonresponse.

1.7.2 Incomplete data in regression

A common feature of large-scale data is incompleteness – missing values scattered across the data matrix of y_i and $\mathbf{x}_i$ values. For regression modeling, missing response values y_i cause no difficulty – the corresponding cases do not contribute to the likelihood and so can be omitted from analysis, as in the single-sample case above. Missing values in some covariates $\mathbf{x}_i$, however, do not render the corresponding cases valueless, so long as they are MAR or MCAR – they still provide information about the response, though it is more difficult to extract it.

An important point in this case is that *we do not need a model for the missingness process* – even if this is MAR, with missingness dependent on other observed variables, the missingness model *cancels* in any likelihood ratios and so *does not need to be explicitly fitted* (though it may be of interest to *describe* the missingness process).

The simplest example is of a single-variable normal regression model in which the cases $i = 1,...,m$ are complete and the cases $i = m+1,...,n$ have missing x_i but observed y_i. The usual model for $y \mid x$ is

$$Y_i \mid x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

However, we are not given x_i for the incomplete cases, only y_i. For these cases to contribute to the likelihood, we have to use the *marginal* distribution of y_i, integrating out x_i. For this we need the *marginal distribution of* x, $f(x)$ – an additional model assumption.

The usual assumption (for a continuous x variable) is that X is normal, with mean μ_X and variance σ_X^2, which are estimated from the observed x sample. The marginal distribution of Y is then $N(\beta_0 + \beta_1\mu_X, \sigma^2 + \beta_1^2\sigma_X^2)$. Full maximum likelihood is achieved using an EM algorithm, which requires the conditional distribution of the missing X values given the observed Y; this is tractable only for the case of normal or categorical X.

For the general case of mixed continuous and categorical variables in $\mathbf{x}$, an EM algorithm approach was developed by Little and Schluchter (1985) when the continuous variables were multivariate normal; these could have a multivariate linear model regression on the categorical variables, which themselves had a log-linear model. Experience with the EM algorithm for the categorical or mixed $\mathbf{x}$ case showed that convergence was generally very slow.

1.7.3 Multiple imputation

These methods are not currently used in many practical data analyses. Instead, *imputation methods* are used, more or less widely. *Multiple imputation* (MI) methods are based on the *data augmentation algorithm*, a Bayesian stochastic version of the EM algorithm in which Markov chain Monte Carlo methods are used to obtain con-

vergence of both the parameters to their full posterior distribution and the missing data to their full conditional distribution given the observed data.

The MI approach, when first proposed, used a small number M of random draws from the converged posterior distribution of the missing data. These draws are used to produce M "completed" *imputed data sets*, which are then analysed in the standard maximum likelihood framework, and the M sets of estimates and standard errors are combined into a single set of estimates and standard errors by standard methods. These estimates should correspond closely to those obtained from the fully Bayesian analysis (which generated them!), in terms of agreement between posterior means and averaged maximum likelihood estimates, and corresponding variability estimates.

In early uses of MI, M was sometimes as small as 5. This was partly due to the computer limitations at the time, and partly due to the high efficiency of quite small values of M relative to $M = \infty$. Current computational power allows much larger values of M, and it should be noted that for $M = 10000$ the parameter draws effectively define the full posterior distribution of the model parameters so that no ML analyses are needed – the (Bayesian posterior) results are already available.

In practical uses of the MI approach, the response and covariates are considered as a joint multivariate set, with missing values across the set, and considerable emphasis is placed on developing a *rich imputation model* – that is, one with all possibly relevant variables included in the model for the conditional distribution of any variables with missing values, given the other variables. The joint multivariate distribution is often treated as a multivariate normal (Schafer 1997) – this greatly accelerates convergence and seems to provide good estimates of parameters and standard errors.

In theory, the methods above can be extended to missing data in more complex survey designs. Clustering is an important extension in nearly all surveys, but the great complexity of the covariance structure, for incomplete data with missingness at all levels of the design, makes this extension very difficult. This leads to difficulties with incomplete questionnaire data in NAEP surveys, as will be seen in later chapters.

Chapter 2
The Current Design and Analysis

2.1 NCES and NAEP

As we stated in the Preface, the work described in this book was supported by NCES, though they are not responsible for the views and opinions expressed in the book, which should not be interpreted as a statement of Department of Education policy. For those readers not familiar with the role of NCES in the design and analysis of the National Assessment of Educational Progress (NAEP), we quote below at considerable length from several NCES publications describing aspects of NCES and the NAEP surveys that were important in our work. From the NCES Website,

> The National Center for Education Statistics (NCES) is the primary federal entity for collecting and analyzing data related to education in the U.S. and other nations. NCES is located within the U.S. Department of Education and the Institute of Education Sciences.
>
> ...The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Assessments are conducted periodically in mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history.
>
> ...NAEP provides results on subject-matter achievement, instructional experiences, and school environment for populations of students (e.g., all fourth-graders) and groups within those populations (e.g., female students, Hispanic students). NAEP does not provide scores for individual students or schools, although state NAEP can report results by selected large urban districts. NAEP results are based on representative samples of students at grades 4, 8, and 12 for the main assessments, or samples of students at ages 9, 13, or 17 years for the long-term trend assessments. These grades and ages were chosen because they represent critical junctures in academic achievement.

The design and analysis of NAEP surveys are described concisely but comprehensively in Chapter 20 of the NCES Handbook of Survey Methods Technical Report (U.S. Department of Education, Institute of Education Sciences, NCES 2003-603), which we refer to in the text as "Handbook". We quote from it extensively. The design varied over time, and we add to or amend (in square brackets [ ]) the 2003 report where necessary for the 1986 survey.

Much greater detail is given in the NAEP Technical Report series; we quote when needed from the Johnson and Zwick (1988) report, referred to in the text as "1988

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_2, © Murray Aitkin and Irit Aitkin 2011

Technical Report", and from the Beaton et al. (1986) report, referred to in the text as "1986 Technical Report". These reports are necessarily technical; a less technical but clear exposition of the design issues in the NAEP and their current analysis can be found in Chapter 7 of Longford (1995).

The publication of NAEP Technical Reports ended with the 1998 report; the NCES Handbook of Survey Methods Technical Report 2003 mentioned above is less detailed and does not give any survey results, but covers all the NCES surveys, not just the NAEP. A Website is now provided for very detailed technical documentation:

```
http://nces.ed.gov/nationsreportcard/tdw/
```

Publication of sample sizes of schools and students is restricted, for the surveys using restricted data, by rounding them to the nearest 10. This affects some basic survey details in Chapters 7 and 8. Footnotes are given for these roundings; the relevant policy may be found in the Statistical Standards Program on the IES Website:

```
http://nces.ed.gov/statprog/instruct_respdata.asp?
resptype=sub
```

Rounding has also been applied to the sample sizes in Chapters 5 and 6, though these chapters report unrestricted public-access data (1986 was the last survey year of the unrestricted provision of NAEP data).

2.2 Design

The NAEP surveys use a multistage clustered and stratified design.

2.2.1 PSUs

> In the first stage of sampling, the United States (the 50 states and the District of Columbia) is divided into [94] geographic (PSUs) [primary sampling units]. The PSUs are classified into four regions (Northeast, Southeast, Central and West) each containing about one-fourth of the US population. In each region, PSUs are additionally classified as metropolitan or non-metropolitan, resulting in [12] subuniverses of PSUs.
>
> For the [1986] assessment, ... [34] of these PSUs were designated as certainty units because of their size. Within each major stratum (subuniverse), further stratification was achieved by ordering the noncertainty PSUs according to several additional socioeconomic characteristics. ... One PSU was selected from each of the ... noncertainty strata, with probability proportional to size. ... To enlarge the samples of Black and Hispanic students, thereby enhancing the reliability of estimates for these groups, PSUs from the high-minority strata were sampled at twice the rate of PSUs from the other strata. (Handbook, p. 192)

2.2.2 Schools

> In the second stage of sampling, public schools (including Bureau of Indian Affairs ... and Department of Defense ... schools) and nonpublic schools (including Catholic schools), within each of the [94] PSUs are listed. ...
>
> ... [T]he schools within each PSU are assigned a probability of selection that is proportional to the number of students per grade in each school. ... Nonpublic schools and schools with high minority enrolment are oversampled. (p. 192)

We do not discuss here the further stage of random sampling for each school "session" (Handbook, p. 192); pilot testing sessions included trial items on the test, and the students in these sessions are not assessed. They are *missing by design* from the overall analysis. This reduces the effective sample size but has no other effect.

2.2.3 Students

> To facilitate the sampling of students, a consolidated list is prepared for each school of all grade-eligible and age-eligible students. ... A systematic selection of eligible students is made from this list – unless all students are to be assessed – to provide the target sample size. ... (Handbook, p. 193).

It is assumed in the NCES analyses that the systematic selection provides a *random* sample of eligible students.

In addition to the test item responses from each student, information was collected for each student on demographics and family background from schools and principals and from the teachers of a sample of students. The teacher and school data were not used in our analyses of the 1986 data, only a small subset of the demographic and family background variables and the test items presented to each student.

2.2.4 Test items

The test items were assigned to test booklets using a balanced incomplete block (BIB) spiraling process since the number of test items needed for a comprehensive test greatly exceeded the school testing time available. This BIB design of test booklets, and the assignment process of booklets to schools, effectively gave to each student a *random sample of the test items*. The allowance in the analysis for this further sampling of test items is discussed in the Analysis section.

2.2.5 *Important design issues*

Several aspects of the design complicate the analysis.

- stratification and oversampling of high-minority PSUs and schools;
- multistage sampling of students within schools within PSUs.

These design processes have to be allowed for in the analysis. An important point in the model-based analysis is that it is able to *allow fully for the multistage cluster design, which the current analysis could not.* This issue is discussed in detail in later sections in this chapter and in Chapter 5.

2.3 NAEP state sample design 2002+

The design was changed for the state NAEP assessments in 2002 and later surveys. The previous national NAEP design gave small samples in the small states, and states wishing to have state-wide information on their students' progress needed to have larger state samples. This resulted in the states becoming the primary sampling units, and the design within the state became a two-stage sample of schools and students within schools. The national NAEP sample became a subset of the state NAEP samples (except for those states not participating in the state surveys). This reduces the complexity of the analysis and also allows for cross-state comparisons of model structures. In our analysis of the 2005 national NAEP math survey, the California and Texas samples are analysed separately.

We comment in Chapter 9 on the possibility of *linking* the state analyses.

2.4 Weighting

> The weighting ... reflects the probability of selection for each student in the sample, adjusted for school and student non-response. The weight assigned to a student's responses is the inverse of the probability that the student would be selected for the sample. Through poststratification, the weighting ensures that the representation of certain sub-populations corresponds to figures from the U.S. Census and the Current Population Survey (CPS). (Handbook, p. 197)

Student base weights.

> The base weight assigned to a student is the reciprocal of the probability that the student was selected for a particular assessment. This probability is the product of the following four factors:
>
> - the probability that the PSU was selected;
> - the conditional probability that the school was selected, given the PSU;

- the conditional probability, given the selected schools in the PSU, that the school was allocated the ... assessment [rather than a trial item session]; and
- the conditional probability, given the school, that the student was selected for the assessment. (Handbook, p. 197)

Nonresponse adjustments of base weights.

The base weight for a selected student is adjusted by two nonresponse factors. The first factor adjusts for sessions that were not conducted. This factor is computed separately within classes formed by the first three digits of the PSU strata. ... The second factor adjusts for students who failed to appear in the scheduled session or makeup session. ... [T]he adjustment classes are based on subuniverse, modal grade status and race class. In some cases, nonresponse classes are collapsed into one to improve the stability of the adjustment factors. (Handbook, pp. 197–198)

2.4.1 Design effect corrections

Because NAEP uses complex sampling procedures, a jackknife replication procedure is used to estimate standard errors. (Handbook, p. 200)

NAEP's jackknife variance estimator is designed for the situation where the first-stage units, or appropriate aggregates of them, are paired within strata. It estimates the sampling variability of any statistic as the sum of components of variability that may be attributed to each of the jackknife pairs. The variance attributed to a particular jackknife pair is measured by estimating how much the value of the statistic would change if the information embodied in the jackknife pair were to be changed. This is done by the computation of a quantity t_i called a pseudoreplicate, which is associated with the ith jackknife pair, and which is an estimate of the statistic of interest t based on an altered sample. Specifically, the ith pseudoreplicate of the statistic t is created by randomly designating the half-sample members of the pair as first and second, eliminating the data from the first half-sample of the pair, replacing the lost information with that from the second half-sample of the pair (so that the second half-sample is included twice), repoststratifying the weights, and then reestimating the statistic for the pseudoreplicates based on this altered set of data.

The component of the sampling variability attributable to a jackknife pair is estimated as the squared difference between the value of the statistic for the complete sample and the pseudoreplicate associated with the pair. The estimated sample variance of the statistic t is the sum of M squared differences (where H is the number of jackknife pairs defined):

$$\widehat{Var(t)} = \sum_{i=1}^{M} (t_i - t)^2.$$

The statistic for the pseudoreplicate associated with a given jackknife pair is the original statistic for the pseudoreplicate recomputed using an altered set of weights, referred to as the student replicate weights. The student replicate weight, $SRWT_i$, for the ith pair of first-stage units is computed as follows:

1. Let W_B be the nonresponse adjusted base weight of a student, where WB accounts for the probabilities of selection and nonresponse but does not include poststratification adjustments.

2. Let W_{B_i} be the nonresponse adjusted replicate base weight formed by replacing the second member of the jackknife pair by the first, specifically:
   - $W_{B_i} = 0$ if the student is in the first set of first-stage units in jackknife pair i
   - $W_{B_i} = JF * W_B$ if the student is in the second set of first-stage units in jackknife pair i
   - $W_{B_i} = W_B$ if the student is in neither of the first-stage units in jackknife pair i

   where JF is a constant multiplier (usually equal to 2) designed to maintain certain population totals.
3. Then the student replicate weight for the jackknife pair i is obtained by applying the poststratification adjustments to the weights W_{B_i} in the associated pseudoreplicate.

> The poststratification adjustments are recomputed for each jackknife replicate to reflect more completely the total effect of replacing one member of a jackknife pair with the other. (Nonresponse adjustments are not recomputed since these are generally performed within the PSU level and therefore their effect is appropriately reflected in the variance estimate.) This estimation technique was used by NAEP to estimate all sampling errors [variances] presented in the various reports. (Technical Report 1988, pp. 208–211)

Details of how the PSUs were assigned into pairs are given on p. 208 of Technical Report 1988. We discuss this approach to variance estimation and the cluster sampling design effects in Chapter 3. We note here that it is the *PSUs* for which the design effect is assessed, not the *schools*.

2.5 Analysis

2.5.1 Item models

We need to make a distinction between student *achievement* on the test, which is measured by the student item responses, and student *ability*, an unobservable or *latent* characteristic of the student that underlies achievement; the nature of this underlying relation is expressed through a (statistical) *psychometric model*.

The test items analysed in the 1986 and 2005 surveys were all multiple choice items, scored with a binary response y for incorrect ($y = 0$) or correct ($y = 1$) answers. They are viewed as imperfect *indicators* of the student's true ability, and the object of analysis is to make statements about student ability, including how this varies, using important *reporting group variables*. (For each student, omitted items *beyond* the last item attempted in the booklet are treated as "not reached" and are ignored; they define the effective sample size of items for, and hence information about, the student. Items omitted *within* the range of attempted items are included and can be scored as incorrect, or as "fractionally correct" with y-value $1/R$, where R is the number of response categories for the item. This is equivalent to assuming a *random guess* for omitted items.)

Student ability on the items comprising the test is treated as a *latent variable*, denoted by θ_i for student i. Ability is allowed to depend, in the model, on student, class, teacher, and school variables, which for the moment are denoted by $\mathbf{x}_i$ for

student i, through a multiple regression function $\beta'\mathbf{x}_i$ (this is discussed at length in Chapter 3). A popular model relates the achievement of the student on the test items to the student ability through a *logistic linear regression model*, which we call the *MIMIC model* (for **M**ultiple **I**ndicators, **M**ult**I**ple **C**auses), though it is usually called the *2PL model* in psychometric theory. (In Chapter 3, we give a detailed discussion of this and other models.)

The probability of a correct answer to item j by student i with ability θ_i is written p_{ij} for $i = 1,...,n$, $j = 1,...,J$. The MIMIC model is

$$\begin{aligned} p_{ij} \mid \theta_i &= \exp[a_j(\theta_i - b_j)]/\{1 + \exp[a_j(\theta_i - b_j)]\}, \\ \theta_i &\sim N(\beta'\mathbf{x}_i, 1), \end{aligned}$$

which is equivalent to

$$\begin{aligned} \operatorname{logit} p_{ij} = \log[p_{ij}/(1-p_{ij})] \mid \theta_i &= a_j(\theta_i - b_j), \\ \theta_i &\sim N(\beta'\mathbf{x}_i, 1). \end{aligned}$$

The *item parameters* a_j and b_j are called the *discrimination* and *difficulty* parameters, respectively. The regression model can be reparametrised to the more conventional statistical form $\alpha_j + \beta_j\theta_i$ by setting $a_j = \beta_j$ and $-a_jb_j = \alpha_j$, that is, $b_j = -\alpha_j/\beta_j$.

An essential feature of this model is the *probability distribution* for ability θ across the population of students. If instead the abilities θ_i are regarded as *fixed effects* – fixed unrelated parameters – the estimation of these parameters may be *inconsistent*. If the sample size of students tested increases but the number of items answered remains fixed, the estimates of student ability from the fixed-effect model do not become more precise – they do not converge towards the true ability values. This is because the information about each student's ability remains fixed by the number of items the student attempts – there are more values θ_i, but no more information about any θ_i.

If, on the other hand, the number of students is fixed but the number of items answered increases, then the ability estimates *do* converge towards their true values. Any real test has a finite number of items, which is usually fixed by the testing time available. The NAEP tests use a large number of items over the tested populations, but each student can answer only a small number of items in the test booklet, which are randomly chosen from the large number of test items available. It cannot be assumed therefore that this number is large enough to provide a consistent estimate of the student's ability. Student abilities have to be *linked* through a probability distribution across students to use the information from other students.

A second essential feature of the model is the *conditional independence of the item responses* y_{ij} for each student *given the student's ability*, that is, the correlation between the binary responses is *completely explained* by the ability used to answer the items. This is a standard assumption in many kinds of *multilevel* or *hierarchical* models, where responses at a "lower level" are assumed to be independent given a common *random effect* shared by the responses at a "higher" level.

The assumption of a *normal* distribution for the student abilities θ_i may appear very strong. It is discussed in Chapter 4 and found to be surprisingly *weak*. The setting of the variance to 1 is necessary to identify the item discrimination parameters; this is discussed in Chapter 3.

The MIMIC model is not used in offical NAEP analysis because it does not allow for *guessing* – the possibility of a correct answer by a random process independent of the student's ability level. The model can be extended to the *three-parameter MIMIC model* (usually called the *3PL model*) by incorporating a third *guessing parameter* c_j for each item. The model is

$$\begin{aligned} p_{ij} \mid \theta_i &= c_j + (1-c_j)\exp[a_j(\theta_i - b_j)]/\{1+\exp[a_j(\theta_i - b_j)]\}, \\ \theta_i &\sim N(\beta'\mathbf{x}_i, 1). \end{aligned}$$

This is the model used in the official NAEP analyses of the two surveys we reanalyse in this book.

The 3PL model can be expressed in a logistic form as

$$\begin{aligned} \log\{[p_{ij}-c_j]/[1-c_j]\} \mid \theta_i &= a_j(\theta_i - b_j), \\ \theta_i &\sim N(\beta'\mathbf{x}_i, 1). \end{aligned}$$

A feature of this model is that the probability of a correct response *cannot fall below the guessing parameter* even for those *not* guessing. This is discussed further in Chapter 3.

2.5.2 Multidimensional ability

The models above assume that all items depend on, or reflect, a *single latent dimension* of ability. However, the items in recent math tests cover five scales (four main scales in the 1986 third grade test). These scale dimensions of ability, as determined by the items designed to assess them, are assumed to be correlated within a student, though the item responses are still assumed to be independent, both within and across scales, given the set of abilities on the scales.

The multidimensional item response models are analogous to the single-dimension models above. We denote the multidimensional case by a *vector* ability variable θ_i for student i. The multidimensional MIMIC model is then

$$\begin{aligned} \operatorname{logit} p_{ij} \mid \theta_i &= \mathbf{a}_j'(\theta_i - \mathbf{b}_j), \\ \theta_i &\sim N(\Gamma'\mathbf{x}_i, \Sigma), \end{aligned}$$

where $\mathbf{a}_j$ and $\mathbf{b}_j$ are vectors of discrimination and difficulty parameters for the multiple dimensions, Γ is the matrix made up of sets of regression coefficients of the covariates on each ability dimension, and Σ is the correlation matrix of the ability dimension variables (the variance of each ability variable is 1).

The multidimensional three-parameter guessing generalisation adds the same guessing parameter as for the single-dimensional ability:

$$p_{ij} \mid \theta_i = c_j + (1 - c_j)\exp[\mathbf{a}_j'(\theta_i - \mathbf{b}_j)]/\{1 + \exp[\mathbf{a}_j'(\theta_i - \mathbf{b}_j)]\},$$
$$\theta_i \sim N(\Gamma'\mathbf{x}_i, \Sigma).$$

The current analysis uses this model in a simpler form but does not report (except in technical manuals) the separate dimensions, only a *composite single dimension* in which the separate dimensions are weighted by the number of items assessing that dimension:

> Using a unidimensional IRT model when the true model is multidimensional captures these overall patterns [different ability levels by subgroup] even though it over- or under-estimates the covariances among responses to items in pairs. (Technical Report 1988, p.234)

So if $\mathbf{w}$ is a vector of relative weights attached to each dimension, for the weighted composite $\theta_{ci} = \mathbf{w}'\theta_i$ the mean will be $\mathbf{E}(\theta_{ci}) = \mathbf{w}'\Gamma'\mathbf{x}_i$ and the variance will be $\mathrm{Var}(\theta_{ci}) = \mathbf{w}'\Sigma\mathbf{w}$. Reporting group differences on each dimension will be averaged.

In the 1986 test, there were four main scales,

- Numbers and Operations (56 items),
- Measurement (27),
- Fundamental Methods (17),
- Data Organization and Interpretation (16),

with a total of 116 items. The large Numbers and Operations scale was itself split into two subscales:

- Knowledge and Skills (30 items),
- Higher-Level Applications (26).

These two subscales and the Measurement scale were also reported on, though in less detail than the composite dimension. (There were also smaller scales: Relations, Functions, and Algebraic Expressions (8), Geometry (6), and Discrete Mathematics (3). These, and the Fundamental Methods and Data Organization and Interpretation scales, had so few items for the age 9/grade 3 students that they were not reported on separately.) The composite dimension is determined as a *weighted sum* of the reported scales, weighted by the number of items on each scale.

An important point in the analysis is that there were many more items – a total of 798 – used in the full NAEP survey across the three age groups, shown below from Technical Report 1986, p. 218, Table 10.1.[1]

[1] There is some ambiguity in this number as elsewhere in this report – p. 216 – the total number of items is given as 537.

Area	Total Items	Number of Booklets	Average Number of items per Booklet	Number booklets with number of items 1-2	3-5	>5
Fundamental Methods	102	25	4.1	9	8	8
Discrete Mathematics	18	11	1.6	10	1	0
Data Organization and Interpretation	96	19	5.1	3	10	6
Measurement	162	28	5.8	9	6	13
Geometry	36	11	3.3	5	5	1
Relations, Functions, and Algebraic Expressions	48	25	1.9	20	5	0
Numbers and Operations: Higher-Level Applications	156	28	5.6	9	6	13
Numbers and Operations: Knowledge and Skills	180	25	7.2	9	0	16

The average number of items per booklet is 34.6. Many items were eliminated because of differential item functioning or because there were too many "not reached" responses: the test was too long for students to reach these items when they were placed at the end of the booklet sequence.

In the model-based analysis of the multidimensional ability item response model, a heavy computational load is imposed by the estimation of the *correlations* of the scales. While the item parameters for an individual item are estimated from all those students who actually attempted the item, the correlations between the scales represented by the items are determined by those students attempting *pairs of items*

from *different scales* – pairs of items from the *same* scale do not contribute to the interscale correlations.

Given the balanced incomplete block spiraling of the items into the test booklets and the sparsity of items from *all* scales that are seen by any individual student, the estimation of these correlations has relatively little data to support it. This is accentuated by the reduction in the number of items included in the third grade test and the corresponding reduction in sample size for the estimation of cross-scale item covariances.

This implies that the precision of estimation of the interitem correlations *is much poorer* than that of the item parameters themselves, and that therefore a *wide range of correlation structures* will be consistent with the observed item responses. This issue is somewhat obscured in the official NAEP analysis since the interscale correlations are estimated by direct correlation of the plausible values for each student on the separate scales.

Technical Report 1986 gives (Table 10.9, p. 231) the estimated interscale correlations of the three scales (without standard errors), based on the first plausible value.

Estimated Correlations between Subscales
(Based on the First Plausible Value)
Grade 3/Age 9

	Measurement	N & 0 (H-L)	N & 0 (K-S)
Measurement	1.00	.63	.60
Numbers and Operations: Higher-Level Applications	.63	1.00	.60
Numbers and Operations: Knowledge and Skills	.60	.60	1.00

The correlations are almost uniform, pointing strongly to a single second-level factor that is simply the sum of the three ability dimensions. It explains 75% of the variance of the three dimensions.

The full multidimensional model is not used in our analyses in Chapter 5, which are restricted to a single-dimension ability scale and the items assessing it. We report some limited analyses with the full set of items for the 1986 survey in Chapter 6 and comment in some detail on the complexity of this model.

2.5.3 Inference and the likelihood function

To draw conclusions about group differences, it is necessary to estimate the parameters in both the psychometric model – the item parameters a_j, b_j, and c_j – and the regression model relating ability to achievement, the regression parameters γ. (We will call this regression model the *ability regression model* to distinguish it from a second logistic regression model, to be discussed below.) This can be achieved by maximum likelihood, given originally for the closely related two-parameter *probit* (2PP) model by Bock and Aitkin (1981).

For the n students tested on the J test items, write z_{ij} for the *indicator* variable, taking the value 1 if student i attempts item j and zero if not. (This requires a coding decision on how omitted items within the range of answered items are to be treated, as discussed above.) Then the likelihood can be expressed in terms of all the parameters λ by

$$L(\lambda) = \prod_{i=1}^{n} \int \left[\prod_{j=1}^{J} p_{ij}^{z_{ij} y_{ij}} (1 - p_{ij})^{z_{ij}(1-y_{ij})} \right] f(\theta_i) \mathrm{d}\theta_i.$$

Thus all items answered by each student are treated in the same way and can contribute to estimation of both item parameters and ability regression model parameters.

The integration over the ability distribution is needed, as the probability of the item responses for student i is *conditional* on the value of ability for this student, but this ability is unobserved and so has to be integrated out of the likelihood contibution for this student. The integration has to be done numerically, and this is the major computational load in the analysis: the integral is replaced by a finite sum over a set of $Q = 41$ discrete ability locations ("quadrature points") θ_q^*, from -5 to 5 in steps of 0.25, with probabilities f_q given by the normal density $N(0,1)$ at θ_q^*, scaled to sum to 1. Reparametrising the MIMIC model by $\theta_i^* = \theta_i - \beta' \mathbf{x}_i$, the likelihood can be written as

$$\begin{aligned} L(\lambda) &\doteq \prod_{i=1}^{n} \sum_{q=1}^{Q} \left[\prod_{j=1}^{J} p_{qij}^{z_{ij} y_{ij}} (1 - p_{qij})^{z_{ij}(1-y_{ij})} \right] f_q, \\ p_{qij} &= \exp[a_j(\theta_q^* + \beta' \mathbf{x}_i - b_j)] / \{1 + \exp[a_j(\theta_q^* + \beta' \mathbf{x}_i - b_j)]\}, \\ f_q &= \frac{\exp[-\frac{1}{2}\theta_q^{*2}]}{\sum_{q=1}^{Q} \exp[-\frac{1}{2}\theta_q^{*2}]}. \end{aligned}$$

The MLEs are found by solving the equations given by setting the first derivatives of this log-likelihood function with respect to the parameters to zero, and the standard errors (SEs) of the MLEs are obtained from the inverse of the information matrix – the matrix of negative second derivatives of the log-likelihood function.

2.5.4 The ability regression model

The current NAEP analysis uses a very large "conditioning" regression model $\beta'\mathbf{x}$, in which the vector $\mathbf{x}$ includes all the main effects and two-way interactions of the variables on which the NCES has to report achievement levels, the *reporting group variables*, and a large number of other variables, including dummy variables for each school. This total number of variables is so large, and the correlations between them so high, that the total set of (up to 1200) variables is not used directly but is first reduced to a set of *uncorrelated principal components*, and a subset $\mathbf{z}$ (of around 300) of these principal variables is used instead in a regression $\gamma'\mathbf{z}$, giving an estimated $\hat{\gamma}$ with estimated covariance matrix $\hat{\Lambda}$. In large samples, it may reasonably be assumed that $\hat{\gamma} \sim N(\gamma, \hat{\Lambda})$.

The purpose of fitting such a large model is *not* to interpret or report the parameters $\hat{\gamma}$ of this model; these are uninformative about β. The aim of the conditioning model fitting is to ensure, as far as possible, that *all possible relevant variables* are included in a *predictive model*, which is then used to *multiply impute* ability for each student:

> NAEP conducts a special form of imputation during the third stage of its analysis procedures. The first stage requires estimating item response theory parameters for each cognitive question. The second stage results in MML [marginal maximum likelihood] estimation of a set of regression coefficients that capture the relationship between group score distributions and nearly all the information from the variables in the teacher, school, or SD/LEP questionnaires, as well as geographical, sample frame, and school record information. The third stage involves calculating imputations designed to *reproduce the group-level results that could be obtained during the second stage*. (emphasis added)

(Handbook, p. 199)

For the convenience of the reader, we summarise this concisely before considering each stage in detail:

- Fit a *null* IRT model with items only – no reporting group variables.
- Hold the item parameters fixed at their estimates.
- Fit a large-scale "conditioning" (regression) model with ~ 300 principal components of many (~ 1200) covariates.
- From the posterior distributions of student ability, given the normal ability distribution and the covariates, generate five *multiple imputations* ("plausible values") of ability for each student.
- Tabulate, or regress, the ability imputations by reporting group variables to give estimates and combine them using the Rubin rules.
- Allow for the design effect of PSU sampling in standard error calculations by jackknifing the PSUs – no allowance is made for the school design effect.

2.5.5 Current model parameter estimation

The item parameters and the conditioning regression model parameters are currently estimated in several steps:

- The item parameters are estimated by maximum likelihood first, with a *null* or *empty* regression model $\gamma'\mathbf{z} = 0$.
- The item parameters are then fixed at these estimates, and the conditioning regression model parameters γ are estimated by (constrained) maximum likelihood.

This process does *not* result in the same estimates as obtained by *simultaneously* maximising the likelihood in *all* the parameters. (Results would be identical if this two-step process were *iterated* – repeated in alternate steps until the results stabilised.) There are two reasons for this approach.

First, the development of maximum likelihood analysis for the 2PP model by Bock and Aitkin (1981) (extended by others to the logit and other psychometric models) dealt only with the null regression model – it was purely for item parameter estimation. Extensions of the approach to a multiple group ability regression structure took some time, and the computational power of computers in the 1980s limited the number of test items that could be analysed, let alone an additional regression structure with a large number of covariates at student and school levels.

Second, the conditioning regression model that is fitted in the current approach is *very* large by conventional regression standards, even after the replacement of the 1200 covariates by several hundred principal variables. The size of this model made it impractical to maximise simultaneously over both the item *and* the conditioning model parameters.

2.5.6 Plausible value imputation

The fitted conditioning model $\hat{\gamma}'\mathbf{z}$ is used to impute $M = 5$ "plausible values" of ability for each student from the posterior distribution of ability given the student's 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 estimated parameter vector and covariance matrix the estimated covariance matrix of the estimated parameter vector:

$$\gamma \sim N(\hat{\gamma}, \hat{\Lambda}).$$

This follows from a diffuse prior distribution on γ and large degrees of freedom for $\hat{\Lambda}$ from the large sample size. A random draw $\gamma^{[m]}$ of the conditioning regression vector γ is then made from this normal posterior distribution and combined with the principal variable value $\mathbf{z}_i$ for individual i to give a random draw

$\gamma^{[m]'}\mathbf{z}_i$ from the posterior distribution of the conditioning model predicted value (of mean ability) for this individual.

2. The posterior distribution of ability $\pi(\theta_q \mid \mathbf{y}_{ij})$ for individual i is then constructed on the discrete grid θ_q by evaluating the likelihood for individual i from the item responses y_{ij}, multiplying this by the discrete normal prior distribution ordinates from $N(\gamma^{[m]'}\mathbf{z}_i, 1)$ on these quadrature points, and scaling the sum to 1.0:

$$\pi(\theta_q \mid \mathbf{y}_{ij}) = \frac{\Pr[\{y_{ij}\} \mid \theta_q] \cdot \pi(\theta_q)}{\sum_{\ell=1}^{K} \Pr[\{y_{ij}\} \mid \theta_\ell] \cdot \pi(\theta_\ell)}.$$

3. A random draw of individual i's ability is then made from this posterior distribution by first drawing at random a quadrature point with probability equal to the quadrature mass and then drawing uniformly a plausible (an imputed) ability value between the upper and lower end points of the interval at which the quadrature point was centred.
4. The three steps above are repeated $M = 5$ times to give M plausible values of ability for each individual. These are rescaled to a common NAEP reporting scale that has standard deviation 35 and mean given by a scaling procedure.

The M plausible values are then used in M analyses, which involve (one-way or two-way) *tabulations* of the ability values for 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 estimates and standard errors.

For example, the report for the 1986 survey gave the following Table 1 (from the Data Appendix, p. 138 of Dossey et al. 1988), summarising trends across the last three surveys. The asterisks indicate significantly different means (at the 5% level) in the earlier surveys relative to the 1986 survey. Additional information is provided in the Report Card for the individual scales, though this is in the form of graphs with confidence intervals rather than tables as above.

A curious feature is visible in this table. Nationally, the mean scaled score increased significantly from 1977–78 to 1985–86 by 4.1 NAEP scale points. However, for the subpopulations defined by level of parental education, only the group with less than a high school education improved, and only by 0.3 points! The college graduate group was unchanged, and the two other groups *decreased*, though not significantly. In 1981–82, *all* the group means decreased relative to 1977–78, but the overall population mean *increased*!

These apparent paradoxes are examples of *Simpson's paradox* and are due to changes in the proportions in the parental education categories over the different survey periods. They show the importance of *disaggregated analysis*, in which the differences among ability levels for one reporting group variable may be examined while keeping other reporting group variables constant – for example, in two-way classifications rather than in separate one-way classifications.

It may seem surprising that such a complex analysis is needed for such relatively simple tabulations. The aim of the imputation of ability for individual students was

Table 2.1 Age 9

WEIGHTED MATHEMATICS PROFICIENCY MEANS
AND JACKKNIFED STANDARD ERRORS

	1977–78	1981–82	1985–86
– TOTAL –	218.6(0.8)*	219.0(1.1)	221.7(1.0)
SEX			
MALE	217.4(0.7)*	217.1(1.2)*	221.7(1.1)
FEMALE	219.9(1.0)	220.8(1.2)	221.7(1.2)
ETHNICITY/RACE			
WHITE	224.1(0.9)	224.0(1.1)	226.9(1.1)
BLACK	192.4(1.1)*	194.9(1.6)*	201.6(1.6)
HISPANIC	202.9(2.3)	204.0(1.3)	205.4(2.1)
REGION			
NORTHEAST	226.9(1.9)	225.7(1.7)	226.0(2.7)
SOUTHEAST	208.9(1.2)*	210.4(2.9)	217.8(2.5)
CENTRAL	224.0(1.5)	221.1(2.4)	226.0(2.3)
WEST	213.5(1.4)	219.3(1.7)	217.2(2.4)
PARENTAL EDUCATION			
LESS THAN H.S.	200.3(1.5)	199.0(1.7)	200.6(2.5)
GRADUATED H.S.	219.2(1.1)	218.3(1.1)	218.4(1.6)
SOME EDUCATION AFTER H.S.	230.1(1.7)	225.2(2.1)	228.6(2.1)
GRADUATED COLLEGE	231.3(1.1)	228.8(1.5)	231.3(1.1)

to allow for secondary data analysis with much more complex regression models, without the need for specialised software to perform the numerical integration and likelihood maximisations needed and allowing for the uncertainty in the imputed ability.

By providing the plausible values in the data file, analysts could bypass the test item responses that provided the achievement information, and carry out their own analyses directly on the plausible values (repeating and combining them according to the Rubin rules), with confidence that all relevant relationships of ability to possible covariates had been built into the plausible values and so could be recovered by quite simple analyses.

This concludes the overview of NAEP design and analysis. Chapter 3 examines in more detail the psychometric models used in NAEP, extends them using an alternative treatment of guessing, and discusses the survey design and its representation by a multilevel model.

Chapter 3
Psychometric and Survey Models

Item response models are a special class of generalised linear models. A detailed discussion of these models, including some of those we use in the NAEP analyses, can be found in De Boeck and Wilson (2004) and Skrondal and Rabe-Hesketh (2004). We give details for only those models we used for the NAEP analyses. We extend them to include the survey structure at the end of the chapter.

3.1 The Rasch model

The Rasch model is not used in NAEP, but it is the simplest item response model, of which all other models are extensions.

For student i with latent ability θ_i (assumed one-dimensional) on the items of the test scale, and covariates (reporting group variables) $\mathbf{x}_i$, attempting item j and giving the binary responses y_{ij} with probabilities p_{ij} for a correct answer, the Rasch model (Rasch 1960, Andersen 1972) is

$$\begin{aligned} \operatorname{logit} p_{ij} \mid \theta_i &= \theta_i - b_j, \\ \theta_i &\sim N(\beta' \mathbf{x}_i, \sigma^2). \end{aligned}$$

Here the *b parameter*, b_j, is the item difficulty: as b_j increases, the probability of answering correctly decreases.

In the formulation of the Rasch model above, the latent ability is regressed on the covariates (that is, *ability varies with reporting group*), and given the student's ability θ_i, the item responses y_{ij} are *independent* of the covariates $\mathbf{x}_i$. We call this version of the Rasch model the *ability* regression model.

However, we can transfer the regression model to the logit scale by defining

$$\begin{aligned} \theta_i' &= \theta_i - \beta' \mathbf{x}_i, \\ \theta_i &= \theta_i' + \beta' \mathbf{x}_i, \end{aligned}$$

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 39
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_3,

and then dropping the prime from θ_i' to obtain the Rasch model in the form

$$\begin{aligned} \text{logit}\, p_{ij} \mid \theta_i &= \theta_i - b_j + \beta' \mathbf{x}_i, \\ \theta_i &\sim N(0, \sigma^2). \end{aligned}$$

Now ability does *not* vary with reporting group (it has a *homogeneous* distribution), and the covariates, through the regression model, *directly affect the probability of answering the items correctly*; that is, the covariates *directly affect achievement*. In this formulation, we call the regression model the *achievement* regression model. For the Rasch model, this is a distinction without a difference, since these two formulations are equivalent and indistinguishable. We will see, however, that for the two-parameter models in more complex multilevel structures, there is a much greater distinction between the two formulations.

It is also clear from this formulation that the intercept term in the reporting group regression is not identifiable: it is aliased (confounded) with one of the item difficulty terms b_j. So the regression model is usually fitted without an intercept.

3.2 The 2PL and MIMIC models

The literature on the two-parameter logistic latent class models comes from two traditions: psychometric and econometric. The model called the "2PL" by psychometricians is called the "MIMIC" model by econometricians and structural equation modelers.

We adopt in this book the econometric "MIMIC" notation and use the term "2PL" to refer to a different model, which we describe below. The distinction between the models, as described in the previous section, comes from the placement of the regression function $\beta' \mathbf{x}$ in the model. For a null regression function, the models are the same.

The null 2PL model was developed by Lord (1952) and Birnbaum (1968), but the same model with covariates in a regression function was given the name MIMIC (Multiple Indicators, MultIple Causes), in its factor analysis form, by Jöreskog and Goldberger (1975). The 2PL model has the form

$$\begin{aligned} \text{logit}\, p_{ij} \mid \theta_i &= a_j(\theta_i - b_j) + \beta' \mathbf{x}_i, \\ \theta_i &\sim N(0, 1), \end{aligned}$$

and the MIMIC model has the form

$$\begin{aligned} \text{logit}\, p_{ij} \mid \theta_i &= a_j(\theta_i - b_j,) \\ \theta_i &\sim N(\beta' \mathbf{x}_i, 1). \end{aligned}$$

Unlike the Rasch model, these models cannot identify the form with a variance parameter σ^2 in the ability distribution. In the apparently more general MIMIC

model, we have

$$\begin{aligned}\theta_i &\sim N(\beta'\mathbf{x}_i, \sigma^2),\\ \theta_i^* &= \theta_i/\sigma\\ &\sim N(\beta'\mathbf{x}_i/\sigma, 1),\\ \text{logit}\, p_{ij} \mid \theta_i^* &= a_j(\sigma\theta_i^* - b_j)\\ &= a_j^*(\theta_i^* - b_j^*),\end{aligned}$$

where $a_j^* = \sigma a_j$, $b_j^* = b_j/\sigma$. So, for the model with scaled abilities θ_i^* with variance 1, the ability regression coefficients and the item difficulties for the "general" model are divided by σ, while the item discrimination parameters are multiplied by σ. The maximised likelihoods are identical for the "general" model and the model with variance 1; the "general" model is strictly *unidentifiable* without a constraint on one of the discrimination parameters or setting the variance to a known constant. In the 2PL model, only the item discrimination parameters are affected, in the same way. In the analyses in this book, we follow the convention of setting the ability variance to 1 rather then setting one discrimination parameter to 1 to "identify" the variance.

The MIMIC model has the *ability regression model* with the reporting group variables "inside" the student ability distribution; the 2PL model has the *achievement regression model* "outside", in the logit model for the item responses. The MIMIC model has the strong property that, given ability, the item response probabilities *do not depend on the covariates* – only the student ability matters (though this may vary by the covariates). This has been an important issue in the design and checking of items: if an item has different success probabilities for students with the same ability but from different ethnic or gender groups, this is regarded as an *item bias* with respect to the ethnic or gender grouping – *differential item functioning*, or DIFF.

However, by transferring the MIMIC reporting group regression model to the logit scale, as for the Rasch model, we have the *equivalent 2PL model*:

$$\begin{aligned}\text{logit}\, p_{ij} \mid \theta_i &= a_j(\theta_i - b_j) + a_j \cdot \beta'\mathbf{x}_i,\\ \theta_i &\sim N(0,1).\end{aligned}$$

Now we have a *homogeneous ability distribution* with *an interaction achievement model on the logit scale*; the success probabilities on an item for students with the same ability now depend on the reporting group variables *as well as* the student's ability.

So the issue of apparent bias in an item is affected by the way in which we express the MIMIC or the Rasch model: if the ability distribution contains the reporting group regression variables, then items should *not* show group differences for students of the same ability. If the reporting group regression is transferred to the 2PL logit scale in these models, items *will* show group differences for students of the same ability.

Since the two forms of the reporting group regression in these models are interchangeable, this makes clear that the usual definition of item bias – that, given

the same ability, the item response probabilities are different for different reporting groups – is not a satisfactory definition.

In the 2PL model, the effects of the reporting group variables and individual items are *additive* on the logit scale, whereas in the MIMIC model item slopes *interact* with the reporting group variables on the logit scale, so that the effects of the reporting group variables on the item response probabilities are scaled by the item discriminations and so are *different* for each item. Thus, in the MIMIC model, for highly discriminating items, the differences on the logit scale for students of the same ability but from different reporting groups will be larger than the differences for poorly discriminating items. It would be unfortunate if highly discriminating items were dropped from a test because of observed reporting group differences for students of the same ability, when this is to be expected under any of these models.

The two models have the same number of parameters but can be discriminated with sufficient data through their maximised likelihoods.

3.3 Three-parameter models

The Rasch, 2PL, and MIMIC models do not allow for the possibility that students may *guess* at an item. The models above can all be extended by an additional parameter, giving a form of *mixture* model to allow for this. The extended models all have the form, for the probability p_{ij} of a correct response by student i on item j,

$$p_{ij} = c_j + (1 - c_j)q_{ij},$$

where the model for q_{ij} is a Rasch, 2PL, or MIMIC model, with the appropriate marginal distribution for ability as given in the previous section. These models have the property that, for students of very low ability, the probability of answering item j correctly is c_j, the *guessing* or *c parameter* of the item. This value is the *minimum* probability of a correct response under the model.

So, for example, the three-parameter version of the 2PL model for the probability p_{ij} of examinee i with ability θ_i correctly answering item j is given by

$$\begin{aligned} p_{ij} &= c_j + (1 - c_j)q_{ij}, \\ \operatorname{logit} q_{ij} \mid \theta_i &= a_j(\theta_i - b_j) + \beta' \mathbf{x}_i, \\ \theta_i &\sim N(0, 1), \end{aligned}$$

while the three-parameter version of the MIMIC model is

$$\begin{aligned} p_{ij} &= c_j + (1 - c_j)q_{ij}, \\ \operatorname{logit} q_{ij} \mid \theta_i &= a_j(\theta_i - b_j), \\ \theta_i &\sim N(\beta' \mathbf{x}_i, 1). \end{aligned}$$

In our experience with the two NAEP surveys reported in the following chapters, we were unable to fit the MIMIC version of the model: despite starting from many different sets of starting values, the successive parameter estimates did not converge to a maximum of the likelihood. However, the 2PL version of the three-parameter model converged without difficulty in all our analyses.

This corresponded with the ETS experience with the three-parameter MIMIC model, that very tight priors on the guessing parameters are needed to identify them and it. This means in effect that the guessing parameters have to be *known* with this model – the full three-parameter MIMIC model itself is *unidentifiable*, an obvious disadvantage. The results we present in later chapters are all based on the well-identified 2PL version of the three-parameter model.

There are several descriptions of the basis for the three-parameter model for a multiple-choice item (Birnbaum 1968; Hutchinson 1991; San Martin, del Pino, and De Boeck 2006). Birnbaum described this (pp. 404–5 in Lord and Novick 1968; we replace Birnbaum's probit by the logit) as

> ... a highly schematized psychological hypothesis. ... [I]f an examinee has [ability] θ, then the probability that he will *know* the correct answer is given by [the logit function]
>
> $$G[a_j(\theta - b_j)] = e^{a_j(\theta - b_j)}/[1 + e^{a_j(\theta - b_j)}]$$
>
> ... [I]f 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, or others,[1] the introduction of the guessing 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. The *two-component mixture* form of these models is more difficult to estimate by maximum likelihood as it has *two* kinds of latent structure. An aspect of this model that is not immediately obvious is that, if many of the test items have nonzero guessing parameters, the estimates of achievement or ability parameters $\widehat{\beta}$ will tend to increase relative to the corresponding estimates for the two-parameter model. This is because the restriction of range of the probability scale will require larger effects to maintain the group differences in success probabilities found on the full [0,1] scale for the two-parameter models.

For the Rasch model, the guessing extension gives a two-parameter model. This has been discussed at length in von Davier and Carstensen (2007). We investigated

[1] We can rewrite the model in the alternative form

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

which has a different interpretation.

this model but found that it gave fits similar to the MIMIC and 2PL models – the guessing parameter compensated for the discrimination parameters in the other models.

The Rasch guessing model is rarely used in practice because it violates a strong measurement assumption of the Rasch model, that the items are equivalent except in difficulty, which leads to the total score on the items as a sufficient statistic for a person's ability. Guessing on some items would imply a failure in the writing of these items (too difficult for the tested population), which would lead, when identified, to the omission of these items from the test. The remaining items would then follow the Rasch model, so no guessing extension would be needed.

Second, the *general mixture* of Rasch models, where the test-taking population is heterogeneous in their responses to the items, is quite widely used with different sets of Rasch item difficulty parameters in different subpopulations taking the test. This family of mixture models is discussed below.

3.4 Partial credit model

NAEP analyses may also use *partial credit* models (Masters 1982), in which the choice of an incorrect response may be given partial credit or the incomplete task requires a sequence of steps for some of which partial credit may be given. These models were not used in our NAEP 1986 or 2005 data analyses, for which only the correct or incorrect response was used.

3.5 The HYBRID model

This model, proposed by Yamamoto (1987) and used in several research applications (Yamamoto 1989, 1995, Yamamoto and Everson 1997, Boughton and Yamamoto 2007), is described in detail in von Davier and Carstensen (2007, §6.3 and Chapter 9). The nature of guessing is different from that in the 3PL model: guessing is assumed to occur only towards the end of the test, when the student realises that he or she cannot finish the test using cognitive ability in the remaining time available. Up to this point, the student answers according to a Rasch model, but beyond this point the student switches to random guessing, independent of ability. The correct response probabilities by guessing are assumed to vary by item but not by person, as in the 3PL model.

An additional complication in this model is the need to identify the (different) item at which each student changes strategy; these points are an additional set of nuisance parameters in the model. The likelihood for this model also requires the *sequence* of items attempted for each student, which may not be clear from the paper responses.

The model has not been extended beyond the Rasch model to the two-parameter IRT models. We have not used it in this work.

3.6 Extensions of the guessing model

3.6.1 A four-parameter guessing model

The 3PL model treats the guessing and nonguessing asymmetrically. A student might follow a different strategy, deciding first whether he or she has enough confidence to try to solve the problem. If so, the probability of a correct answer is q_{ij}. If not, the student guesses and answers correctly with probability d_j. The decision varies with the item: with probability c_j the student decides to guess, and with probability $1 - c_j$ decides not to guess but to try to solve the problem. This formulation of a strategy leads to an explicit two-component mixture model that has two forms, like the three-parameter model. The four-parameter version of the 2PL model is

$$\begin{aligned} p_{ij} &= c_j \cdot d_j + (1 - c_j) \cdot q_{ij}, \\ \operatorname{logit} q_{ij} \mid \theta_i &= a_j(\theta_i - b_j) + \beta' \mathbf{x}_i, \\ \theta_i &\sim N(0, 1), \end{aligned}$$

while the four-parameter version of the MIMIC model is

$$\begin{aligned} p_{ij} &= c_j \cdot d_j + (1 - c_j) \cdot q_{ij}, \\ \operatorname{logit} q_{ij} \mid \theta_i &= a_j(\theta_i - b_j), \\ \theta_i &\sim N(\beta' \mathbf{x}_i, 1). \end{aligned}$$

Interpretation of the 3PL as a formal mixture model and comparison with the four-parameter model shows that the 3PL's single c_j parameter is split into two: the new c_j is now *the proportion of guessers on item* j, and d_j is *the probability of a correct guess on item* j. It follows immediately that the 3PL model is the special case of the four-parameter model with $d_j = 1$ for all j, that is, *all those guessing on item* j *guess correctly* ! This is an unexpected and unreasonable feature of the Birnbaum interpretation of the 3PL model.

The model above has a second, slightly different interpretation. The mixture form of the model is consistent with a *heterogeneous population* in which there are two types of students: one type answers the items according to ability, the other answers items randomly and is not *engaged* in the task. This model is plausible because the NAEP is not a test within the school curriculum and does not count towards the students' grades, though it may well count for state assessment of the school.

However, since the subpopulation structure is unobserved, it has to be inferred from the item responses. It does not follow from the model that *any* student actually

guesses on *all* the items or that any student *does not guess on any item*. The model is not a *classification* of the students into types. What it does is to ensure that items on which a student guesses do not contribute to the estimation of his or her ability, and hence of reporting group differences, since by definition ability is not engaged on items that are guessed. This is ensured by the appearance of the reporting group regression in only the nonguessing component.

In initial analyses with this model, we found it very difficult to identify the d_j parameters separately from the c_j parameters; the d_j parameters had to be *fixed* to identify the model. This consideration led us to different extended models, in which we generalise the form of heterogeneity in the ability population.

3.6.2 The "2-guess" model

We specialise the four-parameter model above by setting the c_j parameters to a *constant* c: the probability of guessing *does not vary by item*. The 2-guess version of the 2PL model is

$$\begin{aligned} p_{ij} &= c \cdot d_j + (1-c) \cdot q_{ij}, \\ \operatorname{logit} q_{ij} \mid \theta_i &= a_j(\theta_i - b_j) + \beta' \mathbf{x}_i, \\ \theta_i &\sim N(0,1), \end{aligned}$$

while the 2-guess version of the MIMIC model is

$$\begin{aligned} p_{ij} &= c \cdot d_j + (1-c) \cdot q_{ij}, \\ \operatorname{logit} q_{ij} \mid \theta_i &= a_j(\theta_i - b_j), \\ \theta_i &\sim N(\beta' \mathbf{x}_i, 1). \end{aligned}$$

The probability of a correct response on item j, for those who are guessing, is d_j. This model we found to be readily identifiable in all the data sets we examined. It appears somewhat unreasonable, however, that the probability of guessing does not vary by item, and we generalise this model further below.

3.6.3 The "2-mix" model – a five-parameter general mixture of logits model

In this two-component mixture model, which we call the "2-mix" model, there is no systematic guessing, but different *strategies* may be followed in the two components, reflecting ability, demographic, or educational level differences in the two subpopulations. (There could be more than two component subpopulations, but

identifiability difficulties occur with large numbers of parameters in multiple components and the sparse item data resulting from the design.)

In the first component that contains a proportion c of the population, the probability of a correct answer on item j is given by the two-parameter model 1 with parameters a_{1j} and b_{1j}, while in the second component, containing the proportion $(1-c)$ of the population, the probability of a correct answer on item j is given by the two-parameter model 2 with parameters a_{2j} and b_{2j}. The probability of a correct answer on item j is then

$$p_{ij} = c \cdot p_{1ij} + (1-c) \cdot p_{2ij},$$

where

$$\begin{aligned} \operatorname{logit} p_{1ij} &= a_{1j}(\theta_i - b_{1j}), \\ \operatorname{logit} p_{2ij} &= a_{2j}(\theta_i - b_{2j}), \end{aligned}$$

and the parameters in the two components are in general unrelated.

Here the reporting group regression and the parameters of the ability distributions in the components are not yet specified. Since in each component students are not systematically guessing but are answering according to their abilities, both components contribute to the estimation of reporting group differences.

We assume that the reporting group regressions are *the same* in the two components. This obviously aids in interpreting the common regression. The most general possibility is that the reporting group regression coefficients are *different* in the two components; we call this model the "2-mix-regressions" model. We were able to identify this model in the 1986 survey but not in the smaller 2005 surveys.

Since the component ability variance and discrimination parameters are confounded (not separately identifiable), the variances in both components need to be specified as 1. It may seem that they should be set to different values, but as we showed in §2 this only changes the scaling of the discrimination parameters and does not give greater generality. For the same reason, the regression model intercept terms are set to zero in both components. It may be possible (we give several examples in later chapters) to identify and estimate a *location shift* in the ability between the components. This requires the "anchoring" of one item's discrimination parameters by equating them in the two components. If this is not done, ability differences in the two components are reflected in differences in the difficulty and discrimination parameters in the two components.

It remains to specify the placement of the reporting group regression relative to the ability distribution. In the "2-mix 2PL" model, the common regression appears on the two logit scales,

$$\begin{aligned} \operatorname{logit} p_{1ij} &= a_{1j}(\theta_i - b_{1j}) + \beta' \mathbf{x}_i, \\ \operatorname{logit} p_{2ij} &= a_{2j}(\theta_i - b_{2j}) + \beta' \mathbf{x}_i, \\ \theta_i &\sim N(0,1), \end{aligned}$$

while in the "2-mix MIMIC" model, the regression appears in the ability distribution:

$$\begin{aligned} \text{logit}\, p_{1ij} &= a_{1j}(\theta_i - b_{1j}), \\ \text{logit}\, p_{2ij} &= a_{2j}(\theta_i - b_{2j}), \\ \theta_i &\sim N(\beta' \mathbf{x}_i, 1). \end{aligned}$$

The *mixture of Rasch models* is a special case of this model with $a_{1j} = a_{2j} = 1$. Multivariate and mixture Rasch models have been discusssed in detail by von Davier and Carstensen (2007) and their contributors. We investigated these models for the 2005 data but found them substantially inferior in fit to the mixture of 2PL and MIMIC models for the NAEP test items, so we do not report on them here.

3.7 Modeling the component membership probability

A final extension of the two-component models above incorporates explanatory variables at the student level *affecting the probability of membership in the components* as well as the probability of correct responses on the items. We model the membership probability by a logistic regression on the explanatory variables, and so the c parameter is replaced by

$$c_i = \exp(\gamma' \mathbf{x}_i)/[1 + \exp(\gamma' \mathbf{x}_i)].$$

To distinguish the regressions, we call the regression $\gamma' \mathbf{x}$ the *membership regression*. The guessing models with this extension we call the "2-guess-prob 2PL" model and the "2-guess-prob MIMIC" model. The general two-component mixture models with this extension we call the "2-mix-prob 2PL" model and the "2-mix-prob MIMIC" model. We report the results from some of these models (where they are identified) in Chapters 5, 6, 7, and 8.

3.8 Multidimensional ability

In Chapter 2, §5.2, we described the current analysis of the multiple scales in the 1986 test. For the three scales for which analyses were reported, we noted that the correlations among the scales were high – around 0.6 – and almost uniform, pointing strongly to a single second-level factor that is simply the sum of the three ability dimensions. It explained 75% of the variance of the three dimensions.

This raises the question of *why* it is necessary to estimate these correlations at all. We first note that there are three possible simple alternatives:

- assume the dimensions (scales) are independent;
- assume the entire test is unidimensional;

- ignore the scale assignments of items, and fit a multidimensional model with independent dimensions.

If the scales are of substantive interest and are to be reported separately as in the 1986 report, then nothing is lost by the first approach, that of estimating the item parameters for each scale separately, since the *marginal* distribution of each scale is correctly specified. (This approach using separate analyses for each scale was followed in the official analysis – see Technical Report 1986, p. 221.)

For the second approach, an assumed unidimensional ability will be automatically weighted across the scales by the number of items for each scale, so the correlations among the scales are irrelevant.

By *combining* these two approaches, both the composite scale and the individual scales can be analysed in the same model analysis, with no need to estimate inter-scale correlations.

The third approach appears to lose precision – we have lost the scale membership of each item. However, the *rotational invariance* of the multidimensional model helps us.

For simplicity, consider the case of two scales. Write the 2PL two-dimensional model in the form

$$\begin{aligned} \operatorname{logit} p_{ij} &= \beta' \mathbf{x}_i + \lambda_{0j} + \lambda_{1j}\theta_{1i} + \lambda_{2j}\theta_{2i}, \\ \theta_i &\sim N(0, D), \end{aligned}$$

where θ_1 and θ_2 are the two correlated abilities with correlation ρ, D is the matrix with diagonal elements 1 and off-diagonals ρ, λ_{0j} are the item intercepts, and λ_{1j} and λ_{2j} are the loadings of the items on the two abilities. The orthogonal rotation

$$\phi_1 = (\theta_1 + \theta_2)/\sqrt{2},\ \phi_2 = (\theta_1 - \theta_2)/\sqrt{2},$$

gives *independent* rotated abilities ϕ_1 and ϕ_2, with

$$\theta_1 = (\phi_1 + \phi_2)/\sqrt{2},\ \theta_2 = (\phi_1 - \phi_2)/\sqrt{2},$$

and the two-dimensional model becomes

$$\begin{aligned} \operatorname{logit} p_{ij} &= \beta' \mathbf{x}_i + \lambda_{0j} + \lambda_{1j}(\phi_{1i} + \phi_{2i})/\sqrt{2} + \lambda_2(\phi_1 - \phi_2)/\sqrt{2} \\ &= \beta' \mathbf{x}_i + \lambda_{0j} + (\lambda_{1j} + \lambda_{2j})\phi_{1i}/\sqrt{2} + (\lambda_{1j} - \lambda_{2j})\phi_{2i}/\sqrt{2} \\ &= \beta' \mathbf{x}_i + \lambda_{0j} + \eta_{1j}\phi_{1i} + \eta_{2j}\phi_{2i}, \\ \phi_{1i} &\sim N(0, 2(1+\rho)), \\ \phi_{2i} &\sim N(0, 2(1-\rho)). \end{aligned}$$

If λ_1 and λ_2 have complementary blocks of zeros (items load on only one ability dimension), η_1 and η_2 will have no zeros, so this transformation will completely lose the "scale purity" of the items on the new independent ability dimensions – all items load on both dimensions. The reporting group parameters can be estimated

unbiasedly, though not fully efficiently, using a general two-dimensional analysis of all the items. This analysis has the additional benefit of guarding against failure of the assumed scale purity of the items and the assumption that they load on only one ability dimension. Thus items can be included in the test even if they load on more than one ability dimension or if their loading pattern is uncertain.

For the two-dimensional MIMIC model, this approach fails because the rotation to uncorrelated abilities also transforms the ability regression model:

$$\begin{aligned}\theta_i &\sim N(\Gamma'\mathbf{x}_i, D),\\ \phi_{1i} &\sim N((\gamma_1+\gamma_2)\mathbf{x}_i/\sqrt{2}, 2(1+\rho)),\\ \phi_{2i} &\sim N((\gamma_1-\gamma_2)\mathbf{x}_i/\sqrt{2}, 2(1-\rho)),\end{aligned}$$

although the original Γ is recoverable from the separate ability dimension analyses.

We illustrate the 2PL approach with the analysis in Chapter 5 of the subset of students who answered any of the 30 items from the Knowledge and Skills scale and a comparison with the NAEP report on this scale, followed by the analysis in Chapter 6 of the full set of 79 items from the three reported scales and a comparison with the corresponding NAEP report.

3.9 Clustering and variance component models

It is well established in educational surveys that the clustering of students in the same classes or same schools leads to greater homogeneity of their test results and other standard measures than for students in different classes or different schools. This homogeneity is represented by *multilevel (variance component) models* in which every student response contains an unobserved *shared random effect* for the class or school in which the student is located. This approach was developed in the school context by Aitkin, Anderson, and Hinde (1981), and the principles of educational survey design and analysis were set out in detail in Aitkin and Longford (1986). Book-length treatments of this approach can be found in Longford (1993), Goldstein (2003), and Raudenbush and Bryk (2002).

All the psychometric models discussed in this chapter are special cases of this approach in which the item responses depend on a student ability random effect *shared by the items* and are *conditionally independent* given this random effect, on which the item responses have a logistic regression. This connection between variance component models and psychometric models was not formally recognised until Adams, Wilson, and Wu (1997).

The variance component model formulation allows the two stages of the 1986 NAEP survey design to be built into a *single model* with binary item responses. No external design effect correction using jackknifing PSUs (or schools) is necessary. We demonstrate this in two stages.

3.9.1 Three-level models

The three-level models represent the clustering of students in schools; these are needed for the 2005 surveys. They include an additional random effect η_k, normally distributed with mean 0 and variance σ^2_{sch} (which is identifiable), for the effect of school k from which student i was sampled; we extend the notation so that the student index i runs over the range $1,...,n_k$ for the n_k students sampled from school k.

We also need to change the notation for ability in the MIMIC model since mean ability varies by school in this model; we write θ_{ik} for the ability of student i in school k.

A further notational change is made for the covariates. As we noted in Chapter 2, these include variables that are recorded on the school and so should be denoted by a k subscript. We adopt the notation $\mathbf{x}_{ik}$ for the covariates in the three-level models whether the variables are in fact at the student or school level.

We give the MIMIC and 2PL model formulations: the Rasch version follows from the MIMIC by setting $a_j = 1$, and the 3PL and mixture versions use the same formulation. For the 2PL model, the full model and likelihood become

$$\begin{aligned} \operatorname{logit} p_{ijk} \mid \theta_i, \eta_k &= a_j(\theta_i - b_j) + \beta' \mathbf{x}_{ik} + \eta_k, \\ \theta_i &\sim N(0,1), \\ \eta_k &\sim N(0, \sigma^2_{sch}), \end{aligned}$$

$$L(\lambda) = \prod_{k=1}^{K} \left\{ \int \prod_{i=1}^{n_k} \int \left[\prod_{j=1}^{k} p_{ijk}^{z_{ij}y_{ij}} (1 - p_{ijk})^{z_{ij}(1-y_{ij})} \right] f(\theta_i) \mathrm{d}\theta_i \right\} f(\eta_k) \mathrm{d}\eta_k,$$

while for the MIMIC model

$$\begin{aligned} \operatorname{logit} p_{ijk} \mid \theta_{ik} &= a_j(\theta_{ik} - b_j), \\ \theta_{ik} \mid \eta_k &\sim N(\beta' \mathbf{x}_{ik} + \eta_k, 1), \\ \eta_k &\sim N(0, \sigma^2_{sch}), \end{aligned}$$

$$L(\lambda) = \prod_{k=1}^{K} \left\{ \int \prod_{i=1}^{n_k} \int \left[\prod_{j=1}^{k} p_{ijk}^{z_{ij}y_{ij}} (1 - p_{ijk})^{z_{ij}(1-y_{ij})} \right] f(\theta_{ik}) \mathrm{d}\theta_{ik} \right\} f(\eta_k) \mathrm{d}\eta_k.$$

The school random effect has different roles in the 2PL and MIMIC models. In the MIMIC model, *the latent abilities of students in the same school are correlated*. In the 2PL model, the latent abilities of students in the same school are *independent*: it is the *propensity to answer the items correctly* that is correlated for students in the same school. A simple interpretation of the MIMIC model is that it represents *school selection* of children with similar abilities, as would be natural in a fully *streamed* (primary) school selection system. The 2PL interpretation is more complicated: the school random effect represents the extent of school *training* of students to answer

the test items. The two models give different maximised likelihoods and so can be discriminated.

However, the MIMIC model can be re-expressed as an interaction 2PL model, as we showed above. Write $\theta^*_{ik} = \theta_{ik} - \beta' \mathbf{x}_{ik} - \eta_k$; then

$$\begin{aligned} \operatorname{logit} p_{ijk} \mid \theta^*_{ik} &= a_j(\theta^*_{ik} + \beta' \mathbf{x}_{ik} + \eta_k - b_j), \\ \theta^*_{ik} \mid \eta_k &\sim N(0,1), \\ \eta_k &\sim N(0, \sigma^2_{sch}), \end{aligned}$$

where now a_j scales the school random effect standard deviation as well as the effects of the covariates. This version of the MIMIC model has a different and more complicated interpretation: that latent abilities are independent and propensities to answer correctly are correlated, as in the 2PL model, but further, the effects of the covariates on outcomes and the among schools variance component *vary by item*.

The three-level model *automatically adjusts the standard errors of all parameter estimates for the clustered school design*. No separate estimation of design effects – allowing for the intraclass correlations of students' ability in the same schools – is necessary. (The inclusion of the school identifiers as fixed effects in the large conditioning model does *not* correct for the school design effect – the schools need to be represented by *random effects* rather than *fixed effects*.)

3.9.2 Four-level models

The four-level model needed for the 1986 survey adds a PSU-level random effect ε_ℓ, normally distributed with mean 0 and variance σ^2_{PSU} (which is identifiable), for the effect of PSU ℓ; the notation for the school effect changes to η_{kl} for the effect of school k in PSU ℓ, and the "among schools" variance component for schools in the three-level model is decomposed into the "among PSUs" component σ^2_{PSU} and the "among schools within PSUs" component σ^2_{sch}. We extend the explanatory variable notation further, to $\mathbf{x}_{ik\ell}$, as there is at least one variable recorded at the PSU level (geographical region). It would be possible also to have covariates at the *item* level, requiring an additional j subscript on $\mathbf{x}$ (for example, *scale membership*), but we do not use any such variables in the analysis. The ability latent variable θ is also indexed for ℓ in the MIMIC model but not in the 2PL model, which has the same homogeneous distribution across the clustered population.

For the 2PL model, the four-level model and likelihood are

$$\begin{aligned} \operatorname{logit} p_{ijk\ell} \mid \theta_i, \eta_{k\ell} &= a_j(\theta_i - b_j) + \beta' \mathbf{x}_{ik\ell} + \eta_{k\ell}, \\ \theta_i &\sim N(0,1), \\ \eta_{k\ell} \mid \varepsilon_\ell &\sim N(\varepsilon_\ell, \sigma^2_{sch}), \\ \varepsilon_\ell &\sim N(0, \sigma^2_{PSU}), \end{aligned}$$

$$L(\lambda) = \prod_{\ell=1}^{L} \left\{ \int \prod_{k=1}^{K} \left\{ \int \prod_{i=1}^{n_k} \int \left[\prod_{j=1}^{k} p_{ijk}^{z_{ij}y_{ij}} (1 - p_{ijk})^{z_{ij}(1-y_{ij})} \right] f(\theta_i) \mathrm{d}\theta_i \right\} f(\eta_k) \mathrm{d}\eta_k \right\} f(\varepsilon_\ell) \mathrm{d}\varepsilon_\ell,$$

while, for the MIMIC model,

$$\begin{aligned} \operatorname{logit} p_{ijk\ell} \mid \theta_{ik\ell} &= a_j(\theta_{ik\ell} - b_j), \\ \theta_{ik\ell} \mid \eta_{k\ell} &\sim N(\beta' \mathbf{x}_{ik\ell} + \eta_{k\ell}, 1), \\ \eta_{k\ell} \mid \varepsilon_\ell &\sim N(\varepsilon_\ell, \sigma^2_{sch}), \\ \varepsilon_\ell &\sim N(0, \sigma^2_{PSU}), \end{aligned}$$

with a similar likelihood expression.

It is of particular importance that the four-level model *automatically adjusts the standard errors of all parameter estimates for the two-stage cluster design with both PSU and school sampling*. As with the three-level model for the 2005 surveys, no separate estimation of design effects – allowing for the intraclass correlations of students' ability in the same schools or PSUs – is necessary. This is important because the NAEP design effect correction for clustering by jackknifing PSUs *only adjusts for the PSU clustering* – it does *not* adjust for the clustering of students in schools. As we will see in Chapter 5, the school clustering is *the major design effect in the 1986 NAEP survey*, and the failure to allow for it *biases downward the standard errors of some of the reporting group differences*.

3.10 Summary of the full model for NAEP analysis

We summarise the fully model-based approach to our NAEP analyses below. We give a *single analysis* for each of the psychometric models discussed; the model incorporates

- provision for guessing, either by item or by personal characteristics, or a more general heterogeneity in the item parameters;
- an achievement (2PL) or ability (MIMIC) regression on the chosen covariates at student, family, and school levels;
- random effects for school and (in 1986) PSU to account for the cluster survey design;
- stratification and oversampling of minority strata accounted for by fitting ethnicity as a model factor;
- a membership regression for the guessing or general heterogeneity model;
- maximum likelihood estimates and correct standard errors from the standard package output.

Before our reports of the analyses in Chapter 5 and later chapters, we give a summary in Chapter 4 of the simulation studies that support our proposal for the general use of this model approach.

Chapter 4
Technical Reports – Data Analyses and Simulation Studies

The arguments that support the model-based analysis we describe in Chapter 3, and carry out in the following chapters on the 1986 and 2005 NAEP surveys, are based partly on the statistical theory set out in the previous chapters but also, and importantly, on small- and large-scale simulation studies we carried out over the period 2003–2009. In this chapter, we summarise the conclusions from these studies; the full details are in the NCES Technical Reports. These can be downloaded from the Web; the file name is given for each report. The following chapters incorporate the relevant details from the reports; Chapters 7 and 8 reproduce almost completely the content of the technical reports on the Texas and California 2005 surveys, so these reports are not provided.

4.1 Research reports

The 12 reports below cover the work done for NCES and IES over this period. Since many of the methods were new to NAEP, the value of these methods has been investigated by simulation studies with smaller versions of NAEP-type structures as well as by application to real NAEP data sets.

Comparison of direct estimation with the conditioning model and plausible value imputation

```
http://www.ms.unimelb.edu.au/~maitkin/TechReports
/NAEP_condition.pdf
```

This project began the NAEP series. It established that in simulations direct maximum likelihood estimation of both item parameters and reporting group parameters was uniformly superior to the current method of estimating item parameters first, generating five plausible values of ability from the item model, fitting a regression

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 55
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_4, © Murray Aitkin and Irit Aitkin 2011

of each plausible value on the reporting group variables, and finally combining the five sets of reporting group parameter estimates.

The importance of this project was that it established the efficiency of direct maximum likelihood estimation of regression model parameters in IRT models relative to the current indirect methods of analysis.

Percentile estimation for the ability distribution in item response models

```
http://www.ms.unimelb.edu.au/~maitkin/TechReports
/NAEP_percentiles.pdf
```

This project examined a range of probability models for the ability distribution. It established that, for reliable inference about percentiles of this distribution, explicit and detailed parametric modeling of it was essential: reliance on the normal distribution was unsound, and nonparametric estimation of the ability distribution was ineffective.

The importance of this project was that NAEP reports percentiles of the ability distribution by major reporting group variables assuming a normal ability distribution. The reported percentiles depend strongly on the distributional assumption for ability, so this needs to be checked.

Multilevel model analysis of the Knowledge and Skills scale of the NAEP 1986 math data

```
http://www.ms.unimelb.edu.au/~maitkin/TechReports
/NAEP_4level.pdf
```

This project set out the generalised linear model framework for IRT models and its extension to multilevel models for clustered survey designs. It established that the four-level maximum likelihood analysis of a 30-item scale of the 1986 NAEP math data using the 2PL model for all items was computationally feasible, and that this analysis properly allowed for the survey design (requiring two extra model levels), giving correct and efficient standard errors. The Gllamm program in Stata was used; it was effective but very slow.

The importance of this project was that it established, for the first time, that full multilevel model-based maximum likelihood analysis was possible for NAEP-scale data.

Comparison of joint and separate estimation analyses of the Knowledge and Skills scale of the NAEP 1986 math data

```
http://www.ms.unimelb.edu.au/~maitkin/TechReports
/NAEP_joint_separate.pdf
```

This project compared the full multilevel approach to the analysis of NAEP data, used in the previous project, with the "separate estimation" approach of first fitting the null item model and then holding the item parameters fixed and estimating the reporting group regression model. The reporting group estimates from the separate estimation approach were good approximations to the full ML estimates, provided that the full four-level survey design was used in the null item model estimation, but their standard errors were seriously underestimated. Separate estimation gave a negligible saving in computation time.

The importance of this project was that it showed that it is essential to allow for the survey design in any analysis of NAEP data and that the separate estimation approach seriously underestimated the standard errors of the regression parameter estimates.

Identification of ability distributions in IRT models for NAEP items

`http://www.ms.unimelb.edu.au/~maitkin/TechReports/NAEP_ability1.pdf`

This project examined the identifiability of the ability distribution underlying the item responses and assessed the extent to which the current analysis could be made more effective and/or simplified as a result of this examination.

The importance of this project was that it showed that the estimates of upper-(individual) level parameters by Gaussian quadrature, used currently in the NAEP analysis for 3PL and other models, were very robust to various degrees of non-normality of the ability distribution, and that more complex semi-parametric and fully nonparametric forms of estimation did not improve the upper-level parameter estimates and were much more computer-intensive.

Investigation of the ability distribution in the NAEP 1986 math survey

`http://www.ms.unimelb.edu.au/~maitkin/TechReports/NAEP_ability2.pdf`

This project examined the effect of varying the distributional model for student ability on the reporting group parameter estimates and standard errors for the 1986 NAEP math scale. Several parametric distributions (including some extreme examples) and the nonparametric maximum likelihood estimate (for the two-level model) were used instead of the normal distribution; these resulted in only small changes in parameter estimates and very small changes in standard errors. We concluded that the normal distribution and Gaussian quadrature provide a robust analysis for the estimation of reporting group parameters that is almost unaffected by the actual form of the ability distribution.

The importance of this project was that it established the robustness of the ML estimation approach to regression coefficient estimation in real NAEP data against

variation in the form of the ability distribution. (This robustness did not, however, extend to percentiles of the ability distribution, as reported above.)

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

`http://www.ms.unimelb.edu.au/~maitkin/TechReports/NAEP_3PL.pdf`

This project found that the 3PL (three-parameter logit) model could be successfully fitted and identified with dense item data – five 2PL items and five 3PL items on 1000 subjects. Fitting the 2PL model to all ten items in the simulations gave biased estimates of reporting group parameters; smaller but still serious biases resulted from fitting 3PL models with incorrectly specified guessing parameters.

The importance of this project was that it raised a serious problem, that not fitting the 3PL model when it was known to be necessary could lead to serious bias in regression model parameter estimates, but the 3PL model might itself be unidentifiable.

Efficient maximum likelihood estimation in large-scale multilevel models

`http://www.ms.unimelb.edu.au/~maitkin/TechReports/NAEP_computation.pdf`

This project examined the facilities of current packages and algorithms for NAEP-scale maximum likelihood multilevel modeling to see if they could be extended or adapted for fitting psychometric models with a multilevel structure. The report gave a list of features that are needed to achieve this in a suitable package and a sequential list of developments that need to be carried out to achieve a suitable analysis system.

The importance of this project was that it defined the necessary features for efficient multilevel IRT ML data analysis.

Development of tools for the analysis of NAEP data

`http://www.ms.unimelb.edu.au/~maitkin/TechReports/NAEP_Analysis_Tools.pdf`

This project aimed to define the properties of a general NAEP data analysis system, assess current packages for this purpose, and evaluate the efficient computational methods needed.

The project began by evaluating Gllamm, the Stata program for fitting very general latent class and latent variable extensions of generalised linear mixed models. The program is remarkably flexible and general, and has been widely used in many contexts, particularly for fitting longitudinal random effect models. It demonstrated

(for the first time, as far as we know) the feasibility of running four-level hierarchical models with *simultaneous* estimation of 2PL item parameters, reporting group regression model parameters and variance components at each level. Its disadvantages were its slow runnng and the need for profile likelihood computations for the 3PL model.

We then evaluated Latent Gold 4 and 4.5. The developers have implemented a recursive multilevel version of the Bock-Aitkin EM algorithm for binary and categorical items that accommodates the four-level structure of the 1986 NAEP survey (and the three-level structure of current state NAEP surveys). This program is very fast, can fit very large models that include mixtures, and has flexible model specification facilities.

Multidimensional ability

`http://www.ms.unimelb.edu.au/~maitkin/TechReports/NAEP_multidimensional.pdf`

This project aimed to examine differences in reporting group estimates resulting from full multidimensional ability estimation, separate estimation of each scale and the combination of estimates, and the estimation of a single overall ability scale for the 70 items in three scales in the 1986 NAEP math data.

This report examined models with two and three latent ability dimensions. The simulation studies showed that:

- The critical issue was that the multifactor structure was recognised: fitting a single-dimension ability model did not give acceptable reporting group parameter estimates except for highly correlated abilities.
- Fitting a model with uncorrelated abilities, with or without the correct loading restrictions, gave very good estimates, closely equivalent in bias and precision to those for the correct model: the failure to account for the interability correlation in the two-ability models did not perceptibly affect the properties of the parameter estimates for these models.
- Fitting independent ability dimensions was fast, simple, and effective compared with the difficult estimation of the interability correlation.

With the NAEP math data with three subscales, the maximised log-likelihood increased dramatically from one to three dimensions, but the changes in the reporting group parameter estimates were relatively small and the changes in standard errors were very small.

The importance of this project is that it showed that it *was* necessary to fit the multidimensional model: fitting a single-dimension ability model could give biased estimates of the reporting group parameters. However, estimation of the interability correlations in a joint estimation of all dimensions was not necessary to obtain good reporting group parameter estimates and standard errors. The *actual* bias in the NAEP reporting group estimates from the one-dimension model was, however, quite small.

Investigation of alternative models for guessing

`http://www.ms.unimelb.edu.au/~maitkin/TechReports/NAEP_guessing.pdf`

This project examined methods for identifying the 3PL guessing model in sparse NAEP data and investigated alternative ability-based models for guesssing which are more easily identified in NAEP data.

We made four major contributions in this project:

- We extended the 2PL model to polynomial models and fitted the three-parameter (quadratic "3QL" model) and four-parameter (cubic) models to both simulated data and the 1986 NAEP math data. The quadratic model can serve as a model diagnostic for failure of the 2PL model and is very easily fitted. It gave a substantial improvement in fit over the 2PL model for the (3-level) 30-item NAEP 1986 math data.
- We were able to identify the 3PL model for the (three-level) 1986 NAEP math data. This gave an even larger improvement over the 2PL model than the 3QL model.
- We were able to identify on the NAEP data an even larger four-parameter model with both guessing and quadratic terms – we called this the 3PQL model. It gave a further substantial deviance reduction over the (three-level) 3PL model.
- We were able to fit a two-component mixture ability model – a mixture of two normal distributions – to the 1986 NAEP math data and later to state samples from the 2005 NAEP math data.

We examined two-component mixtures with several forms:

- One component was for *guessing*, in which the discrimination parameters for all items were zero.
- The second model, a generalisation of the first model, allowed the (logistic) modeling of the probability of being in the guessing component.

Both these models give very substantial improvements over the 3PL model, suggesting that *guessing is not an item process independent for each item but a more systematic process that is individual-specific*.

- The third model, a further generalisation of the first model, allowed for *different but nonzero discrimination parameters in each component*.
- The fourth model, a generalisation of the third model, allowed the (logistic) modeling of the probability of being in each component.

These models gave further substantial improvements.

These analyses were extended to the entire math test set of 70 items, with additional results:

- A *three*-component mixture could be identified in which one component allowed for guessing and the *two* others had different nonzero discrimination parameters.

- In one of these two components, the items were found to be much harder than in the other component.
  - Membership in the guessing component was strongly related to girls.
  - Membership in the "harder" component was strongly related to Black and Hispanic ethnicity and to attendance at low-status metropolitan schools.
  - Membership in the "easier" component was strongly related to White ethnicity, to attendance at high-status metropolitan schools, and to parents who had some college education or were college graduates.

Analyses of the 2005 NAEP math data for Texas

New item models for engagement: simultaneous identification of engagement and adjustment of reporting group differences for non-engagement

We developed complex multilevel models for math achievement on the 2005 NAEP math data. Main-effect analyses were carried out on the California and Texas state samples with a selection of school and teacher variables available in this survey as well as the available reporting group variables. We fitted all the models described in the `NAEP_Guessing` project to the data for these two states for the age 4/grade 10 survey.

We were able to estimate achievement differences using absence from school, teacher qualifications and experience, and school location, as well as the major gender and ethnic reporting groups. With the mixture models, we were able to establish the important variables for identifying membership in the guessing or second component and to assess also the effect on reporting group differences of adjusting these effects for component membership.

Chapter 5
Model-Based Analysis of the 1986 NAEP Math Survey

We describe first the analysis of the 30-item Numbers and Operations – Knowledge and Skills subscale. The item questions are given in Appendix A, Table 1. (This subscale is a small version of the test subscale that had 180 items across the three age groups: most of these items were not used for age 9/grade 3.)

5.1 Data and model specification – subscale

There were 21,290[1] grade 3/age 9 students in the survey, but only about half of these had responses on any of the items on the Knowledge and Skills subscale, so the "full" data set for this subscale had 10,460[1] students clustered in 440[1] schools, which were themselves clustered in 94 PSUs. High-minority-proportion PSUs, and hence high-minority schools and minority students, were oversampled to ensure adequate minority student samples. This oversampling does not require weighting in the analysis as the PSU, the school identifier, and the student ethnicity are retained in all model analyses except those using the two-level model, in which the school is not identified, and the three-level model, in which the PSU is not identified.

The number of students per school varied from 5 to 45, with an average of 24, and an average of seven items per student – the data set is very sparse. The item responses were coded 0 or 1 according to the manual, with items skipped coded zero and items not reached omitted from the data set, as in the original analysis.

We used a set of important reporting variables: sex, race (six levels), geographical region (four levels), size and type of community (stoc, seven levels) and parents' education (pared, six levels), to establish the feasibility of the modeling approach and to give a comparison with the published NCES results. We used a main-effect ability or achievement model with these categorical variables – 20 dummy variables.

[1] This number has been rounded to the nearest 10 in accordance with IES's Statistical Standards Program.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 63
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_5, © Murray Aitkin and Irit Aitkin 2011

For reference, we give here the stoc categories; other variables are defined in the tables.

```
1 Extreme Rural
2 Low Metropolitan
3 High Metropolitan
4 Main City
5 Urban Fringe
6 Medium City
7 Small Place
```

In analysing the test data, we evaluated all the item response models discussed in Chapter 3: the Rasch, 2PL, MIMIC, 3PL, and two-component mixture models, and a two-dimensional 2PL model. The Rasch model is not generally used for NCES test items but is of historical interest and provides a base for assessing the value of the discrimination parameters in the two-parameter models. The 2PL and MIMIC models have different forms for the regression, and although they have the same number of parameters, they can be distinguished with sufficient data. Since they do not allow for guessing, they are also not generally used for NCES test items, but they are included in this chapter to show the improvement in data fit of the three- and four-parameter models. The 3PL and three- and four-parameter mixture models compete for the explanation of guessing or engagement and provide more general forms of response heterogeneity. In subsequent chapters, we present detailed results for only the 3PL model and the mixture models.

We ran all three- and four-level Rasch, MIMIC, and 2PL models in Gllamm. The three- and four-level 3PL and mixture models were fitted at a later stage of the project in Latent Gold. Null models (items but no reporting group variables) were also fitted for all models to establish the size of the variance components and the importance of the regression.

Tables of parameter estimates, standard errors, and maximised log-likelihoods are given in Appendix A, Tables A3, A4, and A7, for the two-, three-, and four-level models, respectively. We give the estimates in some detail, comparing them across the psychometric models at each level of the multilevel model. Variance components are reported on the original normal (0,1) scale.

The parameter estimates and SEs have been rescaled from the $N(0,1)$ scale to the NAEP reporting scale with standard deviation 35. There is no intercept estimate given since it is confounded with one of the item difficulty parameters. The "reference" level is therefore *arbitrary*; this is set by NCES for the definition of the NAEP scale as a population mean of 250 and standard deviation of 35 across the three age groups comprising the population answering items on this scale.

The estimates and SDs we report represent *differences* from the reference (first) category of each variable (male, White, Northeast, extreme rural school, parents did not finish high school) on the NAEP scale. Population estimates aggregated across schools and ethnic groups require weighting, but we do not report weighted estimates because the appropriate weights are difficult to determine from the complexity of their construction in the survey report.

We present first the maximised log-likelihoods and other aspects of the models and follow this with a detailed discussion of the reporting group differences.

5.2 Model aspects

5.2.1 Maximised log-likelihoods

The maximised log-likelihoods for all models reported are given in Table 1 below with the number of parameters.

Table 5.1 Maximised log-likelihoods and number of parameters

Model	two-level	three-level	four-level
Rasch	−40475.22	−40349.04	−40342.85
	51	52	53
MIMIC	−40080.32	−39961.68	−39954.46
	80	81	82
2PL	−40077.26	−39930.05	−39924.60
	80	81	82
3PL		−39848.49	
		111	
2-guess		−39777.88	
		112	
2-guess-prob		−39711.77	
		132	
2-D 2PL		−39685.69	
		111	
2-mix		−39649.40	
		142	
2-mix-regressions		−39623.38	
		162	
2-mix-prob		−39547.41	
		162	

5.2.2 Two- and three-level models

The log-likelihood improvement for the three-level model over the two-level model was large in every case – 126 for the Rasch, 118 for the MIMIC, and 147 for the 2PL. The two-level model does not provide a correct representation of the clustered survey design. (A formal asymptotic test treats the deviance difference – twice the log-likelihood difference – as a mixture distribution: $0.5\chi_0^2 + 0.5\chi_1^2$. The deviance difference vastly exceeds the critical value – for a 1% level test, the 2% point of χ_1^2 is 5.4.)

The intraclass (school) correlations are 0.117 for the Rasch model, 0.113 for the MIMIC, and 0.076 for the 2PL. These are not large values, but they affect the standard errors quite substantially. These are increased by 50% for variables at the school level (such as region and size and type of community). The standard errors for parents' education are also increased by the same proportion, although this is apparently a student-level variable. This probably reflects the homogeneity of educational level amongst parents of children in the same school.

The two-level 2PL model fitted the test data somewhat better than the two-level MIMIC model – a log-likelihood difference of 3.06. For the three-level model, the difference is much greater: 31.63 in favour of the 2PL. There is no current valid test for comparing the MIMIC and 2PL models.

5.2.3 Four-level models

The log-likelihood improvement for the four-level model over the three-level model is much smaller in every case: 6.19 for the Rasch, 7.22 for the MIMIC, and 5.45 for the 2PL. The difference is statistically significant at the 0.1% level for all three. The effect of the fourth-level estimation on the model parameter estimates and SEs was small – the SEs of the region estimates increased by about 20%, while those of the parents' education estimates *decreased* by about 5%, because the school variance component was reduced by the fourth-level modeling. The parents' education estimates all increased consistently (relative to the reference category), by 0.5 for the Rasch model, 0.4 for the MIMIC, and 0.2 for the 2PL. These changes represent 25% of a standard error for the Rasch model, 20% for the MIMIC, and 10% for the 2PL.

The gender and ethnic origin parameters and standard errors were very stable from three to four levels. The other parameters were little affected – the largest change was in the size and type of community variable, with a change of 33% of a standard error in one parameter. The three-level model provided a substantially correct representation of the clustered survey design.

The 2PL model fitted much better than the MIMIC – a log-likelihood improvement of 29.86 for the same number of parameters.

5.2.4 The 3PL model

This model is the main model used by NCES for NAEP binary response data. We fitted it to all 30 items; a (boundary) zero estimate for the guessing parameter reduces the model to the 2PL or MIMIC model for these items. The notorious identification difficulties of this model led us to expect unidentifiability. This occurred in the three-level analysis for the MIMIC version of the three-parameter model, which failed to achieve convergence despite detailed searching over 50 random starting points.

However, we were able to identify the 2PL form of the three-parameter model with no difficulty; it gave a maximised log-likelihood of −39848.49, an improvement of 81.56 (an equivalent χ^2 of 163.12) for the additional 30 parameters. Although there is no formal test for this special two-component mixture model against the single-component 2PL model, there seems little question of the inadequacy of the two-parameter 2PL model.[2]

The guessing parameters are shown in Table A5 together with those from Technical Report 1988 (Table E6, p. 576). In our analysis, the guessing parameters are estimated, with standard errors, on the logit scale. We have converted them to the probability scale; we do not give standard errors on this scale for estimates close to zero. The estimated item parameters are not directly comparable with those given in the Technical Report for any of the item response models because the latter were obtained from a null achievement model across a different population – the combined age group populations taking these items (Technical Report 1988, p. 221). Our (simultaneous) analysis also controls for a regression model for ability or achievement.

The results gave 13 unidentifiable ("infinite") parameters on the logit scale; we interpreted these as guessing parameters that were approaching zero. We refitted the model with these c_j parameters fixed at 0, together with two other items with large logit values (≥ 5 in magnitude); values less than 5 were retained. The maximised log-likelihood for this model was unchanged to 3 dp by the 15 constrained guessing parameters and converged much faster.

It is of interest that the original NAEP analysis identified 12 items with nonzero guessing parameters, 4, 6, 9, 10, 15, 19, 20, 24, 27, 28, 29, and 30, whereas the 3PL analysis here identified 15 items with nonzero guessing parameters, ten of the original 12 plus five others: items 1, 2, 3, 7, and 11. Items 4 and 19 have zero guessing parameters in our analysis.

For the reporting group estimates, those from the 3PL model are almost all larger than those from the 2PL model, as are their standard errors and variance components. The reason is clear: the compressed logit scale for the 3PL model, and the large number of items with nonzero guessing parameters, mean that effects on the 3PL model logit scale must be larger than on the 2PL model logit scale to reproduce the data fit. The two sets of parameter estimates are not directly proportional, but they are very similar in their relations within each set. The improved fit of the 3PL model does not appear to change much the relative differences in the *reporting group categories*.

5.3 Reporting group differences

We first note that the parameter estimates for the MIMIC and 2PL models are not directly comparable since they refer to two different relations of the covariates: *ability*

[2] For the 3PL and mixture models, we did not use the four-level analysis, as the three-level analysis provided essentially the same results for all the parameters except for small changes in the region differences.

(MIMIC) or *achievement* (2PL). Nevertheless, they are closely related in magnitude since, as noted in §3.2, the MIMIC ability regression model can be re-expressed in 2PL form as an interaction achievement regression model on the logit scale:

$$\begin{aligned} \operatorname{logit} p_{ij} \mid \theta_i &= a_j(\theta_i - b_j), \\ \theta_i &\sim N(\beta' \mathbf{x}_i, 1), \end{aligned}$$

is equivalent to

$$\begin{aligned} \operatorname{logit} p_{ij} \mid \theta_i &= a_j(\theta_i - b_j) + a_j \cdot \beta' \mathbf{x}_i, \\ \theta_i &\sim N(0, 1). \end{aligned}$$

If the discriminations do not vary substantially, the effects of the explanatory variables will be very similar across items, and the two sets of regression model parameter estimates will be very similar, though there may be a proportionality factor between them.

A second important point, from comparing the two- and three-level models, is the large increase in standard errors for all the variables above the student level – the region, school location, and parents' education standard errors increase by around 50%. The school variance component is about 12% of the total variance (more for the 3PL model), and so allowing for the school clustering of students greatly reduces the apparent precision of these estimates from that of the two-level model. This is a *major advantage* of the multilevel model approach – it *automatically* corrects for the clustering, without any need to tailor the *design* of the survey to the estimation of the design effect. In the four-level model, the PSU variance component is much smaller – around 2% of the total variance – and so the effect on the standard errors of the region estimates is much smaller, though appreciable. Below we interpret the estimates from the three-level analysis (this also corresponds to the three-level design in the 2005 survey analyses in Chapters 7 and 8).

Thus the complexity of the survey design, and the considerable effort required in the jackknifing of PSUs to allow for the PSU part of the design in the original analysis, were largely wasted, since the clustering at this level is small. *No* allowance for the school clustering was made in the original analysis, and this implies that reported estimates from the conditioning model and plausible value generation are *overprecise*. This is borne out by the few comparisons possible with this subscale, given below.

From Table A4 of Appendix A, it is clear that the four models give very similar reporting group estimates and standard errors, with the 3PL being most different from the others because of the scale restriction imposed by guessing. A notable exception is the sex difference (in favour of girls), which is around 4 NAEP scale points, and very significant, for the Rasch and MIMIC models, but less than 1 point, and nonsignificant, for the 2PL and 3PL models. The better fit of the 2PL and 3PL models to the test data implies that the sex difference appearing in the other models should be discounted. This emphasises the importance of goodness of fit compar-

isons using the maximised likelihood for these models, though there is no general theory for comparing non-nested models (the MIMIC with the 2PL, for example).

Following this principle, we interpret the estimates for the three-parameter model. Following common frequentist practice, we use the term *significant* (strictly, significant at the asymptotic 5% level) to refer to an estimate that is about twice its standard error or greater (in magnitude). This corresponds to rejection of the null hypothesis of a zero regression coefficient for the corresponding covariate using an asymptotic 5% level test in the frequentist framework. Estimates that are less than twice their standard errors in magnitude are called nonsignificant, and may not be mentioned in all summary lists to save space. To summarise the three-parameter model estimates:

- The sex difference was small and nonsignificant.
- The Black mean was 29 points below, and the Hispanic and American Indian means were 21 points below, the White mean. The Asian/Pacific Islander mean was nonsignificantly below the White mean.
- Students in the Central and Western regions were significantly below (7.5 points) those in the Northeast region, and those in the Southeast were nonsignificantly below those in the Northeast.
- Students in low metropolitan schools were significantly below (12 points) those in extreme rural schools, who were not significantly different from those in main city, medium city, urban fringe, and small place schools; students in high metropolitan schools were 14 or more points above the last group, 20 points above those in extreme rural schools, and 32 points above those in low metropolitan schools.
- Students whose parents had at least some college education had a mean 18 points above those whose parents had not finished high school; this difference was nearly significant at the 5% level.

5.3.1 Comparison with NAEP subscale estimates

The NAEP results for the Numbers and Operations – Knowledge and Skills subscale were unpublished but were kindly provided by ETS. The NAEP tables relevant to our results are one-way tabulations of plausible values by Gender, Race/Ethnicity, School Type, Parental Education, and Region.

Our model used Size and Type of Community (stoc), and not School Type. The one-way tabulations will in general give results different from the main-effect model estimates since any correlation between the reporting group variables will affect the marginal mean differences. (This is the reason for fitting all reporting groups together in a main-effect regression model.)

Furthermore, the oversampling of high-minority PSUs and schools means that only the ethnic group one-variable model will give results comparable with the NAEP tabulations – the other one-way models will implicitly collapse over the eth-

nic origin variable and so our model results would need to be weighted for the differential sampling rates. This requires knowledge of the PSUs and schools that are high-minority, but these PSUs and schools are not identified in the data file, so we are unable to correct for the oversampling. For the one-way MIMIC ethnic group model, the maximised log-likelihood is −40082.02, while for the one-way 2PL model it is −40040.14. As for the full model, the 2PL model fits the 30-item scale better than the MIMIC model.

To compare the results, we present the NAEP tables in the same form as the model parameters, with the reference category mean subtracted and standard errors calculated from the variances of the difference between the means. We omit the "other" category for ethnicity, and the "no-response" category for parents' education, which had very few members. We present the 2PL parameters from the three-level one-way model (those from the MIMIC model were very closely similar) and give the 3PL estimates from the full three-level model. None of the model estimates corresponds exactly to those from the NAEP tables since these used the 3PL MIMIC conditioning model (apart from the other differences with plausible values).

	NAEP table		3PL estimates		2PL 1-way	
Variable	MLE	SE	MLE	SE	MLE	SE
male	0		0			
female	2.2	(1.5)	0.8	(1.2)		
white	0		0		0	
black	-31.1	(2.1)	-29.3	(2.1)	-32.5	(2.2)
hisp	-22.4	(2.4)	-21.3	(1.9)	-24.0	(2.2)
asia/pac	- 3.5	(5.9)	- 7.8	(4.9)	-10.4	(6.0)
amerind	-13.0	(3.4)	-21.1	(4.2)	-23.1	(4.7)
NE	0		0			
SE	- 2.4	(2.1)	- 2.6	(3.9)		
Cent	- 1.3	(2.3)	- 7.4	(3.2)		
West	- 4.6	(2.5)	- 7.6	(3.0)		
notfinhs	0		0			
finhs	7.8	(4.0)	2.0	(9.3)		
somcoll	22.8	(4.3)	17.1	(9.3)		
collgrad	24.9	(3.8)	18.3	(9.6)		
DK	13.1	(3.7)	2.0	(9.3)		

There are notable differences among the various sets of estimates and standard errors. First, the standard errors from the one-way 2PL model are uniformly slightly larger than those from the 3PL model (and from the full 2PL model). This is a

consequence of the poorer fit of the one-way models, as shown by their maximised likelihoods – the other variables are very important and reduce the random variation that otherwise inflates the standard errors.[3] Second, the ethnic group differences relative to White are uniformly overstated by the one-way model – when the other variables are taken into account, the ethnic group differences are reduced.

Third, the standard errors of the model-based estimates are substantially larger for the variables above the student level than those for the NAEP table estimates – nearly double for the regional differences and more than double for parents' education. For the student-level sex and ethnic group comparisons, the model-based standard errors are slightly *smaller* than those for the NAEP table estimates. This pattern is very similar to that for the standard errors from the two-level model compared with the three-level model, for the same reason: the school clustering of students is not accounted for in the NAEP table analysis.

It might be thought that the standard errors of estimates in the full model should be *larger* than those in the one-way model because of the correlations of the additional covariates with ethnic group and with each other. However this is not the case: the correlations of the parameter estimates are given for the three-level Rasch model in Appendix A (those for the 2PL model were very similar). The correlations of the reporting group parameter estimates are generally very low, apart from the positive intercorrelations for the parameters for each category of a single variable.[4] So the fitting of the full reporting group model does not induce larger standard errors for this model.

We draw several important conclusions from these comparisons:

- The parameter standard errors for upper-level variables are considerably understated in the NAEP table analysis because they are *not adjusted for the large school design effect*.
- One-way summaries of reporting group differences, whether from plausible values or from direct model fitting, are averages across other important variables and may be different from the results from the *simultaneous fitting of all the reporting group variables of interest*.
- The simultaneous fitting of the important variables gives a fairer picture of their importance than separate one-way tabulations.

5.4 Mixture models

In the 30-item subscale analysis, we fitted five mixture models using the 2PL form of the two-parameter model, as this was uniformly superior in maximised likelihood

[3] The full model *controls each main effect for the others*; the one-way models *aggregate over the other main effects*.

[4] An exception is the positive correlation, around 0.2, of all the stoc category estimates above category 1 with the Central (region 3) estimate. This may reflect a higher proportion of extreme-rural communities in this region than in the other regions.

to the MIMIC form of the model. We fitted the 2-guess, 2-guess-prob, 2-mix, 2-mix-regressions, and 2-mix-prob models, described in §3.5.

5.4.1 2-guess model

This model had a maximised log-likelihood of −39777.88 with 110 parameters, an improvement of 70.61 over the 3PL model, for one extra mixing parameter. The proportion in the guessing component is estimated to be 0.219, with a 95% confidence interval (0.184, 0.259). Reporting group parameter estimates are given in Table A8 (the model headings are abbreviated to 2-g and 2-g-p). The standard errors are uniformly smaller than those for the 3PL model – the large improvement in log-likelihood reduces the random variation and consequent standard errors. The parameter estimates are similar to those of the 2PL model and are mostly smaller than the 3PL estimates, as the 2-guess model uses the full 0–1 scale for the response probability in the engaged group (as we noted in §3.3). A notable change is in the sex effect, which is now 2.8 units in favour of girls and significant.

The estimated guessing parameters, on the probability scale, are given in Table A9 with those for the 3PL and their standard errors from Table A5. For fair comparison with those for the 3PL model, we need to multiply the estimated d_j for the mixture guessing model by the estimated c of 0.219, the (constant) proportion in the guessing component. We do not report standard errors of the d_j or the product $c \cdot d_j$.

The estimates are bounded by the estimated c parameter. There is no clear relation between the two sets of guessing parameters; the (ML) estimates of the probability of a correct answer by guessing for the 2-guess model are considerably more variable than for the 3PL model.

5.4.2 2-guess-prob model

This model further improves over the 2-guess model, with an increase in maximised log-likelihood of 66.11 for the 20 additional parameters in the engagement regression model. The parameter estimates are similar to those for the 2-guess model, though generally slightly smaller. An exception is the sex effect, which changes back to a nonsignificant 0.8 units in favour of girls.

Significant effects associated with engagement from the engagement logistic regression model were Black and Hispanic ethnicity and Southeast region (all negative) and high metropolitan and urban fringe schools (positive). Engagement did not vary by sex and parents' education.

The estimated probabilities of belonging to the engaged component for each category of student can be evaluated by summing the logit estimates for the relevant categories and transforming to the probability scale. For the "most engaged" category (girl, White, Northeast region, high metropolitan school, some college), the

logits are

$$-1.087 + 0.086(\text{girl}) + 0(\text{White}) + 0(\text{NE}) + 0.811(\text{himet}) + 0.497(\text{smcoll}),$$

totaling 0.307. The corresponding probability of being engaged is

$$e^{0.307}/(1 + e^{0.307}) = 0.576.$$

For the "least engaged" category (boy, Black, Southeast region, low metropolitan school, not finished high school), the logit total is

$$-1.087 + 0(\text{boy}) - 0.856(\text{Black}) - 0.349(\text{SE}) - 0.542(\text{lomet}) + 0(\text{nfhs}),$$

totalling -2.834, with corresponding engagement probability 0.056.

These estimates have corresponding variability bounds (standard errors), which we do not give here, though they can be calculated from the program output; the main issue is that engagement has a rather low probability even for the most engaged category. (The estimated logits given above include some for categories that do not show significant variation – sex and parents' education. If the reference category value of zero is used instead for these variables, both the estimates and their standard errors will be less extreme.)

The estimated d_j parameters on the probability scale are also given in the last column of Table A9. The estimated d_j are generally larger than those for the 2-guess model. We do not give the guessing parameters $c_i \cdot d_j$, as these depend on the individual student variables $\mathbf{x}_i$.

5.4.3 Two-dimensional model

We fitted a two-dimensional 2PL model to investigate the "subscale purity" of the items on the Numbers and Operations – Knowledge and Skills subscale. We did not expect to find a significant second dimension, given the relatively small number of items and their assignment to the subscale in terms of their content.

It was therefore surprising to find a really large improvement in maximised log-likelihood, to -39685.69 with 111 parameters. This represents an increase of 244.36 over the 2PL model for 30 extra parameters and an increase of 92.19 over the 2-guess model with one *less* parameter. Compared with the 2-guess-prob model, the improvement is smaller, 26.08, but the number of parameters is 21 less. Compared with the 3PL model, the improvement is very large: 163 with the same number of parameters!

In fitting this model, as discussed in Chapter 3, we made no assumption about the assignment of items to dimensions – the two-dimensional model was completely general. Since this model is rotationally invariant, like a two-factor normal response model, we are not able to identify or assign items to the two dimensions. The

achievement regression parameter estimates and standard errors (given in Table A8) are quite similar to those for the 2-guess model.

5.4.4 2-mix model

This model improved substantially over the 2-guess model, with a maximised log-likelihood of −39649.40, an increase of 128.48 for 30 extra parameters. It also improved by 36.29 over the two-dimensional 2PL model for 31 extra parameters. Estimates and standard errors are given in Table A10. The first component contains an estimated 46% of the tested population and the second contains the other 54%. The achievement regression parameter estimates are quite close to the 2-guess estimates, apart from the sex difference, which is nonsignificant. The standard errors are slightly larger.

The item parameters in the two components of the mixture are given in Table A11 with standard errors. Inspection of the estimates shows that the large majority of items were found easier by those in component 2 than by those in component 1. So component 2 was a subpopulation of more able students. Very few of the items had nonsignificant discrimination parameters: items 27 and 28 in component 1 and items 1, 2, and 12 in component 2. Items 27 and 28 were difficult in both components, items 1, 21, and 22 were very easy in both components, and items 2 and 12 were easy in component 2.

Neither component can be interpreted as a guessing component, as is also indicated by the large change in maximised log-likelihood between the 2-guess and 2-mix models.

5.4.5 2-mix-regressions model

This model improved marginally over the 2-mix model, with a change of 26.02 in maximised log-likelihood for 20 extra parameters. We do not give estimates or interpretation for this model, as it was substantially inferior to the 2-mix-prob model, discussed below. We do not consider it further.

5.4.6 2-mix-prob model

This model improved further over the 2-mix model, with a change of 101.99 in maximised log-likelihood for 20 extra parameters. Estimates and standard errors are given in Table A10. The achievement regression parameters showed some notable changes relative to the 2-mix estimates. The Black, Hispanic, and American Indian

differences from White were all reduced, and neither the school location nor parents' education differences were significant.

Significant effects associated with membership in the "easy items" component from the logistic regression model were Black, Hispanic, and American Indian ethnicity, Southeast and West regions, and low metropolitan schools (all negative) and high metropolitan and urban fringe schools (both positive). Component membership did not vary by sex and parents' education.

The estimated probabilities of belonging to the easy items component for each category of student can be evaluated by summing the logit estimates for the relevant categories and transforming to the probability scale. For the "most easy" category (girl, White, Northeast region, high metropolitan school, some college), the logit total is 0.674, with a corresponding probability of being in the easy items category of 0.842.

For the "least easy" category (boy, Black, Southeast region, low metropolitan school, not finished high school), the logit total is -1.823, with corresponding probability 0.140.

5.4.7 Conclusions from the 30-item analysis

The conclusions from this array of model analyses may appear somewhat confusing. There are several logical sequences of nested models of increasing complexity:

- Rasch $->$ 2PL $->$ 3PL
- 2PL $->$ 2-guess $->$ 2-mix
- 2PL $->$ 2-guess $->$ 2-guess-prob
- 2PL $->$ 2-mix $->$ 2-mix-prob
- 2-guess-prob $->$ 2-mix-prob
- 2PL $->$ two-dimensional 2PL

Other comparisons are not nested:

- 2PL $<->$ MIMIC
- 3PL $<->$ 2-guess

In large samples with small models, where large-sample (asymptotic) theory can be assumed to apply, the nested comparisons can be tested using the likelihood ratio test, which we have quoted frequently above.

For models that are not nested, we lack a general repeated-sampling theory for such comparisons. But even for the nested comparisons, we have the opposite situation: though the sample size is apparently large, the information per student is very limited – an average of seven items – and the larger models are heavily parametrised. There is strong reason therefore to doubt the validity of the direct application of the likelihood ratio test to these comparisons.

This presents a dilemma – if we cannot decide which model is best supported by the data, how can we draw conclusions about group differences? Fortunately there are some external factors, which we have already mentioned:

- The 2-guess model, which has only one more parameter than the 3PL model, is much better supported – by 70.61 for one extra parameter.
- The two-dimensional model is much better supported than the 3PL model with the same number of parameters, and better supported than the 2-guess model with one less parameter.

It is therefore clear, if we can rely on these measures of support, that the 3PL is not the most appropriate model. Amongst those considered, one of the 2-guess-prob, 2-mix, 2-mix-prob, or two-dimensional models, is most appropriate. To be more definite, we need theoretical developments in model comparisons.

In the next chapter, we extend this analysis to all 79 items on the test, from three scales.

Chapter 6
Analysis of All 1986 Math Iems

6.1 The full math test

As we noted in Chapter 2, in the 1986 test for age 9/grade 3 students there were four main scales,

- Numbers and Operations (56 items),
- Measurement (27),
- Fundamental Methods (17),
- Data Organization and Interpretation (16),

with a total of 116 items. The large Numbers and Operations scale was itself split into two subscales:

- Knowledge and Skills (30 items),
- Higher-Level Applications (26).

These two subscales, and the Measurement scale, were also reported on, though in less detail than the composite dimension, which was based on the Numbers and Operations and Measurement scales. In Chapter 5, we found, surprisingly, that the 30-item Numbers and Operations – Knowledge and Skills subscale appeared to be at least two-dimensional.

In this chapter, we report the analysis of the 1986 age 9/grade 3 math survey, with the full set of 79 items from the Measurement scale and the two Numbers and Operations subscales that were reported for the NCES analysis. We used three models of different dimensions for the analysis: a single dimension for all items, a two-dimensional model, and a full three-dimensional model. For the two- and three-dimensional models, the scales of the items were not distinguished. This allows for "nonpurity" of the items. We give parameter estimates for only the 1-D and 3-D models; the three models were very similar.

The multidimensional model was fitted only as a 2PL model; for the MIMIC model we would have had to specify the ability regressions on each dimension.[1]

[1] For the 2PL model, the achievement regression is on the logit scale of item response and is unaffected by the number of dimensions.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 77
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_6, © Murray Aitkin and Irit Aitkin 2011

We also extended the two-component mixture models to three components for both guessing (with one guessing component) and general mixture models.

We fitted the four-level two-parameter model; as with the 30-item scale in Chapter 5, the PSU-level variance component was very small (0.013), and the school variance component was 0.054, considerably smaller than for the 30-item scale. The maximised log-likelihood improved by only 6.37 for the one extra parameter. Both PSUs and schools were more homogeneous with the full set of items. Parameter estimates for the four-level model are given in Table B0, together with those for the three-level two-parameter model. Agreement in both estimates and standard errors is very close, and the other psychometric models were fitted as three-level models, ignoring the PSU sampling level.

The one-dimensional (3PL) and three-dimensional (2PL) analyses correspond to the official analysis of the full survey, apart from a slight difference in the weighting of the three scales in producing a single composite scale: our analysis using a single dimension is automatically weighted by the number of items used on the scales.

The models and approach used are the same as for the analysis of the single 30-item Numbers and Operations – Knowledge and Skills subscale, but the data set used is larger: at least one of the 79 items was answered by 12,180[2] students; the number of schools increased by 0[2] to 440[2]. Most importantly, the average number of items answered per student across the three scales more than doubled, to 15.4. This greatly improves the precision of the information about student ability/achievement, and therefore the precision of parameter estimates in all the regression models, and it allows the identification of some three-component mixture models.

We consider first the 2PL models.

6.2 2PL models

The maximised log-likelihoods for all three-level 2PL models reported are given in Table 1 below; the first line gives the four-level two-parameter model.

Reporting group parameter estimates and SEs are given in Appendix B, Tables B1 and B2; for those models that model the component membership, the logistic model parameters are given in Table B2. A notable feature of all the models is the reduction in the school variance component from the 30-item subscale (14%) to the full test (6%). This suggests that the variation among schools in the items on the other scales in the full test is much less than that on the items in the Numbers and Operations – Knowledge and Skills subscale. This may mean that the teaching of the topics assessed by the Measurement and the Higher-Level Applications items was more consistent across schools than that of the topics assessed by the Numbers and Operations – Knowledge and Skills items.

A statistical concern is the *very* large number of parameters needed for the 3-component mixture models.

[2] This number has been rounded to the nearest 10 in accordance with IES's Statistical Standards Program.

Table 6.1 Three-level 2PL model maximised log-likelihoods and number of parameters

Model	log L	# params.
2-param(4-level)	−103007.14	180
2-param	−103013.51	179
3-param	−102404.15	258
2-guess	−102682.84	259
2-guess-prob	−102513.79	279
2-mix	−102186.27	337
2-mix-prob	−102186.27	337
3-guess	−102380.03	418
3-guess-prob	−101482.05	458
3-mix	−101817.46	497
3-mix-prob	−101320.81	537
2-D	−102219.51	257
3-D	−102014.57	336

6.3 Results

Tables B1 and B2 show a surprising similarity of reporting group parameter estimates and standard errors across the 2PL, 3PL, 2-guess, 2-mix, 3-guess, 3-mix, and 3-D 2PL models, which is surprising considering the large changes in maximised log-likelihood (discussed below). Following Chapter 5, we give the summary results for the 3PL model, with those from the 30-item scale from Chapter 5, where they are different, in square brackets [ ].

- The sex difference was small and nonsignificant.
- The Black mean was 17 [29] points below, and the Hispanic and American Indian means were 14 [21] points below, the White mean. The Asian/Pacific Islander mean was 6 points [nonsignificantly] below the White mean.
- Students in the [Central and] Western regions were significantly below (5.4 [7.5] points) those in the Northeast region, and those in the Southeast were nonsignificantly below those in the Northeast.
- Students in low metropolitan schools were significantly below (8 [12] points) those in extreme rural schools, who were not significantly different from those in main city[, medium city, urban fringe] and small place schools.
- Students in high metropolitan schools were 7–9 [14–16] points above those in main city, urban fringe, or medium city schools, 11 [20] points above those in extreme rural or small place schools, and 19 [32] points above those in low metropolitan schools.
- Students in urban fringe and medium city schools were 3–4 points above those in extreme rural schools and 11–12 points above those in low metropolitan schools.
- Students whose parents had at least some college education had a mean 13–14 [18] points above those whose parents had not finished high school [this difference was nearly significant at the 5% level] and 9 points above those whose parents had finished high school.

The important point here, apart from the expected similarity in conclusions, is that the group differences found on the 30-item subscale in Chapter 5 are all *larger* than those on the full set of items here. It follows that the group differences on the items in the two additional scales in the full test must be much smaller than those on the Numbers and Operations – Knowledge and Skills scale items.

6.3.1 Mixed 2PL models

As we noted above, the 2-guess, 2-mix, 3 guess, and 3-mix 2PL models gave reporting group estimates and standard errors very similar to the 3PL model. Estimated proportions of students in the guessing component, or the "hard items" component (with 95% confidence intervals where available), were

- 2-guess: 0.177 (0.159, 0.197),
- 2-guess-prob: 0.325 (from the sum of the estimated probabilities across all students),
- 2-mix: 0.416 (0.391, 0.442),
- 2-mix-prob: 0.476 (from the sum of the estimated probabilities across all students),
- 3-guess: 0.186 (guessing), 0.426 (hard), 0.388 (easier),
- 3-mix: 0.327 (hardest), 0.363 (easier), 0.310 (easiest).

However, the two models with a two-component mixture that included modeling of the component membership probability gave results quite different from those above though quite similar to each other. In these two models:

- Ethnic group differences from White were *halved* relative to the other models.
- Students in high metropolitan schools were only 4 points above those in main city or urban fringe schools, 6 points above those in extreme rural schools, and 13 points above those in low metropolitan schools.
- Differences in the effects of parents' education (relative to not finished high school) were roughly *halved* relative to the other models.
- *All* other ethnic groups had higher probabilities than Whites of being in the guessing or the lower-ability component.
- Students in low metropolitan schools had higher probabilities of being in the guessing or lower-ability component, and those in high metropolitan schools had higher probabilities of being in the engaged or higher-ability component than students in extreme rural schools.

As in Chapter 5, the estimated probabilities of belonging to the engaged or the "easy items" component for each category of student can be evaluated by summing the logit estimates for the relevant categories and transforming to the probability scale. For the "most engaged" category (White, high metropolitan school, some college), the logit total is 0.933, with the corresponding probability 0.718 of being

engaged. For the "least engaged" category (Black, low metropolitan school, not finished high school), the logit total is −3.238, with corresponding probability 0.038.

Results for the "easy items" membership were very similar: for the "most easy" category (White, Northeast region, high metropolitan school, college graduate), the logit total is 2.148, with the corresponding probability 0.895. For the "least easy" category (Black, West region, low metropolitan school, not finished high school), the logit total is −2.413, with corresponding probability 0.082.

The substantial changes in reporting group differences can be interpreted informally and formally. Informally, the interpretation is clearest for the guessing model. If we exclude students who are guessing from the computation of group differences, and if these guessing students are disproportionately from the non-White ethnic groups, whose mean achievements are below those of Whites, the exclusion process will reduce these group differences.

Formally, we can express the results in terms of *direct* and *indirect* effects in the structural equation modeling sense. The achievement regression gives the direct effects of the reporting group variables on achievement. But there are also indirect effects resulting from the disproportionate membership of the other ethnic groups in the guessing component from the membership regression. These indirect effects, when combined with the direct effects in the achievement model, give the previous results for the 2-guess and 2-mix models. In these models, membership in the guessing or lower-ability component has *the same probability for every student*, so these models cannot represent the full effect of guessing.

6.3.2 Three-component membership models

For the full 1986 survey item set, we were able to fit and identify 2PL three-component mixture models with modeling of the component probabilities. The 3-guess models had one guessing component and *two* engaged components, and the 3-mix models had *three* engaged components.

The 3-guess-prob model improved by 897 in log-likelihood over the 2-guess-prob model for 179 extra parameters and improved by 597 over the 3-guess model for 40 extra parameters.

The achievement and engagement model parameter estimates are given in Table B3, and the item parameter estimates are given in Table B4. There is a striking reduction in *nearly all* the achievement model estimates across the five variables. The Black and Hispanic differences from White are reduced to 4 and 3 NAEP scale units, respectively, and the difference between White and American Indian is reduced to 7 units. There are *no* significant school location differences from the reference extreme rural category (though the overall χ^2_6 test for homogeneity is significant at the 1.1% level), and the parents' education differences from the not finished high school category are reduced to 5 points for the two college education categories.

These are quite dramatic changes. They are produced by the large parameter estimates in the two membership models. The first component is the guessing com-

ponent, and it is clear from the item parameter estimates in Table B4 (given as logit intercepts and slopes) that students in the second component found the items very difficult, while those in the third component found them much easier. The estimated proportions in the three components (the sums of the estimated probabilities of component membership across all students) were 0.214 (1), 0.327 (2), and 0.459 (3).

What defines the component membership? This can be understood from Table B3 by three patterns of signs in the component membership columns: $--$, $+-$, and $-+$. We see that:

- The guessing component ($--$) is strongly identified with girls.
- The "difficult items" component ($+-$) is strongly identified with Black, Hispanic, and American Indian students and less strongly identified with Southeast and Western region students and students in low metropolitan schools.
- The "easier items" component ($-+$) is strongly identified with students in high metropolitan schools and students whose parents had at least some college education.

For the 3-mix-prob model, the maximised log-likelihood increased over the 3-guess-prob model by 161 for 79 extra parameters. The achievement estimates (in Table B3) were similar to those for the 3-guess-prob model but generally larger. The item parameters (not given) in components 2 and 3 were similar to those for the 3-guess-prob model, but those for the first component all had large *negative* slope parameters, and many items also had negative intercept parameters. Apart from being inconsistent with item behaviour, these parameters showed that the estimated model was close to reproducing the guessing model parameters with zero slope. We therefore concluded that, whatever the significance of the (relatively small) improvement in maximised log-likelihood, this model was not well supported. We did not interpret it further.

6.3.3 Multidimensional ability model

The 3-D model gave a vast improvement (1000) in maximised log-likelihood over the 1-D 2PL model for its extra 158 parameters. Despite this improvement, its reporting group parameter estimates and standard errors were very close to those of the 1-D model (and those of the mixed models in Table B1). The 3-D model improved by 172 over the 2-mix model, which has one more parameter. The 2-mix-prob model was better by 221 for 21 extra parameters. The 3-mix model improved by 197 in maximised log-likelihood over the 3-D 2PL for its 161 extra parameters.

6.4 MIMIC models

The maximised log-likelihoods for all MIMIC models reported are given in Table 2 below.

Table 6.2 MIMIC model maximised log-likelihoods and number of parameters

Model	log L	# params.
2-param	−102567.07	179
3-param	-	-
2-guess	−102322.81	259
2-guess-prob	−102211.16	279
2-mix	−101686.16	337
2-mix-prob	−101607.02	337
3-guess	−101548.04	418
3-guess-prob	−101393.33	458
3-mix	-	-
3-mix-prob	-	-
2-D	-	-
3-D	-	-

The MIMIC model improves over the 2PL for all the comparisons available by from 89 to 832 units of maximised log-likelihood. This improvement for the full set of items compared with the improvement of the 2PL over the MIMIC for the 30-item subscale suggests that the items on the other scales have properties different from those on the Knowledge and Skills subscale. We have not investigated this issue.

Reporting group parameter estimates and SEs are given in Appendix B, Tables B3 to B5. Since (as we noted previously) there is no "standard" three-parameter MIMIC model to use as a reference, we give the changes in the two-parameter MIMIC model from the 30-item subscale (in square brackets [ ]) to the full item set first and then give the changes with increasing MIMIC model complexity.

6.5 Results

6.5.1 Two-parameter MIMIC model

- The sex difference was small and nonsignificant [girls were 4 points above boys].
- The Black mean was 25 [22] points below, the Hispanic and American Indian means were 19–20 [16] points below, and the Asian/Pacific Islander mean was 12 points [nonsignificantly] below the White mean.
- Students in the Western region [and Central region] were 7 points below those in the Northeast region.

- Students in low metropolitan schools were 9 [11] points below those in extreme rural schools, who were not significantly different from those in main city[, urban fringe, medium city] and small place schools.
- Students in high metropolitan schools were 13 [10-12] points above those in main city, urban fringe, or medium city schools, 18 [15] points above those in extreme rural or small place schools, and 27 points above those in low metropolitan schools.
- Students in [main city,] urban fringe and medium city schools were 6 points above [not significantly different from] those in extreme rural schools and 16 [15] points above those in low metropolitan schools.
- Students whose parents had at least some college education had a mean 20 [19] points above those whose parents had not finished high school and 13 [12] points above those whose parents had finished high school; students whose parents had finished high school were 7 points [not significantly] above those whose parents had not finished high school.

The MIMIC family reporting group estimates varied more across models than the 2PL family estimates and were larger, though they are not directly comparable.

6.5.2 Mixed MIMIC models

The MIMIC 2-guess model improved over the MIMIC model by 183 in maximised log-likelihood for the 80 extra parameters. The MIMIC 2-mix model gave a further large improvement of 636 for its extra 79 parameters. Modeling the component membership in the 2-guess-prob model improved over the 2-guess model by 111 for the 20 extra parameters. The 2-mix-prob model increased this by a further 604 for the 79 extra parameters. It is clear that the big improvements come from modeling the component probabilities.

For the three-component models, the 3-guess model improved over the 2-guess model by 774 for the extra 159 parameters, and the 3-guess-prob model improved by a further 103 for the 44 extra parameters. The full 3-mix MIMIC models could not be reliably identified.

Parameter estimates for the two three-component models are given in Table B7. Those for the 3-guess-prob model compared with the 3-guess model behave quite differently from those for the corresponding 2PL models. Modeling the component membership reduced the regional and parents education differences slightly, reduced the ethnic group ability differences to a greater extent, and reduced the school location differences substantially, but gave a large (11 point) negative sex difference.

Component membership was more complicated than for the corresponding 2PL model. Membership in the guessing component $(--)$ was strongly related to Black, Hispanic, and Asian/Pacific ethnicity and low metropolitan schools and less strongly to the Central region. Membership in the "easier items" component $(+-)$ was

strongly related to girls, high metropolitan schools, and parents with at least some college education. Membership in the "difficult items" component was not well identified by any variable.

6.6 Comparison with published NAEP results

Results for the full 1986 age 9/grade 3 test can be found in various NCES publications. Table 124 from the 2001 Digest of Education Statistics (National Center for Education Statistics) gives the average mathematics proficiency, by age and selected characteristics of students, for the 1973–1999 surveys. For the 1986 survey, the table gives, for 9-year-olds, the following means and standard errors (in parentheses), which we have extended by giving the mean differences from the reference category with their standard errors:

Male	221.7 (1.1)	
Female	221.7 (1.2)	0.0 (1.6)
White	226.9 (1.1)	
Black	201.6 (1.6)	-25.3 (1.9)
Hispanic	205.4 (2.1)	-21.5 (2.2)
Not high school graduate	200.6 (2.5)	
Graduated high school	218.4 (1.6)	17.8 (3.0)
Some education after high school	228.6 (2.1)	28.0 (3.3)
Graduated college	231.3 (1.1)	30.7 (2.7)
Northeast	226.0 (2.7)	
Southeast	217.8 (2.5)	- 8.2 (3.7)
Central	226.0 (2.3)	0.0 (3.5)
West	217.2 (2.4)	- 8.8 (3.6)

As discussed in Chapter 4, because of the weighting used when averaging over the ethnic categories that were differentially sampled, only the estimated ethnic group differences are consistent between the NAEP report and our analysis. These differences and their standard errors are quite close to those from the two-parameter MIMIC model in Table B5. For the other variables, the standard errors are noticeably *larger* in the NAEP report than those found from the MIMIC model, and some

of the estimates are quite different, though this may be due to the effect of aggregation over the ethnic group categories.[3]

6.7 Discussion

The comparison of the multilevel modeling results with the published tables is not very informative because of the aggregation needed in the presentation of one-way group differences. In the modeling approach, the group differences are *adjusted* for the other variables in the model – for example, the parents' education differences are differences *within* the categories defined by the combinations of the other variables. This reduces the large differences in the one-way tabulation (of 18, 28, and 30 NAEP units) to smaller *within-group* differences (of 7, 20, and 21 units) from the MIMIC model.

The five-variable model we have used can be interpreted as a *five-way tabulation* of mean achievement by the variables, but with a main-effect structure. The possibility of *interactions* between the variables has not been investigated, partly because of the slow computational speed of the analysis in its early stages and later because the mixture models were already heavily parametrised in the item parameters.

With the larger sample and the full set of NAEP items, the issue of whether there really are, on this test, mixtures with two or even three components – subpopulations defined by item parameter heterogeneity – becomes even more pressing. It affects the conclusions we would like to draw about how best to express the group differences in achievement or ability. Unfortunately, we are at the edge of current statistical theory in dealing with this issue. Progress on it depends on the development and application of a statistically sound and general model comparison procedure, discussed in Chapter 9. This is also true for the comparison of 2PL, 3PL, MIMIC, mixture, and multidimensional models.

The 1986 survey is restricted in its model analysis and is of limited relevance educationally after 25 years. In the following chapters, we apply the same methods to more comprehensive models for a recent survey of Texas and California grade 4/age 10 children in the 2005 math survey. We chose these states as they had the largest state NAEP samples, though the samples are substantially smaller than the 1986 national NAEP sample. This leads to identification difficulties for some of the models used in this chapter for the larger NAEP sample; other states with even smaller state samples will give even less precise results, and additional models may not be identifiable. We suggest how to deal with this problem in Chapter 9.

[3] The NAEP report did not publish size and type of community differences, given in our analysis, only the public/private school variable. As we have seen, the community location of the school is an important variable in both the achievement/ability and membership regression models.

Chapter 7
Analysis of the 2005 NAEP Math Survey – Texas

7.1 Population, sample, and test

The data analysed were from the Texas subsample of the 2005 national NAEP math survey for the age 10/grade 4 cohort, using the 70-item Numbers and Operations scale. As we noted in Chapter 1, the survey design was changed after 2001 to use states as the primary sampling units, and schools were sampled within the state with probability proportional to size, with oversampling of high-minority schools and minority students within schools. It is not clear from the NAEP technical manuals how the school design effect is allowed for in the current analysis. For the multilevel model analysis, the separate state analyses mean that there is no PSU sampling level, so only the three-level model is needed.

For the 2005 analyses, we used a selection of covariates at the teacher and school levels from the teacher questionnaire, in addition to the student and family demographic variables and a few school and teacher variables on the student questionnaire. After discussion with NCES staff, these covariates were chosen as relevant variables for achievement from the perspective of educational theory. However, a number of schools did not complete the teacher questionnaire, and so the complete-case analysis we used required the omission of the schools with incomplete teacher questionnaires.

To ensure that the student sample was used effectively in the analysis, we therefore made *two* analyses in Texas and California: one with the "full" school sample but only minimal teacher information obtained from the student questionnaire, and one with comprehensive teacher information but for the "reduced" subset of schools for which this information was available. This was not an optimal solution but one imposed by the difficulty of handling the extensive missingness of the teacher information. The consequences of this compromise are discussed in the final section of this chapter, and suggestions for dealing with it are given in Chapter 9.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 87
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_7, © Murray Aitkin and Irit Aitkin 2011

Without teacher information except for years of experience, the two-stage design of students sampled in Texas schools gave 360[1] schools with 7460[1] children (an average of 21 students per school) who attempted at least one item from the Numbers and Operations scale. The average number of scale items per student was 12.5, giving a moderately dense data set (compared with the 1986 survey). The average number of students attempting each item was 1330[1].

When teacher information in addition to years of experience was included, the number of schools and students with complete teacher data decreased dramatically (24% of the schools but almost 50% of students were lost) because of extensive incompleteness in the teacher questionnaire. Only 260[1] schools with 3860[1] students (an average of 14.7 students per school) had complete teacher data: large schools are under-represented in this analysis. The average number of items per student was, however, the same (12.6). The average number of students per item was only 690[1], so item parameters are probably less well estimated in the reduced data set.

Despite this, the standard errors of the estimated regression parameters were not uniformly larger than in the first data set and were smaller for some of the school-level variables, and this analysis could (potentially) identify the important teacher variables.

7.2 Variable names and codes

The variables used in the models are listed in Tables 1 and 2 below by their NAEP code, number of levels, definition, and name used in this report. Detailed descriptions of the multi-category variables are given in Appendix C (the "count" variables given there are the sample sizes for the full national survey).

7.3 Models fitted

We used the same models examined for the full 1986 math survey, but we report results here for only the five models that can allow for guessing or engagement: the 2PL form of the three-parameter model (the MIMIC version of this is not identifiable) and the four two-component mixture models (2-guess, 2-guess-prob, 2-mix, and 2-mix prob). Three-component mixture models could not be identified with the smaller sample and much larger regression models.

A general feature of the Texas analysis is that, for any of the models we examined except the 3PL (discussed in the next paragraph), the MIMIC model version fitted very much better than the 2PL model version, by several hundred units in maximised log-likelihood for the same number of parameters. This was consistent with the analysis using the 79 test items for the three main scales of the 1986 math survey.

[1] This number has been rounded to the nearest 10 in accordance with IES's Statistical Standards Program.

Table 7.1 Student and school variables

Code	Levels	Definition	Name
IEP	2	Disability	Disable
SEX	2		Sex
TOL9	8	School location	Locate
ACCOM2	2	Accommodation	Accom
SDRACEM	5	Race/ethnicity	Race
SENROL4	4	School enrollment	Enrol
LEP	2	Limited English proficiency	ELL
B017101	2	Computer at home	Comp
B018101	5	Days absent from school last month	Absence
B018201	4	Language other than English spoken in home	Lang
M814701	2	Use computer to play math games	Games
SLUNCH	4	Natl School Lunch Prog eligibility	Lunch
YRSEXP	4	Years taught elementary or secondary	Exp

Table 7.2 Teacher variables

Code	Levels	Definition	Name
T056301	6	Highest academic degree	Degree
T077309	3	Undergrad major/minor mathematics education	Mathed
T077310	3	Undergrad major/minor mathematics	Math
T087601	4	Mathematics education courses	Mathc
T088901	2	Technical support available at school	Tech
T089001	2	Software for math instruction available at school	Soft
T089101	2	Training for computers available at school	Train
T092401	5	Number of students in this class	Studnum
T092301	4	Instructional materials, resources	Resourc

As a consequence, we do not report results and interpretations for the 2PL models; in Appendix C we give a summary table (separately for the data sets with minimal and extensive teacher data) of the maximised log-likelihoods for all the models we examined.

As noted above, the three-parameter MIMIC model proved unidentifiable for the 2005 Texas sample, as for the 1986 sample, while the three-parameter form of the 2PL model was readily identified. Our estimates of the guessing parameters for this may therefore be different from those given in the NCES analysis for the three-parameter MIMIC model[2]; some differences would be expected in any case from the separate estimation of these parameters from the null regression model rather than our joint estimation.

[2] Which are not estimates but are fixed by tight priors.

7.4 Results – limited teacher data

Parameter estimates and standard errors are given in Table C1 of Appendix C. As in previous chapters, the estimates and SEs are given on the NAEP reporting scale, apart from the variance components, which are given on the original $N(0,1)$ scale for ease of assessment of the variance component among schools. Reading across the columns makes it easy to compare the effect sizes in different models and their importance for NAEP reporting.

The three-parameter MIMIC model is again unidentifiable, so we interpret the three-parameter 2PL model. All differences are relative to the first category in NAEP units.

7.4.1 Three-parameter interpretation

Student characteristics

- a disability difference of −29.1 (2.8);
- a limited English proficiency difference of −20.1 (2.3);
- an accommodated difference of −13.1 (2.9);
- a sex difference in favour of boys of 7.0 (1.3);
- a White–Black difference of 39.7 (2.5);
- a White–Hispanic difference of 20.0 (2.2);
- a home computer difference of 6.4 (1.7);
- absence from school in the month effects of
  - 1–2 days: −10.1 (1.5);
  - 3–4 days: −10.7 (2.2);
  - 5–10 days: −15.6 (3.5);
  - $>$ 10 days: −30.1 (3.9).
- other language at home effects (relative to never) of
  - once in a while: +3.5 (1.7);
  - half the time: +4.7 (2.3);
  - all or most of the time: +5.0 (2.0).

School characteristics

- school lunch program
  - reduced price lunch: −14.3 (2.7);
  - free lunch: −22.3 (1.8).

Teacher characteristics

years of teaching experience (relative to 0-4)

- 5–9 years: +5.2 (1.9);
- 10–19 years: +6.1 (1.9);
- $\geq$ 20 years: +6.5 (2.1).

Variables not reported here (school location, school enrolment, playing math games on the computer) were found to be nonsignificant.

7.4.2 Mixture models

The mixture models proved increasingly difficult to fit with increasing dimension. Multiple maxima are a common feature of mixture likelihoods, and we used many sets of starting values for the model parameters, continuing the analysis with those best in log-likelihood.

For the 2-guess model, which fitted the data very slightly better than the 3PL model, the estimated proportion in the guessing component was 0.082, with 95% confidence interval [0.060, 0.112]. It gave very similar reporting group estimates and slightly smaller standard errors. The only major difference was in significant school location differences of −9.4 (3.7) for schools on the fringe of mid-size cities and −11.2 (4.1) for rural (non-Metropolitan Statistical Area) schools. There was a significant positive difference of 10.7 (3.9) for Asians/Pacific Islanders relative to Whites; for the 3PL model, this was 7.9 (4.1).

The 2-mix model fitted better than the 2-guess model by 156 units in maximised log-likelihood for the 70 extra parameters. The estimated proportion in the "hard items" component was 0.362, with 95% confidence interval (0.225, 0.525). It gave generally similar reporting group estimates also, with standard errors very similar to those for the 3PL model.

The 2-guess-prob model improved over the 2-guess model by 112 units in maximised log-likelihood for the 33 extra parameters. The estimated proportion in the guessing component (summing over the probabilities for all students) was 0.113. It gave two *boundary estimates* (discussed in the next section) in the membership model where cell sample sizes were small. Standard errors increased over those for the 2-guess model.

Group differences in the ability regression were reduced relative to the 3PL model, and the teacher experience differences were not significant. Membership in the guessing component was significantly associated with

- girls;
- Black and Hispanic ethnicity;
- Asian/Pacific Islander ethnicity (*negatively*);
- schools on the fringes of large cities and in small towns and rural areas, both MSA and non-MSA;

- teachers with ten or more years of experience (*negatively*);
- no computer at home;
- absent 5–10 days in the last month;
- language other than English spoken at home once in a while (*negatively*);
- schools with free lunches.

We do not list here (as we did in Chapters 5 and 6) the membership probabilities for the extreme categories, though these can be calculated from the tables of parameter estimates. The range of the logit scale covered is much larger because of the larger number of model variables, and the membership probabilities cover the full range from 0 to 1 (though with wide confidence intervals).

However, as with the 1986 full item set, the modeling of membership in the guessing component illuminates the direct and indirect effects of the covariates. In the 2-guess model, students with teachers who had five or more years of experience scored 4–6 NAEP units above students with teachers with less than five years of experience. In the 2-guess-prob model, these score differences were reduced to nonsignificance because they are now assessed on students who were *not* in the guessing component. These students are more likely to be taught by teachers with five or more years of experience and have other advantages that produce the improvement of students taught by these teachers when we eliminate the membership explanation in the 2-guess model.

The 2-mix-prob model fitted better than the 2-mix model by 84 units of maximised log-likelihood for the 33 extra parameters. The estimated proportion in the "hard items" component (summing over the probabilities for all students) was 0.485. It gave one boundary estimate in the membership model where cell sample sizes were small. Standard errors increased further over those for the 2-mix model. Some reporting group estimates were substantially different from those in the 2-guess-prob model in both the ability and membership regressions. In the ability regression,

- girls were 21.2 (5.3) NAEP units below boys;
- students in schools on the fringe of mid-size cities were 17.5 (3.8) units below and those in rural non-MSA schools were 12.7 (3.1) units below those in extreme rural schools;
- students taught by teachers with 5–10 years of experience were 5.0 (2.3) units above those taught by teachers with less than five years of experience.

Membership in the "hard items" component was significantly associated with

- students with a disability;
- boys;
- Black students;
- students with limited English proficiency;
- no home computer;
- absence from school for 1–2 days or more than 10 days;
- a language other than English spoken at home at least half of the time (*negatively*);
- schools with free lunches.

Some of these effects are puzzling – the very large sex effect in the ability model and the reversal of the sex difference in membership, for example. We did not feel that interpretation of this model was reliable and so do not comment further here.

7.5 Boundary values in logistic regression

In both the 2-guess-prob and 2-mix-prob models, the modeling of the component membership probability sometimes led to a *boundary value* on the logit scale (equivalent to $\pm\infty$) for one or more of the component membership model parameters. The boundary value resulted from the sparsity of data in the large models, analogous to a binary response logistic model with a cell with only zero outcomes in a cross-classification model.

The estimate is "infinite", but so is its formal standard error from the information matrix. This makes it difficult to assess the actual importance of this variable category in the logistic regression even if one knows the cell sample size on which it is based.

While this is not formally a model identification difficulty (it is a failure of the usual standard error calculation from the information matrix), it appears to lead to the formal absolute classification of students in these categories as in the appropriate component *with probability 1.0*. This conclusion is unsound, and without an effective standard error or other means of estimating the variability of the estimate we adopt a conservative approach by *omitting these students* from the prediction of the membership model. The membership model prediction is used only for students *not in* the category (or categories) with the infinite estimate(s).

For the larger data set with minimal teacher information, the same school lunch category gave a boundary estimate in both the 2-mix-prob and the 2-guess-prob models, and the school location category 5 (large town) gave a boundary estimate for the 2-guess-prob model. Standard errors were not grossly inflated in either model, and we regard these models as interpretable, excluding the students in the nonparticipating school lunch category (3.9% of students) or in large towns (1.7% of students).

7.6 Results – extensive teacher data

We consider again the three main models above: the 3PL, 2-guess, and 2-mix models. With the smaller sample size but larger model for the extensive teacher data, the 2-mix-prob model became uninterpretable, with very large standard errors indicating near-collinearity of the variables in the information matrix and boundary values in the membership regression model. We therefore do not give estimates for or discuss this model. The estimates and standard errors for the three other models are given in Table C2 of Appendix C.

Parameter estimate differences among these three models were again fairly small, with the standard errors of the 3PL estimates substantially larger than those of the two others. A notable feature of all the models was the large standard errors for the teacher degree, math education, and math variables, suggesting that these are highly intercorrelated. However, omitting any of these variables did not reduce the large standard errors.

The standard errors for the 2-mix model fluctuated wildly, with some much larger than those of the other models and some much smaller. It appeared that this model was near the limit of identifiability, and the standard error calculation was breaking down, giving unreliable standard errors.

As in previous chapters, we report the results for the 3PL model.

Student characteristics

- a disability difference of −22.5 (4.1);
- a limited English proficiency difference of −15.3 (3.0);
- an accommodated difference of −22.5 (4.2);
- a sex difference in favour of boys of 6.4 (1.9);
- a White-Black difference of 45.6 (3.5);
- a White-Hispanic difference of 25.6 (3.0);
- a home computer difference of 6.8 (2.4);
- absence from school in the month effects of
  - 1–2 days: −14.8 (2.1);
  - 3–4 days: −14.8 (3.3);
  - 5–10 days: −14.9 (5.0);
  - $>$ 10 days: −27.0 (6.4).
- other language at home effects (relative to never) of
  - once in a while: +5.3 (2.5);
  - half the time: +6.6 (3.2);
  - all or most of time: +5.3.

School characteristics

- school lunch program
  - reduced price lunch: −14.3 (3.8);
  - free lunch: −20.9 (2.6);
  - not participating: −61.5 (30.3);
- technical support: 9.0 (4.0) (*negative*).

Teacher characteristics

- years of teaching experience (relative to 0–4)
  - 10–19 years: +5.9 (2.8).

Variables not reported here were found to be nonsignificant.

Several points are striking in this analysis of the extensive teacher data set relative to the limited data set:

- Almost no teacher variables from the extensive set were apparently important.
- Standard errors for effects were nearly doubled for many variables relative to the full NAEP sample because of the almost halved student sample size.
- The White–Black and White–Hispanic differences were *increased considerably*, which probably reflects the nonrandom subsample of schools, in which large schools are under-represented.
- The disability and limited English proficiency differences *decreased*, but the accommodated difference *increased*.
- The sex and home computer differences were little changed.

7.6.1 Mixture models

The 2-guess model fitted better than the 3PL by 41 maximised log-likelihood units for one extra parameter. The estimated proportion in the guessing component was 0.069, with 95% confidence interval [0.048, 0.090]. Parameter estimates were generally very similar to those for the 3PL model, but standard errors were generally smaller. One notable difference from the 3PL model was in the class size (number of students in the class): there was a significant negative difference of 8.1 (SE 3.7) for class size 16–18 relative to class size 15 or fewer.

The 2-guess-prob model gave an improvement of 79 in maximised log-likelihood over the 2-guess model for the 56 additional parameters in the membership model. It gave similar ability regression effects, though most were reduced. (The class size difference in the 2-guess model was reduced to 7.3 (3.9).) Additional significant ability effects were:

- Large town schools were 19.9 (9.7) units above extreme rural schools.
- Schools with most resources were 6.6 (2.8) units above those with all resources.

Membership in the guessing component was positively associated with

- girls;
- Black or Hispanic ethnicity;
- large city fringe and rural non-MSA schools;
- teachers with 0–4 years of experience;
- no computer at home;
- absent from school 1–2 days in the last month;
- schools with free lunches;
- no school software for math instruction;
- schools with some or most resources (but not all).

For the 2-guess-prob model, several boundary values occurred in the membership model for categories of school location and teacher degree as well as school lunch program participation. Some standard errors were large, but many effects were important in the membership model, and we regard the 2-guess-prob model as interpretable.

These results were similar to those for the analysis with limited teacher data for the variables common to both data sets. Thus the school-level variables were more importantly associated with membership in the guessing component than with achievement for those *not* in the guessing component.

The 2-mix model fitted better than the 2-guess model by 156 units of maximised log-likelihood for its 70 extra parameters. The proportion in the "harder items" component was 0.321, with 95% confidence interval [0.253, 0.389]. This was close to the estimate for the limited teacher data. Parameter estimates and standard errors were very similar to those for the 2-guess model, apart from language effect estimates, which were larger.

Table C2 in Appendix C does not give the estimates for the 2-mix-prob model for the reduced data set with extensive teacher data. This model gave very large standard errors for almost all the parameters in both the achievement and the membership models. The very large standard errors showed that the variables in the two models gave very high intercorrelations for the corresponding parameters in the information matrix. This model was overly complex for the available data and was unreasonable as a representation of the test results, so we do not give an interpretation of this model.

7.7 Comparison with official NCES analysis

Reporting group results for the Numbers and Operations scale are not published formally. However, they can be computed for this scale (and other scales, as well as the composite score) as one-way tabulations using the NAEP explorer tool on the NCES Website. We give below the tabulations of some of these variables using this tool and the equivalent group differences relative to the first category, which can be compared with the results from the two analyses above for the (MIMIC) 3PL model in Tables C1 and C2. (Note that the NAEP tool gives mean scores only to integer precision.)

As discussed in Chapter 5 (§3.1), only the ethnic group differences are unaffected by the oversampling of high-ethnic-minority schools and minority students. The large models fitted in the multilevel analysis inevitably increase standard errors over those from one-way tabulations. However, the sex of the student tested is unlikely to be related to any of the other variables, so the sex difference in the modeling analysis should not be inflated by the larger models. For the "full" limited teacher data set, the standard error of the sex difference is 20% larger than in the NCES tool tabulation, and the standard errors for the ethnic group differences are 50% larger.

Comparison of NCES and modeling results - Texas

	NCES Explorer tool		Limited data		Extensive data	
Male	243 (0.8)					
Female	238 (0.7)	-5 (1.1)	-7.0 (1.3)		-6.4 (1.9)	
White	252 (1.0)					
Black	227 (1.3)	-25 (1.6)	-39.7 (2.5)		-45.6 (3.5)	
Hispanic	234 (0.7)	-18 (1.2)	-20.0 (2.2)		-25.6 (3.0)	
As/Pac	264 (2.5)	12 (2.7)	7.9 (4.1)		3.5 (6.0)	
Limited English prof.						
Yes	224 (1.1)					
No	243 (0.7)	19 (1.3)	20.1 (2.3)		15.3 (3.0)	
Teacher experience						
0-4	237 (1.1)					
5-9	240 (1.5)	3 (1.9)	5.2 (1.9)		3.3 (2.7)	
10-19	245 (1.2)	8 (1.6)	6.1 (1.9)		5.9 (2.8)	
20+	241 (1.4)	4 (1.8)	6.5 (2.1)		3.9 (3.0)	
Math education level						
Major	230 (6.7)					
Minor/ specialism	238 (3.6)	8 (7.6)			-10.4	(10.4)
Neither	241 (0.8)	11 (6.7)			-7.4 (9.9)	
Math level						
Major	233 (6.3)					
Minor	238 (5.0)	5 (8.0)			-1.1	(10.2)
Neither	241 (0.8)	8 (6.4)			-0.3	(10.1)
Highest degree						
Bachelor	240 (0.7)					
Master	243 (1.6)	3 (1.7)			1.7 (-)	
Education specialist	249 (5.8)	9 (5.8)			-1.3 (-)	
Training in computers						
Yes	242 (1.0)					
No	239 (1.2)	-3 (1.6)			0.9 (2.4)	

As with the 1986 survey, there appears to be no design effect correction for the school clustering of students, so the standard errors of the community and school level variables are probably understated as well in the NCES tool tabulations.

There is one peculiarity in the comparison. For the math education level of the teacher, the tool analysis shows an *increase* in mean score with *decreasing* level of math education, while the modeling analysis shows a *decrease* with decreasing math education, though neither of these trends is statistically significant. This may be simply an effect of aggregation in the tool analysis, but it highlights the importance of disaggregate analysis.

7.8 Conclusion

The analysis of the Texas NAEP sample shows the possibilities, and also the difficulties, of detailed analysis with models that attempt to represent the richness of the information available in the NAEP survey databases.

As with the larger sample in the 1986 survey, with its much smaller set of covariates, we were able to show that for the full data set with only limited teacher information, the 2-guess-prob engagement model gives a richer, and better-supported, interpretation of the test results than the 3PL model. The direct and indirect effects of the covariates on achievement, which are separated in this model, clearly show the importance of the model and the educational importance of dealing appropriately with guessing in formal NAEP analyses.

The more complex 2-mix-prob model could not be reliably interpreted in either Texas sample. The extensive information about teachers and schools was frustratingly missing for 25% of the schools but almost 50% of the students for a complete-case analysis. The importance of the extensive teacher covariates is affected to an unknown extent by the under-representation of large schools in the reduced sample.

This emphasises the importance of obtaining the information about the student and family variables in the incomplete data that is ignored in the complete-case analysis. We discuss this point at length in the final chapter.

Chapter 8
Analysis of the 2005 NAEP Math Survey – California

8.1 Population, sample, and test

The data for this analysis are from the same NAEP survey as for the Texas sample, so we do not repeat the description of variables, though a few changes occur in the variable levels. For example, all teachers in California must have a Bachelor's degree, so the reference category is 3 for this analysis. The full descriptions of variables are given in Appendix C.

We again needed to use two complete-case analyses, depending on whether detailed teacher information was included or not. With only limited teacher data (years of experience), the two-stage design of students sampled in schools gave 410[1] schools with 9260[1] children (an average of 22.4 students per school) who attempted at least one item from the Numbers and Operations scale. The average number of scale items per student was 13.0, giving a moderately dense data set. The average number of students taking each item was 1720[1], so item parameters are better estimated than for the Texas data.

When extensive teacher data in addition to years of experience were included, the number of schools and students again decreased (6.8% of the schools but 21.5% of students were lost) because of incompleteness in the teacher questionnaire.

There were 390[1] schools with 7270[1] students (an average of 18.8 students per school) with complete data; large schools are slightly under-represented in this data set. The average number of items per student was, however, the same (13.0). The average number of students per item was 1350[1], so item parameters are less well estimated in the extensive teacher data set.

[1] This number has been rounded to the nearest 10 in accordance with IES's Statistical Standards Program.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 99
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_8, © Murray Aitkin and Irit Aitkin 2011

8.2 Models

We report here in detail on only the four models that can represent guessing or engagement, though the other models (Rasch, 2PL, MIMIC) were also fitted. Maximised log-likelihoods for all models are given in Table 8.1 below.

Table 8.1 Maximised log-likelihoods – MIMIC and 2PL models

Model	Limited teacher data ($n = 9260^1$)	Extensive teacher data ($n = 7270^1$)
Rasch	−64329.76	−50510.92
# params	108	128
2-param (2PL)	−64058.94	−50240.10
# params	177	197
2-param (MIMIC)	−63494.47	−49844.81
# params	177	197
3-param (2PL)	−63399.96	−49732.73
#params	247	267
MIMIC 2-guess	−63385.73	−49743.69
# params	248	268
MIMIC 2-guess-prob	−63216.12	−49610.72
# params	285	324
MIMIC 2-mix	−63148.74	−49569.81
# params	318	338
MIMIC 2-mix-prob	−63047.61	−49443.73
# params	354	398

As in the Texas sample, the MIMIC versions of the models were very much superior to the 2PL versions, and so we report only the results from the former family (except for the three-parameter model, for which the MIMIC version could not be identified).

8.3 Results

Results are tabulated in Appendix D.

8.3.1 Limited teacher data

Parameter estimates and standard errors for the MIMIC and the four models for guessing, together with the maximised log-likelihoods, are shown in Table D1. We follow the Texas analysis and give the interpretation of the California 3PL estimates, with the Texas results in square brackets [ ].

8.3.2 3PL interpretation

Student characteristics

- a disability difference of −26.9 (2.7) [−29.1 (2.8)];
- a limited English proficiency difference of −29.5 (1.6) [−20.1 (2.3)];
- an accommodated difference of −32.4 (3.4) [−13.1 (2.9)];
- a sex difference in favour of boys of 8.4 (1.2) [7.0 (1.3)];
- a White–Black difference of 32.7 (2.5) [39.7 (2.5)];
- a White–Hispanic difference of 17.9 (1.9) [20.0 (2.2)];
- a White–American Indian difference of 24.3 (5.9) [18.7 (11.1)];
- a *negative* White–Asian/Pacific Islander difference of 5.1 (2.1) [7.9 (4.1)];
- a home computer difference of 6.8 (1.6) [6.4 (1.7)];
- absence from school in the month effects of
  - 1–2 days: −9.0 (1.3) [−10.1 (1.5)];
  - 3–4 days: −14.2 (1.9) [−10.7 (2.2)];
  - 5–10 days: −12.0 (2.7) [−15.6 (3.5)];
  - $>$ 10 days: −33.6 (3.3) [−30.1 (3.9)].
- other language at home effects (relative to never) of
  - once in a while: 0.5 (1.6) [3.5 (1.7)];
  - half the time: 9.1 (2.2) [4.7 (2.3)];
  - all or most of the time: 9.4 (1.7) [5.0 (2.0)];
- did not use computer to play games: 6.9 (1.3) [not significant].

School characteristics

- school location (relative to large city)
  - fringe of mid-sized city: −7.8 (3.6) [not significant]
- school enrollment (relative to 1–299)
  - size 700+: −11.6 (5.7) [not significant]
- school lunch program
  - reduced price lunch: −9.1 (2.3) [−14.3 (2.7)];
  - free lunch: −18.9 (1.6) [−22.3 (1.8)].

Teacher characteristics

years of teaching experience (relative to 0–4)

- 5–9 years: 3.0 (1.7) [5.2 (1.9)];
- 10–19 years: 5.8 (1.8) [6.1 (1.9)];
- $\geq$ 20 years: 7.4 (2.1) [6.5 (2.1)].

The group differences for California relative to those for Texas are mixed (some larger, some smaller), but the general structure is very similar.

8.4 Mixture models

The 2-guess model improved in maximised log-likelihood by 14 over the 3PL for one extra parameter. The estimated proportion in the guessing component was 0.095, with 95% CI [0.074, 0.119]. Parameter estimates were very similar to those for the 3PL model, and the standard errors were smaller. The only notable difference was in the accommodation effect, which was smaller in the 2-guess model.

We report in Table D3 the item parameter estimates in the guessing and engaged components, and compare them in terms of a "cut" value. We expect that the probability of a correct answer on an item for a student in the guessing component should be surpassed by the probability of a correct answer for a student in the engaged component, for a sufficiently high ability of the engaged student.

This is easily determined from an inequality on their logit functions: if $\phi_1 = \alpha_1$ is the logit function for the non-engaged student, and $\phi_2 = \alpha_2 + \beta_2\theta$ is the function for the engaged student, then $\phi_2 > \phi_1$ when the ability of the engaged student exceeds the cut value $\theta_C = (\alpha_1 - \alpha_2)/\beta_2$, where $\theta = 0$ is the average ability.

In Table D3, nine items have *positive* cut values (the largest is 0.9), indicating that above-average ability is required to score "better than chance" on these items. However the great majority of items have quite large *negative* values, showing that for most items even low ability students scored better than chance on most items by engaging in the item task.

The 2-guess-prob model improved by 170 in maximised log-likelihood over the 2-guess model for the 36 additional parameters in the engagement model. The estimated proportion in the guessing component was 0.169, summing over the individual student membership probabilities. There was one boundary estimate for the large town category in the membership model where cell sample sizes were small. Standard errors of the ability regression estimates increased over those for the 2-guess model.

Group differences in the ability regression were similar to those of the 2-guess model, except for a large reduction in the disability effect. Membership in the guessing component was significantly associated with

- students with a disability or an accommodation;
- girls;
- Black or Hispanic ethnicity;
- no computer at home;
- absent 1–2 or 5 or more days in the last month;
- language other than English spoken at home all or most of the time (*negatively*);
- using the computer to play games;
- teachers with 20 or more years of experience (*negatively*);
- schools with free lunches.

All of these variables appeared in the corresponding Texas analysis except for the disability and accommodation variables.

The 2-mix model improved further over the 2-guess model by 237 for 70 extra parameters. The estimated proportion in the "hard items" component was 0.362,

with 95% confidence interval [0.255, 0.526]. The estimates and standard errors for the 2-mix model were very similar to those of the 3PL model, including the accommodation estimate.

The 2-mix-prob model improved in maximised log-likelihood over the 2-mix model by 101 for the 36 extra parameters. The estimated proportion in the "hard items" component (by averaging the estimated membership probabilities over all students) was 0.493. There was one boundary estimate in the membership model for the same large town category. Standard errors of the ability regression estimates generally increased over those for the 2-mix model. For this model, we could estimate a *location shift* in the ability distribution between the two components of 85.8 (SE 42.3) NAEP units – more than two NAEP scale standard errors.

Substantial changes in the ability group regression occurred:

- The disability difference decreased to 12.4 (2.7).
- The sex difference *increased* to 20.1 (1.9).
- The differences between large city schools and schools on the fringe of large cities, −4.5 (2.1), or mid-size cities, −11.8 (3.6), became significant.
- The White–Asian/Pacific Islander difference decreased to 0 (2.5).

(Other smaller decreases occurred as well.) Membership in the "hard items" component was significantly associated with

- disability or limited English proficiency;
- girls;
- Blacks or Hispanics;
- Asian/Pacific Islanders (*negatively*);
- no computer at home;
- absence from school 3 or more days in the month;
- language other than English spoken at home all or most of the time (*negatively*);
- teachers with 5–20 years of experience (*negatively*);
- schools with free or reduced-price lunches.

The very large change in the sex effect, which also appeared in the Texas sample, is hard to believe. As for the Texas sample, we do not feel that the interpretation of this model is reliable and do not comment further.

8.5 Extensive teacher data

Parameter estimates and standard errors for the MIMIC and the four models for guessing, together with the maximised log-likelihoods, are shown in Table D4. We follow the reporting framework for the limited data analysis and give the interpretation of the significant California 3PL estimates, with the corresponding Texas results in square brackets [ ].

8.5.1 3PL interpretation

Student characteristics

- a disability difference of −22.8 (3.0) [−22.5 (4.1)];
- a limited English proficiency difference of −28.0 (1.8) [−20.1 (2.3)];
- an accommodated difference of −35.4 (4.1) [−22.5 (4.2)];
- a sex difference in favour of boys of 7.8 (1.3) [6.4 (1.9)];
- a White–Black difference of 33.2 (2.8) [45.6 (3.5)];
- a White–Hispanic difference of 17.9 (2.0) [25.6 (3.0)];
- a White–American Indian difference of 19.8 (7.5) [12.5 (11.7)];
- a home computer difference of 6.5 (1.8) [6.8 (2.4)];
- absence from school in the month effects of
  - 1–2 days: −8.0 (1.5) [−14.8 (2.1)];
  - 3–4 days: −13.4 (2.1) [−14.8 (3.3)];
  - 5–10 days: −10.0 (2.6) [−14.9 (5.0)];
  - $>$ 10 days: −29.0 (3.0) [−27.0 (6.4)].
- other language at home effects (relative to never) of
  - once in a while: 2.2 (1.8) [5.3 (2.5)];
  - half the time: 9.8 (2.4) [6.6 (3.2)];
  - all or most of the time: 7.9 (1.4) [5.3 (2.7)].
- did not use computer to play games: 7.0 (1.5) [not significant].

School characteristics

- school enrollment (relative to 1–299)
  - size 700+: −16.1 (5.5) [not significant]
- school lunch program
  - reduced price lunch: −8.6 (2.6) [−14.3 (3.8)];
  - free lunch: −18.1 (1.8) [−20.9 (2.6)].
- technical support available: 5.6 (2.2) [9.0 (4.0)] *negative*;
- student numbers in this class (relative to $\leq$ 15)
  - 16–18: 28.7 (10.5) [not significant];
  - 26 or more: 12.9 (6.3) [not significant].

Teacher characteristics

- years of teaching experience (relative to 0–4)
  - 10–19 years: 5.0 (2.1) [5.9 (2.9)];
  - $\geq$ 20 years: 5.2 (2.6) [not significant].
- teacher degree (relative to Bachelor's)

– Master's: 8.7 (1.7) [not significant]

A comparison of Tables D1 and D4 shows that there is close agreement between the 3PL estimates and standard errors for the common variables in the limited and extensive teacher data analyses. Agreement between the estimates for California and Texas is less close than for the limited teacher data, but the structure is again similar.

A notable exception is class size, which gives a 28-point higher mean for students in class sizes 16–18 and a 13-point higher mean for students in classes of 26 or more relative to students in class sizes 15 or less. In the Texas analysis, these differences are smaller and nonsignificant. The class size differences were quite puzzling, as the Texas study showed the opposite effect (in the 2-guess and 2-mix models), of *negative* (and nonsignificant) differences relative to the smallest class size.

8.6 Mixture models

The 2-guess model improved in maximised log-likelihood by 26 over the 3PL for one extra parameter. The estimated proportion in the guessing component was 0.100, with 95% confidence interval [0.075, 0.131]. The reporting group estimates and standard errors for the 2-guess model were again very similar to those of the 3PL model.

We report in Table D6 the item parameter estimates in the guessing and engaged components, and compare them in terms of a "cut" value, as in Table D3 for the limited teacher data set. In Table D6, only three items have positive cut values (the largest is 1.9), while the great majority of items have even *larger* negative values than those in Table D3. The differences for individual items are probably a consequence of the different student sample in the restricted data set, in which large schools are under-represented.

The 2-guess-prob model improved by 133 in maximised log-likelihood over the 2-guess model for the 59 additional parameters in the engagement model. It gave one boundary estimate for the PhD teacher degree category in the membership model where cell sample sizes were small. Standard errors of the ability regression estimates increased over those for the 2-guess model.

Group differences in the ability regression were generally reduced relative to the 3PL model; the reduction was large for American Indian students, the largest-enrolment schools, no home computer, and schools with free lunches. A dramatic difference occurred with the class size effect. This *reversed* relative to the 2-guess model, with significant *negative* differences for class sizes 21–25, 22.0 (7.6), and 26 or more, 15.5 (7.0).

Membership in the guessing component was significantly associated with

- students with a disability or an accommodation, or with limited English proficiency;
- Blacks, Hispanics, and American Indians;
- no computer at home;

- absent 1–2 or 5 or more days in the last month;
- language other than English spoken at home all or most of the time (*negatively*);
- using the computer to play games;
- students in all class sizes larger than 15;
- teachers with a Master's degree (*negatively*);
- teachers with 20 or more years of experience (*negatively*);
- schools with free lunches.

The class size reversal in the ability regression is a consequence of the higher probability of membership in the guessing component of students in *all* classes larger than 15. When these students are taken out of the assessment of ability differences, the class size differences in ability reverse to what one might expect a priori: that students in smaller classes achieve more than students of the same ability in larger classes.

The 2-mix model improved further over the 2-guess model by 173 for 70 extra parameters. The estimated proportion in the "hard items" component was 0.482, with 95% confidence interval [0.451, 0.513]. Despite this improvement in fit, the estimates and standard errors for the 2-mix model were again very similar to those for the 3PL model. In particular, the class size effect for the 16–18 class size remained a significant 20 points above the smallest class size.

The 2-mix-prob model improved in maximised log-likelihood over the 2-mix model by 126 for the 59 extra parameters. The estimated proportion in the "hard items" component (by averaging the estimated membership probabilities over all students) was 0.453.

Standard errors of the ability regression estimates generally increased over those for the 2-mix model. For this model we could (poorly) estimate a location shift in the ability distribution between the two components of 11.7 (14.9) NAEP units. Substantial changes in the ability group regression occurred:

- the accommodation difference decreased to 18.2 (4.2);
- the sex difference was *reversed*, to 6.4 (2.1) in favour of girls;
- the differences between large city schools and schools in mid-size cities, −8.2(3.6), on the fringe of large cities, −7.0 (3.5), on the fringe of mid-size cities, −12.0 (5.0), and in rural (MSA) areas, −31.3 (11.5), became significant;
- the White–Black difference decreased to 21.4 (3.8);
- the White–Hispanic difference decreased to 10.1 (3.1);
- the White–American Indian difference became nonsignificant;
- school lunch differences were greatly decreased;
- math education differences became very large and significant.

Other smaller changes occurred as well. The class size difference in the 2-mix model for size 16–18 *increased* to 26.2 (11.5).

Membership in the "hard items" component was significantly associated with

- limited English proficiency and accommodated students;
- girls;
- Blacks, Hispanics and American Indians;

- absence from school 3–4 days in the month;
- rural schools of both types (*negatively*);
- schools with free or reduced-price lunches.

These results are rather puzzling:

- The counter-intuitive class size difference for class size 16-18 persists.
- The *reversed* sex effect combined with the strong membership of girls in the hard item component produces a Simpson's paradox for girls: in both components, girls were superior to boys by 6.4, but the much smaller proportion in the easy items component gave a marginal sex difference of 8.2 in favour of boys in the 2-guess model.

The conclusions from the 2-mix-prob model are difficult to accept.

We review these results, and the implications for NAEP analysis present and future, in the final chapter.

8.7 Comparison with official NCES analysis

Reporting group results for the Numbers and Operations scale are not published formally. However, they can be computed for this scale (and other scales, as well as the composite score) as one-way tabulations using the NAEP explorer tool on the NCES Website.

Below we give the tabulations of some of these variables using this tool and the equivalent group differences relative to the first category, which can be compared with the results for the 3PL model from the two analyses above in Tables D1 and D2. (The NAEP tool gives mean scores only to integer precision.) As discussed in Chapter 6, only the ethnic group differences are unaffected by the oversampling of high-ethnic-minority schools and minority students.

Comparisons are generally similar to those in the Texas analysis. For the "full" limited teacher data set, the standard error of the White–Black and White–Hispanic differences from the modeling analysis are 50% larger than those from the tool analysis. As with the 1986 survey, there appears to be no design effect correction for the school clustering of students, so the standard errors of the community and school-level variables are probably understated as well in the NCES tool tabulations.

The curious math education reversal found in the Texas comparison is also present here. For the math education level of the teacher, the tool analysis shows an *increase* in mean score with *decreasing* level of math education, while the modeling analysis shows a *decrease* with decreasing math education, though neither analysis gives a statistically significant group difference at the 5% level. As we commented in the Texas analysis, this may be simply an effect of aggregation in the tool analysis, but it highlights the importance of disaggregate analysis.

Comparison of NCES and modeling results - California

	NCES Explorer tool		Limited data		Extensive data	
Male	231 (0.8)					
Female	228 (0.7)	-3 (1.1)	-8.4 (1.2)		-7.8 (1.3)	
Ethnicity						
White	245 (0.9)					
Black	214 (1.5)	-31 (1.7)	-32.7 (2.5)		-33.2 (2.8)	
Hispanic	218 (0.8)	-27 (1.2)	-17.9 (1.9)		-17.9 (2.0)	
Asian/ Pacific Islander	250 (1.6)	5 (1.8)	5.1 (2.1)		3.8 (2.5)	
American Indian	228 (4.6)	-17 (4.7)	-24.3 (5.9)		-27.3 (6.8)	
Limited English proficiency						
Yes	212 (1.0)					
No	238 (0.7)	26 (1.2)	29.5 (1.6)		28.0 (1.8)	
Teacher years of experience						
0-4	227 (1.4)					
5-9	227 (1.0)	0 (1.7)	3.0 (1.7)		3.2 (2.0)	
10-19	234 (1.3)	7 (1.9)	5.8 (1.8)		5.0 (2.1)	
20+	235 (1.7)	8 (2.2)	7.4 (2.1)		5.2 (2.6)	
Math education level						
Major	218 (6.2)					
Minor/ specialism	228 (2.9)	10 (6.8)			-27.2	(14.0)
Neither	230 (0.8)	12 (6.3)			-20.1	(13.8)
Math level						
Major			-			
Minor	227 (3.9)					
Neither	230 (0.7)	3 (4.0)			0.8	(-)
Highest degree						
Bachelor	229 (0.8)					
Master	232 (1.3)	3 (1.5)			8.7	(1.7)
Educ. specialist	237 (3.3)	8 (3.4)			6.9	(3.6)
Prof- essional	235 (9.5)	6 (9.5)			-7.2	(6.9)
Training in computers						
Yes	231 (1.2)					
No	230 (0.8)	-1 (1.4)			-4.2	(1.7)

8.8 Conclusion

The analysis of the California sample shows the same possibilities and difficulties as the Texas sample: detailed analysis with complex models on relatively small samples runs into identifiability difficulties and may give results that seem paradoxical and of doubtful validity.

With the larger California sample, we were able to show that for both the restricted data set with only limited teacher information and the school subsample with extensive teacher information, the 2-guess-prob engagement model gives a richer, and better supported, interpretation of the test results than the 3PL model. The direct and indirect effects of the covariates on achievement, which are separated in this model, clearly show the importance of the model and the educational importance of dealing appropriately with guessing in formal NAEP analyses.

As with the Texas sample, the interpretation of the importance of the extensive teacher covariates is affected to an unknown extent by the (slight) underrepresentation of large schools in the reduced sample.

This emphasises again the importance of obtaining the information about the student and family covariates in the incomplete data that is ignored in the complete-case analysis. We discuss this point at length in Chapter 9.

Chapter 9
Conclusions

In this chapter, we set out the conclusions from our studies

- on the nature and structure of models;
- on our results from modeling;
- on the current analysis of sparse NAEP binary item data;
- on how analyses are reported; and
- on the future analysis and use of NAEP data for both personal academic research and official publications.

9.1 The nature and structure of models

The complexity of the NAEP survey designs and the psychometric models underlying the binary item responses means that *no analysis of the test items is possible without some form of model*. The important issue for NCES and secondary users of their data is *to what extent* models should be incorporated in the analysis and to what extent analysis should be based on the survey design without reference to models.

We have used, and propose generally, a *fully integrated single model* that incorporates both the survey design and the psychometric model by extending the traditional form of the psychometric model to accommodate the design structure while allowing for student, teacher, and school covariates.

Our analyses used the 2PL and MIMIC models and their mixture generalisations. The math items in the 1986 and 2005 surveys did not require *graded response* or *partial credit* models, which we mentioned briefly in Chapters 2 and 3. These models could be extended straightforwardly to multiple survey design levels and could also be extended to mixture models, though with some additional complexities as the various threshold parameters in these models may also vary across mixture components.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,*111
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_9, © Murray Aitkin and Irit Aitkin 2011

9.2 Our modeling results

9.2.1 Comparisons with published NAEP tables

A principal advantage of the fully model-based approach is that it accounts automatically for the clustered structure of the survey design. The jackknifing of PSUs in the 1986 survey did not correct for the school clustering, and it appears that the 2005 survey also did not allow for the school clustering. The failure to do so means that standard errors for estimates at all levels have to be based on a *compromise* variance instead of *separate variances* for the variables at each level, provided by the variance component model.

From the limited comparisons possible of official NCES results with our modeling results, it is clear that the standard errors of the reporting group parameter estimates *above the student level* are *understated considerably* by the current analysis, while those of the reporting group parameter estimates *at the student level* are *overstated slightly*. So one important conclusion from the comparison is the following:

> *The school design effect appears not to be appropriately allowed for* in the current NAEP analysis; the multilevel model approach is a straightforward way of achieving this.

9.2.2 Main effects and interactions

For both formal NCES publications and secondary analysis, the simultaneous fitting of all the reporting group variables has a major benefit – it *disaggregates* the population into reporting group *strata* defined by the cross-classification by these variables, within which the differences on each reporting group variable are *consistent across all the other reporting group variables*; that is, they are *adjusted* for the effects of the other variables.

These constant differences are an *assumption* implied by the model and are open to investigation by fitting additional *interaction* terms – this is a standard and powerful aspect of modeling. In our 2005 survey analyses, we have not used such interaction terms, as the sample sizes in the California and Texas state samples are not large enough to identify the many possible interactions in the already complex models being fitted.

The search for, and interpretation of, multilevel interaction models in large-scale surveys was discussed in detail, with examples, in Aitkin and Zuzovsky (1994).

9.2.3 Mixtures and latent subpopulations

A major contribution of our work is the use of mixture models to represent heterogeneity in students' responses to the items, especially the identification of engagement. The idea of a mixture of responses to the items – expressed as different item parameters in different sub-populations – is described in detail in von Davier and Carstensen (2007) but runs counter to current analysis methods and the philosophy on which they are based.

The assumption of a single population of students over which there are "true" values of the psychometric model item parameters is central to current NAEP analysis. The analyses in this book using a mixture model for guessing, or the more general two-component mixture with different difficulty and discrimination parameters in each component, contradict this assumption. The mixture structure of the 3PL model as it is currently used is different because it assumes that guessing is an *item-based* property, while the guessing mixture model (especially in its component probability modeling form) is a *person-based* property, which in turn allows us to assess "engagement".

An important limitation of these models follows from the relatively small state sample sizes derived from the national NAEP sample. In the Texas sample with extensive teacher data, the 2-mix-prob model was near the limit of identifiability, and this is to be expected with large models (especially at the teacher or school level) relative to the sample size. For a national analysis that could examine *interactions* between student and school or teacher variables, state models would have to be closely related (for example, by a state intercept term in a model otherwise invariant over states).

Including *all* math items in the 1986 analysis helps greatly in increasing the sample size, increasing the information per student, and extending model identification, though allowing for ability multidimensionality again increases the number of model parameters.

A sceptical reader may question this approach, asking:

- Do we really need to worry about item parameter heterogeneity?
- What about differential item functioning (DIFF)?
- Could there be a quite different explanation for apparent heterogeneity, for example the ability underlying the items being multidimensional?
- Could the very large number of parameters be reduced and simplified, for example by a random effect model allowing for individual variations in the item parameters?
- What *evidence* is there for the existence of the *engaged* and *non-engaged* subpopulations of students, or for the more general mixture of two sub-populations that see the items very differently?
- Does the two-component mixture *really* fit better than the 3PL model, and if it does, is this sufficient reason to give up the current standard model?

We deal with these questions in turn.

Heterogeneity and DIFF

A general aspect of our analysis is that item responses are only *indicators* of ability or achievement, and their parameters in the psychometric model are not of primary interest. What matters is the reporting group parameters, which are at a level above the item responses.

So item issues such as DIFF and heterogeneity over subpopulations, which are a cause for concern in current NCES analyses (and even more so in Rasch modeling), are recognised as model choice issues in our approach, and are resolved or investigated by comparing different models.

In the representation of heterogeneity, from the viewpoint of national school educational assessment, the 2-guess model alternative to the 3PL is much easier to accept than the general 2-mix model. The difficulties of *engaging* students in the NAEP task (especially at the later ages) are well documented, with a number of experimental studies of ways to improve participation. Participation is not in itself a guarantee of engagement, and it seems quite reasonable that national comparisons of test scores across reporting group variables should be based on students who were *actually trying* on the test rather than randomly checking boxes. Since guessed items do not give information about a student's ability (under either the 3PL or the 2-guess model), basing the reporting group regression on the component that *is* engaged appears to be a very reasonable way of assessing group differences for those who are engaged in either of these models.

Furthermore, the 2-guess model may be able to provide a valid analysis for part of the population when this is severely heterogeneous because of language or other difficulties of subpopulations with the items. (Of course, this is not a *remedy* for such difficulties.)

In the 3PL and 2-guess models, we cannnot identify students who are guessing on specific items except by examining their item responses for inconsistency – students who have difficulty with easy items should not be able to answer difficult items correctly. In these models, the probability of guessing is *constant* across individuals. In the 2-guess-prob model, this probability varies by item *and* individual characteristics, and it is striking that this extension of the 2-guess model gives very large improvements in maximised log-likelihood in all cases.

Despite these improvements, even the guessing model raises troubling difficulties in interpretation and publication. We have found that membership in the guessing component is strongly associated (in the 1986 *and* the 2005 surveys) with Black, Hispanic, and American Indian ethnicity and with low metropolitan schools. Membership in the engaged component is strongly associated with White ethnicity and high metropolitan schools.

If the reporting group estimates are adjusted for membership in the engaged component, these estimates are changed, sometimes substantially. Typically, the differences from White for the other ethnic groups are substantially reduced. Should both sets of group differences be published, and, if so, how should the difference between them be explained? If not, what should be done with the guessing models?

The more general two-component mixture of logits raises even more difficulties. Maximised log-likelihoods again improve substantially over those for the 2-guess and 2-guess-prob models. Membership in the "difficult items" component for the 2-mix-prob models has a structure very similar to that for the guessing component membership. From the viewpoint of educational psychology and sociology, it is certainly a very serious issue if a large part of the tested population sees the items very differently from the other part. A large location shift in the ability distribution between the two components in the general 2-mix models (or the equivalent large change in item difficulties and discriminations between the components) suggests that the test items require a level of cognitive understanding that is developed differently in parts of the student population.

Multidimensional ability

Our simulation studies showed that, if a real multidimensional structure underlies ability, with different subsets of items identifying these dimensions, then ignoring it in a unidimensional analysis may lead to biases in reporting group estimates. However, it is not necessary to know which items define which dimensions: a general multidimensional covariance structure provides effectively fully efficient estimation.

The multidimensional model is easy to include in the linear predictor for the 2PL model form but complicated in the MIMIC model form, where an ability regression model has to be specified for each dimension; if each dimension has *the same regression model*, then so does the *sum across dimensions* of the separate abilities. In this case, a unidimensional model will give consistent, though possibly inefficient, estimates, and one might question the existence of real multidimensionality.

A separate question, as described below, is whether it is multidimensionality or population heterogeneity that better represents the test item responses.

Random effect models

A random effect model for item parameter heterogeneity extends the mixture concept even further – each individual now has his or her own item parameters, though these are constrained by a distribution over individuals, for example a bivariate normal. We have not used these models in our analyses, though the analyses could be further extended to incorporate them.[1]

The real existence of mixtures

It is obvious that before we try to address these issues, we need to establish beyond statistical doubt that the 3PL model, the guessing models, or the general mixture models (or multidimensional models, or random effect models) *are necessary to represent the test item data*; that is, that the apparent improvements with the mixture or other models are *real*, in the sense that they can be evaluated by statistical theory.

To address these important questions, we have in statistics the major tool of *model comparisons* through the observable data. If we can express the observed item

[1] The mixture model can be thought of as a form of discrete *nonparametric* model for the item parameters that is an alternative to a continuous bivariate model.

responses through different statistical models, then we can express the evidence for these models in terms of their *likelihoods* – the probabilities of the observed data under the models. This is a standard matter when comparing nested models with large samples, when the likelihood ratio test has its usual asymptotic χ^2 distribution. The failure of this distribution to apply, even asymptotically, to non-nested comparisons means that current statistical theory is not adequate to answer the question, as we have repeatedly stated.

At the end of this chapter, we give a discussion of how this difficulty can be resolved.

9.3 Current analysis

The current analysis requires a major computational effort. We give a set of suggestions for changes to the analysis that have the prospect of reducing both the cost of the analysis and its time.

9.3.1 Dependence of design on analysis

A general difficulty with standard errors in the design-based approach is the inability to obtain them from a model information matrix.[2] Sampling variability in the reporting group (and any other) estimates has to be obtained from complex jackknifing, which requires extensive calculations from the pairing of PSUs or schools. This in turn affects the design of the survey in requiring a design in which pairs of PSUs (or schools) can be treated as equivalent.

In the four-level structure of the 1986 survey, this could not be carried out at both the school and PSU levels, and only the PSU design effect was allowed for. This resulted in overstatements of precision – standard errors that were too small – for the upper-level variables. A similar effect appears to be occurring with the 2005 surveys.

A great advantage of the model-based approach is that no such pairing or other constraints on the sample design are necessary: the standard errors are obtained automatically from the information matrix for the variance component model.[3]

[2] As we noted in Chapter 1, the general recommendation to use *weights* – the reciprocals of the sample inclusion probabilities – in model-based analyses with complex survey designs is both *ineffective* and potentially *seriously misleading.*

[3] As we noted in Chapter 4, the model assumption of a normal distribution for the ability or achievement variable is a *very weak* one in the sense that highly skewed or discrete forms for the true ability distribution make very little difference to the maximum likelihood estimates and standard errors of *parameters at the reporting group level*.

9.3.2 Multilevel modeling

We see a major benefit to both NCES and secondary analysts in NCES's moving to a fully multilevel model-based analysis of NAEP surveys. The large conditioning model with hundreds of principal components[4] of more than 1200 covariates, including several hundred school dummy variables, separate estimation of item parameters and regression model coefficients, and the generation of plausible values could all be replaced by direct joint estimation of item parameters and the necessary regression model coefficients.

For the present publications of NCES, the necessary tables could be provided by very small regression models with post-analysis reweighting over the ethnic and other oversampled strata if required. The comparisons across states could be handled by including in the model a state factor that interacts with the reporting group variable(s) if necessary.

For secondary analysis, NCES could provide, or refer to, the appropriate software for full modeling (the current explorer tool could be further developed for this purpose) and might wish to provide a suitable manual for this purpose. The current use of plausible values can provide a reasonably fast though inefficient estimate of group differences, but the process of *generating* them is very complicated and inefficient, both statistically and computationally.

9.3.3 The limitations of NAEP data for large-scale modeling

Our experience with the Texas and California samples from the 2005 survey shows that in separate state analyses the sparsity of the state NAEP data matrices limits the complexity of large-scale regression models relating achievement to (the very large number of) potential explanatory variables on the state data files.

In our analyses, there are no *interactions* among variables in the regression models. We *could* fit such terms, but the number of possible interactions would increase dramatically as the number of variables in the main model increased. We went in a different direction, to an *implicit interaction* of a latent student group variable with the item parameters (and implicit state by variable interactions).

This also requires large numbers of parameters; for tests with large numbers of items, it may be beneficial to relate the item parameters across items through a higher-level model, for example a bivariate normal distribution for the two-parameter models.

For full national NAEP analyses that incorporate state analyses and comparisons, it is clearly necessary to use the full survey data in all participating states; that is, to include all math items in the psychometric model. This may require multidimensional models to accommodate the different math scales, or mixture models if these

[4] Which do not necessarily represent the measured covariates, as only a subset is used.

are better fitting, and a set of state fixed or random effects to allow for overall state differences in achievement, as well as the specific reporting group variables.

Fixed state effects would add of the order on 50 parameters to the model, while random state effects would require a four-level analysis as used for the 1986 survey. This scale of analysis is not beyond the capacity of existing statistical packages (such as Latent Gold, for example), though changes to such packages might be required for efficiency; for example, using random effects for item slopes and intercepts and starting reduced model analyses from the results of the previous model rather than by refitting the reduced models from the beginning each time.

For assessing interactions, especially of state fixed effects with covariates, further developing the capacity of existing programs or sparse data matrix computations might be needed.

9.4 The reporting of NAEP data

In official NCES publications, there are few presentations of NAEP assessment outcomes that go beyond one-way tabulations; for example, by sex and by ethnic group. Many of these presentations would be enriched by multiway tables; for example, of sex differences for each ethnic group. The two-way tabulation might not show the same pattern of sex or ethnic group differences *within* the other variable as shown by the marginal tabulations by sex and by ethnic group. This may at first be confusing to the reader, but it illustrates an important point, that needs explanation. Furthermore, even a common pattern of sex differences across all ethnic groups does not necessarily give the same *marginal difference* between sexes summing over ethnic groups (the Simpson paradox illustrated in Chapter 2).

Multiway tabulations by many variables are complex to present as tables: the regression model framework we have used in the earlier chapters provides an economical way of presenting them that can be extended to include interactions if these are fitted and shown to be necessary.

This illustrates the point made earlier that NCES could not only publish results, but also inform and guide the secondary analysis of their data, by providing not only the software to achieve this, but also a user guide for both statistical modeling principles and the interpretation of multilevel and multiway models. This could also deal with model validation issues for NAEP results that depend *strongly* on model assumptions. An important example is the reporting of percentiles discussed in the research report *Percentile estimation for the ability distribution in item response models* listed in Chapter 4.

The percentiles of a probability distribution depend *strongly* on the tail behaviour of the distribution, much more strongly than means and variances. The fact that the ability distribution is unobserved makes inference about its percentiles much more difficult than for an observed distribution. The assumption of a normal distribution, which is a *very weak one* for estimation of reporting group parameters, is *a very strong one* for estimation of percentiles, and needs to be validated by modeling the

ability distribution using explicit parametic forms that allow for possible skewness and heavy tails, both of which strongly affect percentiles (especially extreme ones).

9.5 The future analysis and use of NAEP data

The administration of the NAEP assessments is a huge task – the spiraling of items into booklets, the adminstration and supervision of the tests themselves, and the test scoring and production of the NAEP data files before their analysis are all very major tasks, quite apart from the analysis on which we have concentrated in this book.

The current developments in *adaptive testing* that are being implemented in some state-wide assessment programs may have an impact on NAEP assessments as well:

> The Secretary of Education has set aside up to $350 million of Race to the Top funds for the potential purpose of supporting States in the development of a next generation of assessments.

This comes from the NCES Website,

`www.ed.gov/programs/racetothetop-assessment`

In adaptive testing, the student uses a local (usually school) computer that is networked to the server of a testing organisation that downloads test items to the student. The items are presented sequentially to the student. If the first item is answered correctly, the program provides successively harder items, continuing until the student answers an item incorrectly. Then the program provides successively easier items, continuing until the student again answers correctly. This process *adapts to the student's ability* and continues until this is sufficiently accurately defined using a server program that performs an analysis on the set of items answered.

With a much smaller set of items than are necessary in a paper-based test, the abilities of all students can be assessed; this happens *simultaneously* for all students in the test session. The assessment session can then provide statistics for the individual students *and the tested group as a whole*, often within 24 hours.

This approach completely avoids the current administrative burden of NAEP assessments, though of course it has an administrative cost of its own in terms of local and server computers for the school assessments and the adaptive testing and administrative reporting software. It would be much more complex to implement such a process for NAEP, but the possibility of greatly simplifying the test administration, especially through omitting completely the production of paper-based booklets and their scoring, would certainly be worth careful investigation.

We return finally to the comparison of item response models and what to do when we have incomplete data on the explanatory variables.

9.6 Resolution of the model-comparison difficulties

We have noted repeatedly the lack of adequate model comparison methods based on the (frequentist) likelihood ratio test. However, the Bayesian paradigm for statistical inference is more flexible in providing general methods for model comparison that do not need models to be nested or that sample sizes be large compared with the model complexities. This approach using simple noninformative priors is developed at book length in Aitkin (2010).

We believe that this approach will be able to provide answers to the model comparison questions that we have raised repeatedly, and we will report the results in due course.

9.7 Resolution of the problems with incomplete data

The Bayesian approach also provides solutions for this problem, for example in the Texas sample, where 25% of the teacher questionnaires were incomplete and the complete-case analysis lost these teachers but also 50% of the students.

One solution, which is based on the data augmentation (DA) algorithm of Tanner and Wong (1987), is to use *multiple imputation* (Rubin 1987) to provide multiple *completed data sets* using an *imputation model*, which provides a number of random draws of the missing variable values from their conditional distribution given the other variables in the data set, and the model parameter estimates. These imputed values of the missing variables are used to *complete* the incomplete data set, and the completed data sets are then analysed in the usual frequentist way and the estimates and their covariance matrix are combined to give a single set of estimates, standard errors, and covariances, which reflect the variations among the sets of estimates.

A more formal Bayesian analysis fully extends the DA analysis to provide the full posterior information about all the model parameters without requiring the multiple analyses of the imputation approach.

Both of these approaches can provide adequate inference about the model parameters without the need for discarding incomplete data in the complete-case approach, giving an important gain in precision. We will also be following this approach and will report the results in due course.

Appendix A
1986 Survey Results, 30 Item Subscale

Table A.1 NAEP items - answers in ()

Report item	NAEP block	NAEP item	
1	M1	4	35 + 42 = (77)
2	M1	5	55 + 37 = (92)
3	M1	6	59 + 46 + 82 + 68 = (255)
4	M1	11	? represents nine tens (90)
5	M1	13	Number 10 more than 95 (105)
6	M1	15	The digit in the thousands place in 45,372 (5)
7	M1	16	Product of 21 and 3 (63)
8	M1	17	Product of 314 and 12 (3768)
9	M2	3	Which is greater: 2573, 2537, or 2735 (2753)
10	M2	6	Which is correct: $7 > 5$, $7 = 5$, or $7 < 5$
11	M2	8	7 + 24 + 9 = (40)
12	M2	9	64 − 27 = (37)
13	M2	10	604 − 207 = (397)
14	M2	11	231 − 189 = (42)
15	M2	12	Number of birds in picture [< 100, (100-1000), > 1000, $> 10{,}000$]
16	M2	21	15 / 5 = (3)
17	M2	22	52 / 4 = (13)
18	M2	23	29 − (13) = 16
19	M3	15	One dollar and 86 cents means: [$.186, ($1.86), $10.86, $18.60, $186.00]
20	M3	17	The digit in the tens place in 3058 (5)
21	M4	8	39 −26 (13)
22	M4	9	79 − 45 (34)
23	M4	10	65 − 7 (58)
24	M4	11	If 10 in each bag, 150 marbles in [10, (15), 25, 140, 150, 160] bags
25	M4	17	Three-fourths is (3/4)
26	M4	20	If N*13 = 13, N = (1)
27	M4	22	4.32 is [forty-three and 2/10, four hundred 32, (four and 32/100), forty-three hundred]
28	M6	21	152 − 59 − 93 = [four possibilities]
29	M7	15	Which picture shows 3/4 shaded [four possibilities]
30	M7	23	82 − 39 is closest to [80 − 30, (80 − 40), 90 − 30, 90 − 40]

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5,

Table A2

Rasch variance component estimates and SEs for the null models

	2-level	3-level	4-level
s^2_PSU	0	0	0.059(.016)
s^2_sch	0	0.325(.029)	0.257(.024)
s^2	1.558(.043)	1.246(.038)	1.243(.038)
log L	-40977.29	-40605.49	-40601.08

2PL variance component estimates and SEs for the null models

	2-level	3-level	4-level
s^2_PSU	0	0	0.020(.010)
s^2_sch	0	0.171(.020)	0.130(.015)
s^2	1.0	1.0	1.0
log L	-40559.10	-40183.23	-40180.08

Table A3

Rasch, MIMIC, and 2PL estimates and (SE)s for the two-level model

	Rasch	MIMIC	2PL
male	0	0	0
femal	3.9 (1.1)	3.9 (1.1)	0.5 (1.0)
white	0	0	0
black	-29.0 (1.6)	-25.1 (2.9)	-27.7 (1.4)
hispa	-19.1 (1.6)	-16.9 (2.2)	-18.0 (1.4)
as/pa	-7.5 (4.4)	-6.0 (3.8)	-8.4 (4.0)
amind	-21.0 (4.0)	-19.9 (3.9)	-20.6 (3.2)
other	-4.2(27.4)	7.4(21.6)	-13.3(29.9)
NE	0	0	0
SE	-1.2 (1.8)	-1.4 (1.6)	0.1 (1.6)
Cent	-8.6 (1.8)	-7.4 (1.8)	-7.7 (1.6)
West	-7.3 (1.6)	-6.8 (1.6)	-6.8 (1.5)
extru	0	0	0
lomet	-7.8 (2.9)	-6.7 (2.7)	-4.4 (2.6)
himet	17.8 (2.8)	16.2 (3.0)	16.6 (2.6)
manct	4.9 (2.8)	5.6 (2.5)	4.7 (2.5)
urbfr	4.7 (2.8)	5.6 (2.6)	6.1 (2.6)
medct	2.8 (2.6)	2.5 (2.3)	1.1 (2.3)
smplc	-2.3 (2.5)	-1.2 (2.3)	-1.9 (2.3)
nfnhs	0	0	0
finhs	5.8 (4.9)	6.2 (4.4)	1.6 (4.8)
smcol	20.8 (5.4)	19.1 (5.1)	16.1 (5.2)
colgr	20.7 (5.0)	19.1 (4.8)	15.9 (4.8)
DK	4.2 (5.1)	4.7 (4.5)	2.1 (4.9)
nores	-3.7 (5.6)	-2.8 (5.0)	-7.2 (5.3)
s^2	1.311 (.039)	1.0	1.0
log L	-40475.22	-40080.32	-40077.26
#params	51	80	80

Table A4

Rasch, MIMIC, 2PL and 3PL estimates and (SE)s
for the three-level model

	Rasch	MIMIC	2PL	3PL
male	0	0	0	0
femal	3.7 (1.1)	4.2 (1.1)	0.4 (1.0)	0.8 (1.2)
white	0	0	0	0
black	-24.6 (1.8)	-22.1 (2.8)	-23.3 (1.6)	-29.3 (2.1)
hispa	-17.2 (1.6)	-15.8 (2.2)	-16.1 (1.5)	-21.3 (1.9)
as/pa	-6.9 (4.5)	-5.7 (4.3)	-7.1 (4.1)	-7.8 (4.9)
amind	-17.1 (3.7)	-16.5 (4.0)	-16.5 (3.3)	-21.1 (4.2)
other	-1.9(25.1)	5.7(24.3)	-7.0(24.3)	-1.2(28.7)
NE	0	0	0	0
SE	-2.0 (2.8)	-2.6 (2.7)	-2.7 (2.7)	-2.6 (3.3)
Cent	-7.1 (2.6)	-6.6 (2.4)	-6.0 (2.6)	-7.4 (3.2)
West	-7.1 (2.6)	-6.6 (2.4)	-6.4 (2.4)	-7.6 (3.0)
extru	0	0	0	0
lomet	-10.9 (4.4)	-10.8 (4.3)	-7.0 (4.0)	-12.2 (4.8)
himet	17.9 (4.4)	15.2 (4.5)	17.4 (4.1)	20.5 (5.1)
manct	5.0 (4.2)	4.4 (4.1)	5.3 (3.7)	5.7 (4.7)
urbfr	6.6 (4.4)	5.8 (4.2)	5.5 (3.9)	6.0 (5.3)
medct	4.0 (3.9)	3.3 (3.9)	3.2 (3.4)	4.0 (4.3)
smplc	-1.3 (3.8)	-1.5 (3.7)	-0.7 (3.3)	1.0 (4.1)
nfnhs	0	0	0	0
finhs	7.7 (7.2)	7.1 (7.1)	0.9 (6.9)	2.0 (9.3)
smcol	20.6 (7.2)	18.8 (7.4)	13.4 (6.9)	17.1 (9.3)
colgr	21.2 (7.5)	19.3 (7.6)	13.9 (7.2)	18.3 (9.6)
DK	6.5 (7.3)	5.7 (7.2)	1.6 (7.0)	2.0 (9.3)
nores	-1.2 (7.6)	-0.7 (7.4)	-6.3 (7.2)	-6.5 (9.6)
s^2_sch	0.154 (.019)	0.126 (.030)	0.139 (.017)	0.233 (.028)
s^2	1.168 (.037)	1.0	1.0	1.0
log L	-40349.04	-39961.68	-39930.05	-39848.49
#params	52	81	81	111

Table A5

Guessing parameters for 3PL models, NAEP and multilevel

item	NAEP	S.E.	multilevel	S.E.
1	0	0	0.462	0.203
2	0	0	0.045	-
3	0	0	0.043	-
4	0.238	0.015	0	0
5	0	0	0	0
6	0.208	0.013	0.036	-
7	0	0	0.053	-
8	0	0	0	0
9	0.280	0.014	0.363	0.034
10	0.352	0.015	0.098	0.034
11	0	0	0.188	0.031
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0
15	0.225	0.013	0.185	0.030
16	0	0	0	0
17	0	0	0	0
18	0	0	0	0
19	0.198	0.020	0	0
20	0.257	0.018	0.061	-
21	0	0	0	0
22	0	0	0	0
23	0	0	0	0
24	0.232	0.012	0.013	-
25	0	0	0	0
26	0	0	0	0
27	0.197	0.006	0.108	0.014
28	0.164	0.006	0.127	0.018
29	0.247	0.013	0.257	0.040
30	0.243	0.013	0.266	0.032

Table A6

Parameter correlation matrix for three-level Rasch model

	sex2	race2	race3	race4	race5	race6	regi2	regi3	regi4
race2	-.010	1.0							
race3	.029	.259	1.0						
race4	.017	.100	.113	1.0					
race5	.020	.119	.126	.042	1.0				
race6	.012	.013	.012	.009	.005	1.0			
regi2	-.003	-.080	-.013	-.009	-.001	-.001	1.0		
regi3	-.002	.001	.014	-.000	.009	-.008	.497	1.0	
regi4	.009	.023	-.069	-.026	-.032	-.007	.543	.500	1.0
stoc2	.000	-.139	-.108	-.034	-.017	-.005	.132	.219	.060
stoc3	.003	-.088	-.021	-.008	.014	-.009	.098	.243	.013
stoc4	.013	-.080	-.053	-.027	.008	-.007	.064	.175	-.007
stoc5	.011	-.040	-.008	-.006	.023	-.004	.100	.241	.009
stoc6	.005	-.072	-.038	-.002	.015	-.007	-.049	.200	.013
stoc7	-.002	-.033	-.017	-.003	.022	-.006	.018	.245	.044
pare2	-.005	.034	.004	.032	.006	.059	.003	-.039	.008
pare3	-.000	.027	-.001	.034	-.003	.058	.013	-.047	.008
pare4	.003	.025	.003	.027	-.001	.060	.009	-.042	.006
pare5	.001	.026	.005	.032	-.004	.062	.016	-.043	.011
pare6	-.008	.032	.005	.034	-.001	.061	.010	-.045	.003

	stoc2	stoc3	stoc4	stoc5	stoc6	stoc7	pare2	pare3	pare4	pare5
stoc2	1.0									
stoc3	.680	1.0								
stoc4	.677	.665	1.0							
stoc5	.646	.687	.624	1.0						
stoc6	.658	.701	.650	.678	1.0					
stoc7	.713	.730	.701	.707	.783	1.0				
pare2	-.017	-.022	-.030	-.034	-.059	-.045	1.0			
pare3	-.005	-.020	-.029	-.034	-.061	-.044	.928	1.0		
pare4	-.012	-.021	-.027	-.037	-.058	-.041	.904	.935	1.0	
pare5	-.010	-.038	-.038	-.045	-.070	-.050	.937	.972	.945	1.0
pare6	-.015	-.031	-.040	-.043	-.066	-.049	.941	.975	.948	.985

Table A7

Rasch, MIMIC, and 2PL estimates and SEs for the four-level model

	Rasch	MIMIC	2PL
male	0	0	0
femal	3.7 (1.1)	4.2 (1.1)	0.4 (1.0)
white	0	0	0
black	-24.7 (2.0)	-22.2 (2.8)	-23.5 (1.6)
hispa	-17.2 (1.6)	-15.6 (2.2)	-15.9 (1.5)
as/pa	-7.0 (4.5)	-6.0 (4.3)	-7.2 (4.1)
amind	-17.3 (3.7)	-16.7 (3.9)	-16.5 (3.3)
other	-1.5(25.1)	6.7(24.3)	-6.1(25.9)
NE	0	0	0
SE	-1.8 (3.3)	-2.4 (3.0)	-1.3 (2.8)
Cent	-7.1 (3.4)	-6.1 (3.2)	-6.2 (3.0)
West	-6.4 (3.1)	-6.2 (3.1)	-6.9 (2.7)
extru	0	0	0
lomet	-10.5 (4.4)	-11.2 (4.1)	-6.6 (4.2)
himet	16.5 (4.4)	13.8 (4.1)	16.8 (4.2)
manct	4.8 (4.3)	4.5 (3.8)	6.6 (4.2)
urbfr	6.3 (4.4)	5.6 (4.0)	6.2 (4.3)
medct	2.4 (3.6)	2.3 (3.6)	2.1 (3.8)
smplc	-1.8 (3.7)	-2.2 (3.4)	-0.9 (3.6)
nfnhs	0	0	0
finhs	9.5 (7.1)	8.3 (6.9)	1.7 (6.7)
smcol	22.3 (7.1)	20.0 (7.1)	14.1 (6.7)
colgr	22.9 (7.4)	20.5 (7.4)	14.8 (7.0)
DK	8.2 (7.2)	6.9 (6.9)	2.3 (6.8)
nores	0.6 (7.5)	0.6 (7.1)	-5.4 (7.0)
s^2_PSU	.036 (.013)	.030 (.013)	.019 (.010)
s^2_sch	.116 (.019)	.100 (.030)	.119 (.017)
s^2	1.169 (.037)	1.0	1.0
log L	-40342.85	-39954.46	-39924.60
#params	53	82	82

Table A8

Reporting group estimates and SEs -
3PL, 2PL, guessing, and two-dimension models

	3PL	2-g	2-g-p	engaged membership	2-D 2PL
intercept				-1.09 (.54)	
male	0	0	0	0	0
femal	0.8 (1.2)	2.8 (1.0)	0.8 (1.0)	0.09 (.10)	1.2 (1.1)
white	0				
black	-29.3 (2.1)	-23.6 (1.6)	-21.6 (1.7)	-0.86 (.17)	-24.2 (1.7)
hispa	-21.3 (1.9)	-15.9 (1.5)	-14.5 (1.5)	-0.59 (.15)	-16.5 (1.6)
as/pa	-7.8 (4.9)	-4.3 (4.1)	-2.4 (4.2)	-0.62 (.41)	-6.9 (4.4)
amind	-21.1 (4.2)	-18.2 (3.3)	-16.8 (3.3)	-0.70 (.39)	-17.1 (3.5)
other	-1.2(28.7)	-4.7(22.2)	0.5(20.9)	-0.27 (2.4)	-10.0(28.3)
NE	0	0	0	0	0
SE	-2.6 (3.3)	1.9 (2.7)	4.1 (2.8)	-0.35 (.16)	-0.9 (2.7)
Cent	-7.4 (3.2)	-5.1 (2.9)	-5.8 (2.8)	0.10 (.16)	-6.7 (2.7)
West	-7.6 (3.0)	-3.4 (2.7)	-6.3 (2.8)	-0.10 (.14)	-6.9 (2.5)
extru	0	0	0	0	0
lomet	-12.2 (4.8)	-8.1 (4.6)	-6.7 (4.1)	-0.54 (.30)	-7.7 (4.1)
himet	20.5 (5.1)	15.0 (4.3)	13.4 (4.1)	0.81 (.25)	17.5 (4.1)
manct	5.7 (4.7)	5.4 (4.0)	5.3 (3.5)	0.11 (.25)	5.0 (3.9)
urbfr	6.0 (5.3)	5.9 (4.5)	4.7 (4.0)	0.57 (.25)	5.5 (4.0)
medct	4.0 (4.3)	2.7 (3.7)	2.6 (3.4)	0.37 (.23)	2.7 (3.6)
smplc	1.0 (4.1)	-0.2 (3.7)	-0.6 (3.3)	0.24 (.23)	-1.2 (3.5)
nfnhs	0	0	0	0	0
finhs	2.0 (9.3)	0.7 (6.1)	-0.3 (6.2)	-0.07 (.51)	1.9 (6.8)
smcol	17.1 (9.3)	13.2 (6.4)	9.1 (6.4)	0.50 (.53)	14.7 (7.1)
colgr	18.3 (9.6)	13.9 (6.1)	10.4 (6.1)	0.35 (.50)	14.1 (6.8)
DK	2.0 (9.3)	2.4 (6.2)	-2.6 (6.0)	0.13 (.50)	1.2 (6.7)
nores	-6.5 (9.6)	-6.2 (6.4)	-8.2 (6.4)	-0.32 (.57)	-6.5 (7.1)
s^2_sch	.233 (.028)	.143 (.016)	.132 (.015)		.137 (.018)
s^2	1.0	1.0	1.0		1.0
log Lmax	-39848.49	-39777.88	-39711.77		-39685.69
#params	111	112	132		111

Table A9

Guessing parameters

	3PL		2-g		2-g-p
item	c_j	SE	d_j	c.d_j	d_j
1	0.462	0.203	1.000	0.219	0.960
2	0.045	-	0.889	0.195	0.876
3	0.043	-	0.428	0.183	0.445
4	0	0	0.677	0.148	0.800
5	0	0	0.682	0.149	0.757
6	0.036	-	0.524	0.115	0.569
7	0.053	-	0.387	0.085	0.509
8	0	0	0.009	0.002	0.039
9	0.363	0.034	0.666	0.146	0.706
10	0.098	0.034	0.599	0.131	0.637
11	0.188	0.031	0.559	0.122	0.578
12	0	0	1.000	0.219	0.985
13	0	0	0.760	0.166	0.762
14	0	0	0.824	0.180	0.794
15	0.185	0.030	0.449	0.098	0.488
16	0	0	0.646	0.141	0.688
17	0	0	0.068	0.015	0.110
18	0	0	0.300	0.066	0.364
19	0	0	0.834	0.183	0.895
20	0.061	-	0.684	0.150	0.757
21	0	0	0.719	0.157	0.817
22	0	0	0.719	0.157	0.818
23	0	0	0.648	0.142	0.735
24	0.013	-	0.882	0.193	0.915
25	0	0	0.405	0.089	0.469
26	0	0	0.472	0.103	0.529
27	0.108	0.014	0.102	0.022	0.123
28	0.127	0.018	0.214	0.047	0.231
29	0.257	0.040	0.614	0.134	0.572
30	0.266	0.032	0.520	0.114	0.544

Table A10

Reporting group estimates and SEs
- mixed 2PL models

	2-m	2-m-p	easy items membership
male	0	0	0
femal	1.2 (1.0)	0.3 (1.2)	0.17 (.09)
white	0	0	0
black	-23.6 (1.7)	-16.0 (2.0)	-0.94 (.12)
hispa	-15.8 (1.5)	-11.0 (1.8)	-0.61 (.12)
as/pa	-5.2 (4.3)	-2.6 (4.9)	-0.30 (.31)
amind	-16.8 (3.4)	-10.7 (3.9)	-0.72 (.27)
other	-9.0(26.7)	-6.5(32.8)	-0.02 (1.8)
NE	0	0	0
SE	1.0 (2.6)	4.1 (2.8)	-0.40 (.14)
Cent	-6.2 (2.7)	-5.8 (2.8)	-0.05 (.15)
West	-4.7 (2.4)	-2.7 (2.6)	-0.30 (.13)
extru	0	0	0
lomet	-7.9 (3.9)	-3.9 (4.2)	-0.49 (.22)
himet	12.6 (4.3)	7.8 (4.4)	0.74 (.21)
manct	4.2 (3.7)	3.2 (4.0)	0.13 (.20)
urbfr	3.8 (3.9)	-0.6 (4.2)	0.55 (.22)
medct	0.4 (3.5)	-1.4 (3.8)	0.29 (.19)
smplc	-2.7 (3.4)	-3.5 (3.7)	0.13 (.19)
nfnhs	0	0	0
finhs	-1.1 (6.4)	-3.3 (7.1)	0.27 (.39)
smcol	12.1 (6.7)	5.7 (7.4)	0.76 (.41)
colgr	11.0 (6.3)	6.3 (7.0)	0.64 (.38)
DK	-2.1 (6.3)	-6.2 (7.0)	0.54 (.38)
nores	-8.5 (6.7)	-11.0 (7.5)	0.20 (.43)
s^2_sch	.130 (.016)	.121 (.015)	
s^2	1.0	1.0	
log Lmax	-39649.40	-39547.41	
#params	142	172	

Table A11

Intercept and slope parameters for the 2-mix 2PL model

item	comp1 int (SE)	comp2 int (SE)	comp1 slope (SE)	comp2 slope (SE)
1	2.74 (0.30)	4.56 (0.54)	1.13 (0.21)	-0.67 (0.47)
2	0.82 (0.24)	2.75 (0.26)	0.96 (0.16)	0.10 (0.20)
3	-1.61 (0.26)	0.64 (0.23)	0.82 (0.23)	0.41 (0.13)
4	-0.11 (0.23)	0.63 (0.23)	1.05 (0.17)	0.97 (0.15)
5	-0.33 (0.23)	0.44 (0.22)	1.03 (0.18)	0.66 (0.13)
6	-0.98 (0.23)	0.04 (0.23)	0.57 (0.16)	1.07 (0.16)
7	-0.82 (0.24)	0.85 (0.23)	0.59 (0.17)	0.75 (0.16)
8	-4.28 (0.52)	-2.55 (0.31)	0.21 (0.60)	1.08 (0.25)
9	0.47 (0.22)	1.27 (0.22)	0.86 (0.13)	0.89 (0.13)
10	-0.19 (0.22)	0.90 (0.22)	0.86 (0.13)	0.65 (0.11)
11	-0.43 (0.22)	0.72 (0.22)	1.02 (0.15)	0.77 (0.11)
12	-1.17 (0.25)	2.71 (0.25)	0.64 (0.16)	0.63 (0.22)
13	-7.75 (2.65)	0.75 (0.22)	2.90 (1.25)	0.76 (0.14)
14	-3.19 (0.38)	1.01 (0.22)	0.83 (0.35)	0.78 (0.14)
15	-0.30 (0.22)	0.14 (0.21)	0.77 (0.12)	0.46 (0.09)
16	0.26 (0.23)	1.39 (0.24)	1.46 (0.22)	1.14 (0.16)
17	-3.49 (0.44)	-1.78 (0.25)	2.10 (0.36)	1.26 (0.18)
18	-1.38 (0.24)	-0.49 (0.23)	1.05 (0.17)	1.09 (0.15)
19	0.11 (0.26)	1.69 (0.25)	1.32 (0.35)	0.77 (0.23)
20	-0.32 (0.27)	1.27 (0.25)	0.90 (0.26)	0.91 (0.23)
21	2.67 (0.38)	2.45 (0.29)	3.53 (0.41)	1.92 (0.28)
22	2.93 (0.45)	2.43 (0.28)	4.40 (0.90)	1.95 (0.24)
23	-0.40 (0.24)	1.14 (0.23)	1.69 (0.33)	1.14 (0.18)
24	0.24 (0.23)	1.95 (0.28)	0.75 (0.15)	1.11 (0.21)
25	1.17 (0.27)	-0.06 (0.22)	0.98 (0.26)	0.69 (0.14)
26	-1.23 (0.27)	0.16 (0.23)	1.01 (0.29)	0.84 (0.16)
27	-1.71 (0.26)	-1.59 (0.24)	-0.27 (0.16)	0.39 (0.16)
28	-1.16 (0.28)	-0.94 (0.25)	-0.12 (0.23)	0.70 (0.21)
29	0.14 (0.24)	0.82 (0.23)	0.35 (0.17)	0.56 (0.15)
30	-0.35 (0.25)	0.59 (0.24)	0.41 (0.21)	0.82 (0.19)

Table A12

Intercept and slope parameters, 2-mix-prob 2PL model

item	comp1 int (SE)	comp2 int (SE)	comp1 slope (SE)	comp2 slope (SE)
1	3.25 (0.35)	3.93 (0.36)	1.45 (0.24)	-0.19 (0.37)
2	1.33 (0.26)	2.43 (0.26)	1.48 (0.19)	0.36 (0.20)
3	-0.96 (0.28)	0.28 (0.24)	1.66 (0.25)	0.63 (0.14)
4	-0.61 (0.24)	1.25 (0.25)	0.41 (0.12)	0.81 (0.18)
5	-0.78 (0.25)	0.92 (0.24)	0.60 (0.13)	0.38 (0.14)
6	-1.16 (0.25)	0.35 (0.24)	0.32 (0.15)	0.99 (0.17)
7	-0.19 (0.25)	0.48 (0.24)	1.23 (0.18)	1.23 (0.18)
8	-3.67 (0.44)	-3.29 (0.44)	1.37 (0.34)	1.67 (0.34)
9	0.49 (0.23)	1.38 (0.24)	0.83 (0.12)	0.86 (0.13)
10	-0.11 (0.23)	1.00 (0.23)	0.81 (0.12)	0.68 (0.11)
11	-0.32 (0.23)	0.79 (0.23)	1.04 (0.15)	0.80 (0.12)
12	-0.94 (0.25)	2.88 (0.26)	0.68 (0.16)	0.51 (0.26)
13	-9.01 (7.07)	0.95 (0.24)	3.19 (2.66)	0.69 (0.15)
14	-3.17 (0.48)	1.17 (0.24)	1.06 (0.46)	0.63 (0.14)
15	-0.29 (0.23)	0.25 (0.23)	0.77 (0.12)	0.40 (0.09)
16	0.31 (0.24)	1.50 (0.25)	1.42 (0.22)	1.13 (0.17)
17	-3.11 (0.40)	-1.62 (0.26)	1.67 (0.33)	1.27 (0.19)
18	-1.37 (0.25)	-0.31 (0.24)	0.93 (0.17)	1.08 (0.15)
19	-0.02 (0.28)	1.92 (0.27)	1.55 (0.40)	0.61 (0.22)
20	-0.22 (0.26)	1.37 (0.27)	0.89 (0.22)	0.97 (0.24)
21	2.46 (0.33)	2.62 (0.31)	3.45 (0.38)	1.91 (0.25)
22	2.42 (0.33)	2.60 (0.31)	3.71 (0.51)	1.89 (0.23)
23	-0.36 (0.25)	1.26 (0.25)	1.79 (0.29)	1.03 (0.17)
24	0.20 (0.24)	2.28 (0.30)	0.71 (0.12)	1.05 (0.23)
25	-1.22 (0.27)	0.12 (0.24)	0.97 (0.20)	0.62 (0.12)
26	-1.20 (0.26)	0.32 (0.25)	0.94 (0.21)	0.76 (0.14)
27	-1.60 (0.26)	-1.57 (0.26)	-0.28 (0.14)	0.53 (0.18)
28	-1.15 (0.27)	-0.81 (0.26)	-0.12 (0.20)	0.68 (0.19)
29	0.28 (0.24)	0.82 (0.24)	0.28 (0.14)	0.59 (0.14)
30	-0.21 (0.25)	0.62 (0.24)	0.48 (0.17)	0.82 (0.18)

Appendix B
1986 Survey Results, Full 79 Items

Table B0

Reporting group estimates and SEs - two-parameter 2PL models, 79 items (3- and 4-level models)

	3-level	4-level
male	0	0
femal	-1.8 (0.7)	-1.8 (0.7)
white	0	0
black	-18.7 (1.1)	-18.5 (1.1)
hispa	-13.8 (1.0)	-13.8 (1.0)
as/pa	-9.4 (2.9)	-9.4 (2.9)
amind	-16.0 (2.2)	-16.0 (2.1)
other	-15.6(14.1)	-14.7(13.8)
NE	0	0
SE	-0.6 (1.8)	-0.1 (2.1)
Cent	-2.1 (1.8)	-1.9 (2.2)
West	-4.2 (1.6)	-3.8 (2.0)
extru	0	0
lomet	-4.5 (2.5)	-4.9 (2.8)
himet	12.6 (3.1)	13.3 (2.9)
manct	6.0 (2.3)	4.3 (2.7)
urbfr	7.5 (2.5)	7.3 (2.7)
medct	3.9 (2.1)	2.8 (2.5)
smplc	1.5 (2.2)	0.9 (2.3)
nfnhs	0	0
finhs	4.8 (1.8)	4.7 (1.8)
smcol	14.1 (2.1)	14.2 (2.1)
colgr	15.0 (1.6)	14.9 (1.6)
DK	5.4 (1.6)	5.4 (1.6)
nores	2.1 (4.3)	0.7 (4.4)
s^2_PSU	-	.013 (.006)
s^2_sch	.061 (.006)	.054 (.007)
s^2	1.0	1.0
log Lmax	-103013.51	-103007.14
#params	179	180

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 133
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5, © Murray Aitkin and Irit Aitkin 2011

Table B1

Reporting group estimates and SEs - 2PL models, 79 items
(three-level models)

	2PL	2-guess	2-mix	3-guess	3-mix	3-D
male	0	0	0	0	0	0
femal	-1.8(0.7)	-0.6(0.7)	-1.5(0.6)	-0.1(0.6)	-1.9(0.7)	-1.8(0.7)
white	0	0	0	0	0	0
black	-18.7(1.1)	-17.2(1.0)	-16.8(1.0)	-16.6(1.0)	-17.3(1.0)	-19.9(1.1)
hispa	-13.8(1.0)	-14.0(0.9)	-13.6(1.0)	-12.3(1.0)	-13.9(1.0)	-14.8(1.1)
as/pa	-9.4(2.9)	-8.1(2.7)	-7.6(2.9)	-7.7(2.9)	-7.8(2.8)	-9.6(3.0)
amind	-16.0(2.2)	-17.0(2.2)	-16.5(2.2)	-14.6(2.1)	-16.4(2.5)	-17.5(2.3)
other	-15.6 (14)	-17.2 (14)	-19.0 (15)	-25.7 (13)	-23.9 (13)	-25.4 (21)
NE	0	0	0	0	0	0
SE	-0.6(1.8)	-1.2(1.9)	-0.0(1.5)	-2.1(1.7)	-1.0(1.5)	-1.3(1.8)
Cent	-2.1(1.8)	-2.7(1.7)	-3.2(1.5)	-3.3(1.6)	-3.9(1.6)	-1.5(1.6)
West	-4.2(1.6)	-4.0(1.9)	-2.4(1.5)	-5.5(1.8)	-5.1(1.5)	-3.8(1.9)
extru	0	0	0	0	0	0
lomet	-4.5(2.5)	-4.0(2.3)	-7.8(2.3)	-5.9(2.1)	-6.4(2.3)	-7.1(2.6)
himet	12.6(3.1)	10.4(2.6)	10.1(2.3)	11.0(3.0)	12.9(2.4)	13.6(3.3)
manct	6.0(2.3)	4.5(2.1)	0.2(2.2)	0.9(2.3)	2.6(2.3)	3.6(2.6)
urbfr	7.5(2.5)	2.9(2.6)	3.4(2.6)	3.6(2.4)	3.0(2.8)	3.7(2.8)
medct	3.9(2.1)	3.8(2.1)	-2.0(2.3)	0.5(2.7)	-1.2(2.4)	2.6(2.5)
smplc	1.5(2.2)	1.7(2.1)	-2.1(2.1)	1.7(1.9)	0.0(2.0)	-0.2(2.4)
nfnhs	0	0	0	0	0	0
finhs	4.8(1.8)	6.8(1.7)	9.2(1.7)	6.0(1.7)	6.0(1.8)	7.2(1.9)
smcol	14.1(2.1)	14.2(1.9)	14.5(2.0)	12.3(2.0)	14.8(2.0)	15.6(2.2)
colgr	15.0(1.6)	15.7(1.6)	17.2(1.6)	13.7(1.6)	15.4(1.6)	16.4(1.8)
DK	5.4(1.6)	5.7(1.5)	7.6(1.5)	4.8(1.5)	6.9(1.6)	7.0(1.7)
nores	2.1(4.3)	0.6(3.5)	4.0(3.9)	1.2(3.6)	0.2(4.1)	5.6(4.2)
s^2_sch	.061(.01)	.059(.01)	.057(.01)	.061(.01)	.062(.01)	.057(.01)
s^2	1.0	1.0	1.0	1.0	1.0	1.0
logLmax	-103013.5	-102682.8	-102186.3	-102380.0	-101817.5	-102014.6
#params	179	258	338	418	497	337

Table B2

Reporting group estimates and SEs - 2PL and mixed models, 79 items
(three-level models)

	3PL	2-guess-prob	engaged membership	2-mix-prob	easy items membership
1			-1.12 (.23)		-0.05 (.22)
male	0	0	0	0	0
femal	0.2 (0.6)	-1.0 (0.7)	-0.12 (.07)	-1.3 (0.8)	-0.09 (.077)
white	0	0	0	0	0
black	-17.3 (0.9)	-6.6 (1.1)	-1.72 (.12)	-5.4 (1.1)	-1.65 (.10)
hispa	-13.8 (0.9)	-7.7 (1.1)	-0.91 (.10)	-5.7 (1.1)	-1.06 (.09)
as/pa	-6.0 (2.4)	-1.7 (2.8)	-0.56 (.25)	-3.2 (3.0)	-0.71 (.24)
amind	-13.7 (2.0)	-7.7 (2.4)	-0.95 (.24)	-7.5 (2.7)	-0.99 (.23)
other	-10.4 (9.3)	-12.6(12.6)	-0.24 (1.1)	-23.5(13.7)	0.01 (1.3)
NE	0	0	0	0	0
SE	-1.8 (1.5)	-2.7 (1.5)	0.05 (.11)	-0.6 (1.8)	-0.16 (.11)
Cent	-0.2 (1.5)	-5.5 (1.5)	0.12 (.11)	-4.5 (1.7)	0.05 (.11)
West	-5.4 (1.4)	-3.2 (1.4)	-0.13 (.10)	-5.4 (1.5)	-0.27 (.10)
extru	0	0	0	0	0
lomet	-7.9 (2.2)	-7.2 (2.2)	-0.41 (.18)	-2.3 (2.6)	-0.44 (.17)
himet	11.2 (1.8)	5.7 (2.2)	0.79 (.15)	2.3 (2.8)	1.13 (.17)
manct	2.2 (1.8)	1.9 (2.3)	0.12 (.15)	3.5 (3.2)	0.16 (.15)
urbfr	4.4 (2.1)	2.0 (2.4)	0.20 (.16)	-0.5 (2.7)	0.50 (.16)
medct	3.4 (1.7)	-0.5 (2.1)	0.29 (.14)	-0.3 (2.5)	0.26 (.14)
smplc	0.1 (1.6)	-1.0 (2.0)	-0.01 (.13)	0.9 (2.4)	0.11 (.14)
nfnhs	0	0	0	0	0
finhs	4.6 (1.5)	3.7 (1.9)	0.45 (.20)	5.5 (2.0)	0.28 (.18)
smcol	13.3 (1.8)	5.3 (2.2)	1.26 (.22)	8.5 (2.3)	0.94 (.21)
colgr	14.0 (1.4)	8.1 (1.8)	1.06 (.19)	8.9 (1.9)	1.07 (.17)
DK	5.6 (1.3)	3.5 (1.7)	0.51 (.19)	4.9 (1.8)	0.41 (.16)
nores	-1.7 (3.6)	0.1 (5.8)	0.71 (.38)	1.8 (3.9)	0.17 (.31)
s^2_sch	.063 (.006)	.057 (.005)		.050 (.005)	
s^2	1.0	1.0		1.0	
log Lmax	-102404.15	-102396.61		-101793.69	
#params	258	279		358	

Table B3

Reporting group estimates and SEs - 2PL and mixed models, 79 items
(three-level models)

	3-g-p	diff. items membership	easi. items membership	3-m-p	diff. items membership	easi. items membership
1		0.61 (.30)	0.68 (.302)		-0.20 (.31)	-0.17 (.30)
male	0	0	0	0	0	0
femal	- 1.8 (0.8)	-0.84 (.10)	-0.70 (.097)	0.1 (0.8)	-0.88 (.10)	-0.75 (.10)
white	0	0	0	0	0	0
black	-3.8 (1.2)	0.42 (.13)	-1.51 (.140)	-5.8 (1.2)	0.49 (.13)	-1.73 (.15)
hispa	-2.8 (1.1)	0.38 (.13)	-0.99 (.131)	-4.9 (1.1)	0.32 (.13)	-1.10 (.13)
as/pa	-0.1 (3.1)	0.51 (.36)	-0.46 (.374)	2.6 (3.0)	0.04 (.34)	-1.17 (.37)
amind	-6.7 (2.4)	0.05 (.29)	-1.06 (.281)	-8.5 (2.4)	0.27 (.31)	-0.90 (.28)
other	-25.5(13.9)	-0.52 (1.9)	-0.58 (1.88)	-22.6(14.3)	0.34 (2.4)	0.18 (2.3)
NE	0	0	0	0	0	0
SE	-0.5 (1.7)	0.44 (.15)	0.15 (.149)	0.2 (1.6)	0.22 (.16)	-0.03 (.14)
Cent	-2.7 (1.6)	0.29 (.16)	0.24 (.150)	-3.2 (1.8)	0.40 (.17)	0.26 (.15)
West	-4.8 (1.4)	0.33 (.14)	-0.02 (.133)	-3.9 (1.7)	0.27 (.14)	-0.16 (.13)
extru	0	0	0	0	0	0
lomet	-2.7 (2.6)	0.52 (.23)	-0.10 (.232)	-4.0 (2.5)	0.55 (.24)	-0.19 (.24)
himet	2.7 (2.9)	-0.21 (.26)	0.92 (.235)	7.5 (2.7)	-0.01 (.28)	1.16 (.22)
manct	1.4 (2.7)	0.05 (.22)	0.17 (.211)	1.5 (2.5)	0.08 (.24)	0.30 (.21)
urbfr	-1.3 (2.7)	-0.24 (.24)	0.36 (.222)	-0.2 (3.7)	0.11 (.26)	0.66 (.22)
medct	-1.6 (2.6)	-0.03 (.22)	0.20 (.198)	-2.4 (2.5)	0.06 (.23)	0.37 (.19)
smplc	2.5 (2.3)	0.30 (.21)	0.18 (.197)	1.1 (2.3)	0.55 (.22)	0.37 (.20)
nfnhs	0	0	0	0	0	0
finhs	3.4 (1.9)	-0.14 (.22)	0.32 (.249)	2.1 (1.8)	-0.04 (.22)	0.54 (.25)
smcol	5.0 (2.2)	-0.21 (.29)	1.07 (.300)	6.4 (2.2)	-0.47 (.29)	1.07 (.29)
colgr	5.2 (1.7)	-0.41 (.21)	1.00 (.234)	5.1 (1.8)	-0.40 (.21)	1.34 (.24)
DK	0.8 (1.7)	-0.51 (.20)	0.26 (.224)	1.5 (1.7)	-0.46 (.20)	0.48 (.24)
nores	0.7 (4.2)	0.37 (.45)	0.60 (.492)	-4.2 (4.5)	0.62 (.49)	1.22 (.51)
s^2_sch	.055 (.006)			.053 (.006)		
s^2	1.0			1.0		
log Lmax	-101482.05			-101320.81		
#params	458			537		

Table B4

Item parameter estimates, 3-guess-prob 2PL model

comp	1	2		3		comp	1	2		3	
item	a1	a2	b2	a3	b3	item	a1	a2	b2	a3	b3
1	2.71	0.77	1.35	3.66	1.60	41	0.36	-0.30	0.82	2.30	0.65
2	-0.16	-0.91	0.94	2.18	0.75	42	0.20	-0.80	0.39	0.98	0.63
3	4.70	2.82	1.44	3.69	1.24	43	-0.40	-2.10	-0.01	0.98	0.69
4	2.51	0.53	1.11	2.05	0.54	44	-0.42	-2.30	0.38	0.39	0.64
5	0.66	-2.07	0.87	0.10	0.75	45	-0.77	-1.82	0.50	2.00	0.85
6	-0.46	-1.18	0.72	1.51	0.80	46	-0.61	-2.07	0.97	0.45	0.50
7	-0.14	-2.10	1.05	0.97	0.44	47	-2.65	-1.63	-0.05	-1.52	0.63
8	-0.32	-1.07	0.85	0.81	0.58	48	-3.84	-2.73	0.11	-1.43	0.82
9	-0.69	-1.65	0.47	0.31	0.92	49	0.27	-0.84	0.35	1.23	0.36
10	0.49	-1.11	0.71	0.48	0.89	50	-0.35	-0.86	-0.09	1.16	0.64
11	-2.42	-4.89	0.19	-2.70	1.07	51	-0.02	-1.02	0.36	1.14	0.54
12	-1.01	-1.50	0.52	1.09	0.81	52	-1.06	-1.50	-0.02	-0.44	0.24
13	-1.59	-1.42	0.41	-1.25	0.16	53	-1.82	-2.23	-0.41	-0.05	0.65
14	-1.62	-1.38	0.79	-0.99	0.75	54	-0.99	-0.91	-0.02	-0.11	0.18
15	-1.85	-1.50	0.12	-1.26	0.56	55	-1.34	-1.57	0.01	0.17	0.83
16	-2.90	-2.02	0.22	-0.47	0.93	56	-1.64	-1.78	-0.35	-0.34	1.23
17	-2.21	-2.03	0.08	-1.48	0.75	57	-0.81	-0.80	0.19	0.02	0.84
18	-1.71	-1.37	0.15	-2.15	-0.03	58	-0.96	-1.30	0.13	1.26	0.94
19	1.71	0.76	1.56	3.67	0.82	59	0.02	-1.26	0.52	1.22	0.71
20	0.05	-0.81	0.89	3.01	1.09	60	0.47	-0.47	0.41	1.14	0.60
21	0.59	-0.16	0.68	1.47	0.62	61	-0.11	-0.26	0.23	0.63	0.48
22	0.19	-0.66	0.66	1.08	0.76	62	-1.05	-0.65	0.25	0.03	0.77
23	-0.10	-1.64	1.50	1.45	0.85	63	-0.60	-0.69	0.22	-0.16	0.18
24	0.08	-1.03	0.92	0.87	0.64	64	-0.77	-0.78	0.22	0.54	0.58
25	2.80	-1.32	0.92	1.70	1.42	65	-1.09	-0.76	0.39	-0.83	0.21
26	0.37	-11.7	-4.26	0.19	1.77	66	-1.71	-1.53	-0.01	-0.63	0.91
27	0.75	-3.40	0.00	0.41	1.91	67	-1.05	-2.06	-0.85	-0.40	0.57
28	-0.52	-1.01	0.47	0.54	0.13	68	-1.24	-0.83	0.08	-0.61	0.61
29	-0.56	-1.24	0.97	1.24	0.39	69	-37.2	-1.75	-0.27	-2.80	0.65
30	0.58	-0.69	1.64	1.64	0.59	70	1.60	0.81	0.73	2.74	0.38
31	-2.16	-4.13	1.50	-1.34	0.97	71	1.43	0.73	1.10	3.09	0.43
32	-1.53	-2.20	0.72	0.08	0.73	72	-0.01	-0.57	0.38	1.34	0.66
33	-3.11	-2.03	-0.34	-1.41	0.37	73	0.32	-0.25	0.34	0.82	0.59
34	1.19	-1.14	1.53	2.04	0.73	74	-0.99	-1.87	-0.34	-0.03	0.62
35	1.01	-0.72	1.11	2.40	0.78	75	-1.21	-1.33	-0.62	-0.64	0.67
36	0.36	-0.87	0.80	1.51	1.04	76	-0.64	-0.69	-0.27	0.53	0.70
37	2.86	0.83	2.83	2.33	1.46	77	-1.28	-1.57	-0.40	0.09	0.66
38	2.54	0.84	30.2	2.50	1.62	78	-0.06	-0.76	0.26	0.66	0.80
39	0.64	-1.10	1.26	1.35	1.16	79	-1.64	-1.84	-0.94	-3.22	1.19
40	0.67	-0.35	1.00	2.37	0.60						

Table B5

Reporting group estimates and SEs - MIMIC and mixed models, 79 items (three-level models)

	MIMIC	2-g	2-m	2-g-p	engaged membership
1					1.46 (.25)
male	0	0	0	0	0
femal	0.3 (0.8)	-2.3 (1.0)	-2.2 (0.9)	-5.1 (1.1)	0.43 (.08)
white	0	0	0	0	0
black	-25.0 (2.0)	-32.8 (1.5)	-31.4 (1.5)	-28.2 (1.6)	-0.87 (.10)
hispa	-20.4 (1.8)	-24.2 (1.5)	-24.3 (1.3)	-23.1 (1.6)	-0.71 (.10)
as/pa	-11.8 (3.9)	-9.2 (4.4)	-16.5 (3.3)	-2.5 (5.4)	-1.01 (.24)
amind	-18.8 (3.0)	-27.4 (3.1)	-24.8 (3.0)	-28.7 (3.5)	-0.28 (.28)
other	-14.3(15.0)	-24.5(20.0)	-35.8(15.5)	-25.3(19.9)	-0.56 (1.3)
NE	0	0	0	0	0
SE	-1.1 (2.2)	-2.6 (2.5)	-5.9 (2.2)	-3.0 (2.1)	-0.30 (.13)
Cent	0.8 (2.2)	-0.5 (2.4)	-1.5 (2.3)	-0.2 (2.1)	-0.27 (.13)
West	-7.2 (2.2)	-5.6 (2.2)	-6.5 (2.2)	-3.9 (1.9)	-0.35 (.12)
extru	0	0	0	0	0
lomet	-9.9 (3.1)	-8.5 (3.2)	-13.0 (4.5)	-5.6 (3.6)	-0.49 (.19)
himet	18.6 (2.9)	20.3 (3.1)	15.4 (4.1)	25.0 (3.6)	0.60 (.22)
manct	4.5 (2.6)	8.8 (3.3)	-0.3 (4.1)	7.0 (3.2)	0.05 (.19)
urbfr	7.1 (2.8)	6.2 (3.0)	1.9 (4.5)	5.8 (3.4)	0.25 (.21)
medct	5.7 (2.4)	8.0 (2.7)	0.8 (4.0)	6.9 (3.0)	0.13 (.18)
smplc	1.1 (2.5)	3.2 (2.7)	-5.1 (3.9)	2.6 (3.0)	-0.22 (.17)
nfnhs	0	0	0	0	0
finhs	7.1 (2.2)	8.0 (2.6)	11.5 (2.6)	8.3 (3.2)	0.25 (.18)
smcol	20.4 (2.9)	24.5 (3.2)	25.6 (2.8)	24.3 (3.6)	0.57 (.21)
colgr	21.1 (2.4)	22.3 (2.4)	26.1 (2.4)	21.0 (2.9)	1.01 (.17)
DK	8.6 (2.0)	8.4 (2.3)	11.8 (2.3)	6.7 (2.8)	0.57 (.16)
nores	-3.5 (4.3)	9.3 (6.5)	-3.5 (5.1)	10.8 (5.9)	-0.12 (.31)
s^2_sch	.132 (.020)	.154 (.013)	.111 (.009)	.054 (.005)	
s^2	1.0	1.0	1.0	1.0	
log Lmax	-102562.07	-102322.81	-101686.16	-102211.16	
# params	178	258	338	279	

Table B6
Reporting group estimates and SEs - MIMIC and mixed models,
79 items (three-level models)

	MIMIC	2-m	2-m-p	easier items membership	3-g
1				0.04 (.22)	
male	0	0	0	0	0
femal	0.3 (0.8)	-2.2 (0.9)	-7.7 (1.1)	0.60 (.07)	-2.6 (1.0)
white	0	0	0	0	
black	-25.0 (2.0)	-31.4 (1.5)	-26.6 (1.8)	-0.51 (.10)	-31.7 (1.6)
hispa	-20.4 (1.8)	-24.3 (1.3)	-23.1 (1.6)	-0.32 (.10)	-25.1 (1.5)
as/pa	-11.8 (3.9)	-16.5 (3.3)	-18.7 (3.3)	-0.02 (.24)	-17.4 (3.6)
amind	-18.8 (3.0)	-24.8 (3.0)	-24.7 (2.9)	-0.20 (.22)	-25.6 (2.9)
other	-14.3(15.0)	-35.8(15.5)	-41.1(17.6)	0.33 (1.4)	-42.5(17.5)
NE	0	0	0	0	0
SE	-1.1 (2.2)	-5.9 (2.2)	-3.0 (2.3)	-0.15 (.11)	-2.0 (2.1)
Cent	0.8 (2.2)	-1.5 (2.3)	-0.5 (2.3)	-0.24 (.11)	-2.0 (2.1)
West	-7.2 (2.2)	-6.5 (2.2)	-0.8 (2.0)	-0.30 (.10)	-5.0 (2.2)
extru	0	0	0	0	0
lomet	-9.9 (3.1)	-13.0 (4.5)	-3.2 (3.2)	-0.39 (.17)	-5.9 (3.1)
himet	18.6 (2.9)	15.4 (4.1)	22.9 (3.3)	0.29 (.17)	24.0 (3.1)
manct	4.5 (2.6)	-0.3 (4.1)	8.1 (3.3)	0.08 (.16)	9.0 (2.8)
urbfr	7.1 (2.8)	1.9 (4.5)	17.0 (3.8)	0.15 (.17)	18.4 (3.7)
medct	5.7 (2.4)	0.8 (4.0)	10.6 (3.1)	0.07 (.16)	12.2 (2.9)
smplc	1.1 (2.5)	-5.1 (3.9)	8.5 (3.0)	-0.21 (.15)	4.5 (2.9)
nfnhs	0	0	0	0	0
finhs	7.1 (2.2)	11.5 (2.6)	8.8 (2.8)	0.13 (.17)	10.0 (2.7)
smcol	20.4 (2.9)	25.6 (2.8)	19.5 (3.2)	0.66 (.21)	25.0 (3.0)
colgr	21.1 (2.4)	26.1 (2.4)	21.9 (2.6)	0.52 (.16)	25.0 (2.5)
DK	8.6 (2.0)	11.8 (2.3)	7.0 (2.6)	0.44 (.16)	9.9 (2.4)
nores	-3.5 (4.3)	-3.5 (5.1)	-1.6 (6.7)	0.38 (.31)	-3.9 (7.3)
s^2_sch	.132 (.020)	.111 (.009)	.134 (.013)		.131 (.014)
s^2	1.0	1.0	1.0		1.0
log Lmax	-102562.0	-101686.16	-101605.83		-101548.04
# params	178	338	358		418

Table B7

Reporting group estimates and SEs -
3-mixed MIMIC models, 79 items (three-level)

	3-g	3-g-p	easi.items membership	diff.items membership
1			1.60 (.33)	1.59 (.36)
male	0	0	0	0
femal	-2.6 (1.0)	-11.4 (1.2)	0.79 (.11)	0.26 (.12)
white	0	0	0	0
black	-31.7 (1.6)	-29.0 (1.9)	-1.17 (.15)	-0.40 (.15)
hispa	-25.1 (1.5)	-19.7 (1.8)	-0.97 (.14)	-0.42 (.15)
as/pa	-17.4 (3.6)	-11.7 (5.1)	-1.13 (.30)	-1.31 (.38)
amind	-25.6 (2.9)	-25.4 (4.0)	-0.36 (.39)	0.19 (.42)
other	-42.5(17.5)	6.0(27.9)	-1.82 (1.7)	-1.02 (1.6)
NE	0	0	0	0
SE	-2.0 (2.1)	-1.2 (2.0)	-0.36 (.18)	-0.31 (.20)
Cent	-2.0 (2.1)	1.5 (2.2)	-0.49 (.18)	-0.38 (.19)
West	-5.0 (2.2)	-4.1 (1.9)	-0.43 (.17)	-0.21 (.18)
extru	0	0	0	0
lomet	-5.9 (3.1)	-6.4 (3.6)	-0.78 (.25)	-0.42 (.27)
himet	24.0 (3.1)	19.4 (3.7)	0.91 (.30)	0.57 (.33)
manct	9.0 (2.8)	2.3 (3.6)	0.03 (.25)	-0.06 (.28)
urbfr	18.4 (3.7)	5.0 (3.6)	0.37 (.28)	0.17 (.31)
medct	12.2 (2.9)	3.5 (3.4)	0.25 (.24)	0.23 (.27)
smplc	4.5 (2.9)	3.1 (3.3)	-0.24 (.23)	-0.07 (.26)
nfnhs	0	0	0	0
finhs	10.0 (2.7)	10.0 (3.4)	0.24 (.24)	0.02 (.25)
smcol	25.0 (3.0)	23.1 (3.9)	0.71 (.29)	0.01 (.32)
colgr	25.0 (2.5)	23.1 (3.2)	1.17 (.23)	0.68 (.25)
DK	9.9 (2.4)	8.3 (3.1)	0.62 (.22)	0.26 (.23)
nores	-3.9 (7.3)	15.7 (8.5)	-0.39 (.38)	-0.85 (.42)
s^2_sch	.131 (.014)	.052 (.005)		
s^2	1.0	1.0		
log Lmax	-101548.04	-101393.33		
# params	418	459		

Appendix C
Model Parameter Estimates and SEs, 2005 Texas Survey

1. Variable names and definitions, national NAEP sample

```
Student classified as having a disability
 VALUE  LABEL
    1   Yes
    2   No

Sex of subject
 VALUE LABEL
    1   Male
    2   Female

School location
 VALUE LABEL
    1   Large city
    2   Mid-size city
    3   Fringe/large city
    4   Fringe/mid-size city
    5   Large town
    6   Small town
    7   Rural (MSA)
    8   Rural (non-MSA)
  [ 9   DoD Dependents ] (not used)

Accommodated
 VALUE LABEL
    1   Accommodated
    2   Not accommodated

Race/ethnicity
 VALUE LABEL
    1   White
    2   Black
    3   Hispanic
    4   Asian Amer/Pacif Isl
    5   American Indian
  [ 6   Unclassified  ]  (not
  [ 8   Omitted       ]  used)
```

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 141
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5, © Murray Aitkin and Irit Aitkin 2011

```
School enrollment
 VALUE LABEL
    1   1-299
    2   300-499
    3   500-699
    4   700+
  [ 8   Omitted       ] (not
  [     Missing       ] used)

Computer at home
 VALUE LABEL
    1  Yes
    2  No
  [ 8  Omitted        ] (not
  [ 0  Multiple       ]
  [    Missing        ] used)

Days absent from school last month
 VALUE LABEL
    1   None
    2   1-2 days
    3   3-4 days
    4   5-10 days
    5   More than 10 days
  [ 8   Omitted       ] (not
  [ 0   Multiple      ]
  [     Missing       ] used)

Language other than English spoken in home
 VALUE LABEL
    1   Never
    2   Once in a while
    3   Half the time
    4   All or most of time
  [ 8   Omitted       ] (not
  [ 0   Multiple      ]
  [     Missing       ] used)

Use computer to play math games
 VALUE LABEL
    1   Yes
    2   No
  [ 8   Omitted       ] (not
  [ 0   Multiple      ]
  [     Missing       ] used)
```

```
Years taught elementary or secondary
 VALUE LABEL
    1   0-4 years
    2   5-9 years
    3   10-19 years
    4   20 years or more
  [ 8   Omitted       ] (not
  [      Missing      ] used)

Natl School Lunch Prog eligibility
 VALUE LABEL
    1   Not eligible
    2   Reduced-price lunch
    3   Free lunch
  [ 4   Info not available  ] (not
  [ 5   School refused info ] used)
    6   Not participating   (recoded to 4)

Highest academic degree
 VALUE  LABEL
   [ 1   High-school diploma  ] (not used)
    2   Assoc deg/voc cert
    3   Bachelor's degree
    4   Master's degree
    5   Education specialist
    6   Doctorate
    7   Professional degree
  [ 8   Omitted               ] (not
  [ 0   Multiple              ]
  [     Missing               ] used)

Undergrad major/minor mathematics education
 VALUE LABEL
    1   Major
    2   Minor/spec emphasis
    3   No
  [ 8   Omitted               ] (not
  [ 0   Multiple              ]
  [     Missing               ] used)
```

```
Undergrad major/minor mathematics
 VALUE LABEL
    1   Major
    2   Minor/spec emphasis
    3   No
  [ 8   Omitted               ] (not
  [ 0   Multiple              ]
  [     Missing               ] used)


Mathematics education courses
 VALUE LABEL
    1   None
    2   1 or 2 courses
    3   3 or 4 courses
    4   5 courses or more
  [ 8   Omitted               ] (not
  [ 0   Multiple              ]
  [     Missing               ] used)


Technical support available at school
 VALUE LABEL
    1   Yes
    2   No
  [ 8   Omitted               ] (not
  [ 0   Multiple              ]
  [     Missing               ] used)


Software for math instruction available at school
 VALUE LABEL
    1   Yes
    2   No
  [ 8   Omitted               ] (not
  [ 0   Multiple              ]
  [     Missing               ] used)
```

```
Training for computers available at school
 VALUE LABEL
    1   Yes
    2   No
  [ 8   Omitted              ] (not
  [ 0   Multiple             ]
  [     Missing              ] used)


Number of students in this class
 VALUE LABEL
    1   15 or fewer
    2   16 - 18
    3   19-20
    4   21-25
    5   26 or more
  [ 8   Omitted              ] (not
  [ 0   Multiple             ]
  [     Missing              ] used)


Instructional materials, resources
 VALUE LABEL
    1   Get all resources
    2   Get most resources
    3   Get some resources
    4   Don't get resources
  [ 8   Omitted              ] (not
  [ 0   Multiple             ]
  [     Missing              ] used)
```

C.1 Parameter estimates and SEs – limited teacher data

Table C1

Reporting group estimates and SEs, 70-item scale

		3-PL	2-guess	2-mix	2-guess-prob	2-mix-prob
Shift				24.2(94.9)		30.8(26.8)
Disable		29.1 (2.8)	26.0 (2.5)	30.8 (2.7)	24.5 (2.8)	15.5 (3.7)
Sex		-7.0 (1.3)	-7.6 (1.2)	-7.6 (1.4)	-4.1 (1.4)	-21.2 (5.3)
Locate	(2)	2.4 (3.1)	1.6 (2.8)	1.7 (2.8)	1.5 (3.2)	-2.8 (3.2)
	(3)	-0.3 (2.6)	-3.5 (2.5)	-3.2 (2.6)	3.6 (2.7)	-1.5 (2.7)
	(4)	-7.7 (4.7)	-9.4 (3.7)	-11.2 (4.6)	-2.0 (4.3)	-17.5 (3.8)
	(5)	-3.1 (8.8)	2.6 (6.4)	0.4 (7.6)	7.2 (7.1)	7.1 (7.4)
	(6)	-3.9 (5.1)	-4.4 (3.8)	-7.2 (4.7)	3.1 (4.3)	-7.2 (6.2)
	(7)	-4.3 (6.0)	-4.8 (4.9)	-6.4 (4.7)	2.9 (5.5)	-9.3 (4.7)
	(8)	-7.5 (4.1)	-11.2 (4.1)	-11.5 (5.6)	-4.0 (4.5)	-12.7 (3.1)
Accom		13.1 (2.9)	11.1 (2.6)	9.5 (2.8)	11.3 (3.0)	13.6 (4.9)
Race	(2)	-39.7 (2.5)	-36.4 (2.2)	-40.6 (2.7)	-31.7 (2.5)	-25.4 (3.9)
	(3)	-20.0 (2.2)	-20.1 (2.0)	-22.5 (2.4)	-15.6 (2.4)	-17.0 (3.5)
	(4)	7.9 (4.1)	10.7 (3.9)	12.4 (4.7)	3.2 (5.6)	7.9 (6.3)
	(5)	-18.7(11.1)	-17.6 (8.9)	-24.4(10.2)	-19.9(11.2)	-2.7(14.7)
Enrol	(2)	2.4 (4.6)	1.9 (4.0)	2.1 (4.3)	0.1 (4.3)	5.3 (3.7)
	(3)	2.8 (4.6)	3.7 (3.9)	3.7 (4.4)	0.1 (4.3)	2.1 (4.1)
	(4)	6.0 (4.6)	7.0 (4.0)	5.6 (4.4)	5.6 (4.3)	4.0 (4.3)
Exp	(2)	5.2 (1.9)	4.2 (1.8)	3.8 (2.0)	2.7 (1.9)	0.9 (2.8)
	(3)	6.1 (1.9)	5.7 (1.8)	4.6 (1.9)	2.7 (2.0)	5.0 (2.3)
	(4)	6.5 (2.1)	5.8 (1.9)	5.1 (2.2)	2.1 (2.2)	4.4 (2.5)
LEP		20.1 (2.3)	17.5 (2.0)	19.6 (2.2)	15.8 (2.2)	2.3 (2.9)
Computer		-6.4 (1.7)	-5.5 (1.5)	-5.7 (1.6)	-3.8 (1.6)	1.4 (2.5)
Absence	(2)	-10.1 (1.5)	-10.3 (1.4)	-10.2 (1.5)	-9.8 (1.5)	-5.7 (2.1)
	(3)	-10.7 (2.2)	-10.2 (2.0)	-11.0 (2.2)	-9.2 (2.2)	-8.4 (2.6)
	(4)	-15.6 (3.5)	-13.4 (3.3)	-13.2 (3.4)	-11.0 (3.4)	-21.8 (2.9)
	(5)	-30.1 (3.9)	-25.9 (3.9)	-27.2 (4.6)	-25.1 (4.3)	-10.2 (6.2)
Lang	(2)	3.5 (1.7)	3.4 (1.6)	3.8 (1.8)	0.4 (1.8)	2.2 (2.1)
	(3)	4.7 (2.3)	4.0 (2.2)	5.6 (2.4)	4.0 (2.4)	-2.1 (3.2)
	(4)	5.0 (2.0)	5.4 (1.8)	6.0 (2.0)	6.0 (2.0)	0.4 (2.5)
Games		0.2 (1.5)	0.3 (1.4)	0.1 (1.5)	-2.0 (1.6)	0.9 (1.9)
Lunch	(2)	-14.3 (2.7)	-14.2 (2.6)	-16.1 (2.7)	-10.8 (2.9)	-16.4 (3.5)
	(3)	-22.3 (1.8)	-22.5 (1.7)	-24.5 (2.0)	-13.9 (1.9)	-13.7 (2.5)
	(4)	-36.6(23.0)	-26.1(23.5)	-40.4(19.5)	-20.3(22.4)	-8.6(23.8)
sigma^2_sch		.142(.021)	.129(.017)	.129(.021)	.119(.018)	.147(.020)
sigma^2		1.0	1.0	1.0	1.0	1.0
log L		-48555.32	-48554.45	-48398.01	-48441.81	-48313.76
# params		244	245	315	278	348

Table C1 (continued)

Reporting group estimates and SEs, 70-item scale

term	2-mix-prob	easi. items membership	2-guess-prob	engaged membership
1/shift	30.8(26.8)	-1.98 (.59)		-3.70 (1.7)
Disable	15.5 (3.7)	1.24 (.34)	24.5 (2.8)	0.59 (.50)
Sex	-21.2 (5.3)	1.03 (.34)	-4.1 (1.4)	-0.60 (.14)
Locate (2)	-2.8 (3.2)	0.37 (.21)	1.5 (3.2)	-0.30 (.22)
(3)	-1.5 (2.7)	-0.19 (.19)	3.6 (2.7)	-0.78 (.18)
(4)	-17.5 (3.8)	0.54 (.32)	-2.0 (4.3)	-2.48 (1.5)
(5)	7.1 (7.4)	-0.80 (.50)	7.2 (7.1)	-43. **
(6)	-7.2 (6.2)	0.01 (.41)	3.1 (4.3)	-1.18 (.46)
(7)	-9.3 (4.7)	0.05 (.39)	2.9 (5.5)	-2.01 (.88)
(8)	-12.7 (3.1)	-0.09 (.28)	-4.0 (4.5)	-0.78 (.32)
Accom	13.6 (4.9)	0.02 (.38)	11.3 (3.0)	1.02 (1.1)
Race (2)	-25.4 (3.9)	-1.39 (.31)	-31.7 (2.5)	-1.38 (.29)
(3)	-17.0 (3.5)	-0.53 (.28)	-15.6 (2.4)	-0.66 (.21)
(4)	7.9 (6.3)	0.37 (.51)	3.2 (5.6)	1.34 (.44)
(5)	-2.7(14.7)	-1.51 (.91)	-19.9(11.2)	0.05 (.82)
Enrol (2)	5.3 (3.7)	-0.08 (.29)	- 0.1 (4.3)	0.42 (.49)
(3)	2.1 (4.1)	0.27 (.29)	0.1 (4.3)	0.41 (.49)
(4)	4.0 (4.3)	0.21 (.31)	5.6 (4.3)	0.12 (.50)
Exp (2)	0.9 (2.8)	0.33 (.23)	2.7 (1.9)	0.39 (.21)
(3)	5.0 (2.3)	0.13 (.18)	2.7 (2.0)	0.55 (.20)
(4)	4.4 (2.5)	0.23 (.20)	2.1 (2.2)	0.69 (.20)
LEP	2.3 (2.9)	1.34 (.23)	15.8 (2.2)	1.10 (.93)
Computer	1.4 (2.5)	-0.69 (.18)	-3.8 (1.6)	-0.81 (.30)
Absence(2)	-5.7 (2.1)	-0.41 (.15)	-9.8 (1.5)	-0.22 (.15)
(3)	-8.4 (2.6)	-0.22 (.22)	-9.2 (2.2)	-0.22 (.24)
(4)	-21.8 (2.9)	0.64 (.34)	-11.0 (3.4)	-1.02 (.49)
(5)	-10.2 (6.2)	-1.76 (.52)	-25.1 (4.3)	-1.15 (.82)
Lang (2)	2.2 (2.1)	0.16 (.17)	0.4 (1.8)	0.44 (.17)
(3)	-2.1 (3.2)	0.52 (.26)	4.0 (2.4)	0.03 (.27)
(4)	0.4 (2.5)	0.46 (.21)	6.0 (2.0)	-0.36 (.34)
Games	0.9 (1.9)	-0.06 (.15)	-2.0 (1.6)	0.26 (.15)
Lunch (2)	-16.4 (3.5)	0.00 (.26)	-10.8 (2.9)	-0.50 (.27)
(3)	-13.7 (2.5)	-0.93 (.19)	-13.9 (1.9)	-1.91 (.38)
(4)	-8.6(23.8)	-49. **	-20.3(22.4)	-46. **
sigma^2_sch	.147 (.020)		.119 (.018)	
sigma^2	1.0		1.0	
log L	-48313.76		-48441.81	
# params	348		278	

** "infinite" SE

C.2 Parameter estimates and SEs – extensive teacher data

Table C2
Reporting group estimates and SEs, 70-item scale

		3PL	2-guess	2-mix	2-guess-prob	engaged membership
1/Shift						-6.36 (12.7)
Disable		22.5 (4.1)	21.0 (3.5)	23.9 (3.8)	19.4 (3.7)	0.33 (0.45)
Sex		-6.4 (1.9)	-7.4 (1.6)	-7.2 (0.9)	-5.3 (2.0)	-0.44 (0.21)
Locate	(2)	3.3 (3.8)	6.8 (3.4)	9.1 (1.9)	7.0 (3.6)	-0.42 (0.29)
	(3)	-5.9 (3.6)	-3.8 (3.5)	-4.0 (12.0)	3.3 (3.3)	-1.04 (0.27)
	(4)	-7.2 (5.4)	-5.6 (5.4)	-3.2 (2.7)	2.6 (6.3)	-42. **
	(5)	17.4(11.0)	16.7 (7.7)	12.2 (7.0)	19.9 (9.7)	-1.06 (1.91)
	(6)	-10.0 (7.3)	-8.6 (6.6)	-12.8 (5.3)	12.1 (7.7)	-44. **
	(7)	2.0(10.1)	-8.6 (6.6)	3.9 (3.3)	0.1 (10.5)	-0.67 (0.99)
	(8)	-4.8 (5.8)	-7.4 (5.3)	-7.0 (1.8)	1.2 (5.6)	-1.45 (0.50)
Accom		22.5 (4.2)	16.2 (3.7)	12.7 (5.4)	17.5 (3.7)	1.52 (1.01)
Race	(2)	-45.6 (3.5)	-40.9 (3.2)	-44.2 (6.8)	-35.6 (3.6)	-1.55 (0.37)
	(3)	-25.6 (3.0)	-25.1 (2.8)	-30.4 (6.8)	-18.9 (3.3)	-0.85 (0.29)
	(4)	3.5 (6.0)	6.4 (5.7)	5.5 (4.7)	2.5 (8.2)	0.80 (0.64)
	(5)	-12.5(11.7)	-20.5(17.4)	-30.0(13.2)	-15.5(13.3)	0.11 (1.53)
Enrol	(2)	7.3 (6.5)	4.0 (5.6)	5.5 (2.9)	1.2 (6.0)	0.44 (0.60)
	(3)	7.9 (6.2)	5.4 (5.4)	4.1 (2.3)	2.7 (5.8)	0.27 (0.59)
	(4)	11.0 (6.4)	6.3 (5.7)	5.9 (2.5)	5.7 (6.0)	-0.01 (0.61)
Exp	(2)	3.3 (2.7)	1.9 (2.3)	2.8 (1.6)	-4.2 (2.7)	1.10 (0.29)
	(3)	5.9 (2.8)	4.1 (2.5)	6.3 (3.6)	0.4 (2.9)	0.92 (0.28)
	(4)	3.9 (3.0)	3.0 (2.7)	2.1 (1.9)	-1.5 (3.1)	0.82 (0.33)
ELL		15.3 (3.0)	3.9 (2.6)	13.3 (2.2)	10.8 (2.8)	1.36 (0.84)
Computer		-6.8 (2.4)	-6.5 (2.1)	- 6.2 (1.3)	-3.4 (2.2)	-1.19 (0.55)
Absence	(2)	-14.8 (2.1)	-13.9 (1.8)	-13.0 (1.2)	-13.0 (2.0)	-0.52 (0.21)
	(3)	-14.8 (3.3)	-13.8 (2.9)	-15.2 (2.7)	-13.4 (3.1)	-0.18 (0.33)
	(4)	-14.9 (5.0)	-15.8 (4.7)	-12.0 (1.9)	-9.2 (5.4)	-1.72 (1.29)
	(5)	-27.0 (6.4)	-21.6 (5.2)	-16.8 (9.3)	-17.0 (5.7)	-1.51 (1.33)
Lang	(2)	5.3 (2.5)	4.3 (2.2)	7.6 (1.8)	2.2 (2.5)	0.22 (0.22)
	(3)	6.6 (3.2)	6.6 (2.9)	9.3 (2.1)	6.8 (3.2)	0.00 (0.34)
	(4)	5.3 (2.7)	5.6 (2.4)	9.5 (1.4)	5.3 (3.1)	-0.20 (0.60)
Games		-1.2 (2.1)	-0.4 (1.9)	0.8 (1.1)	-1.4 (2.2)	0.07 (0.26)
Lunch	(2)	-14.3 (3.8)	-15.0 (3.5)	-12.1 (2.8)	-11.6 (4.4)	-0.37 (0.38)
	(3)	-20.9 (2.6)	-21.6 (2.3)	-24.4 (1.6)	-11.9 (2.6)	-1.91 (0.34)
	(4)	-61.5(30.3)	-47.8(22.3)	-47.8(22.3)	-41.3(22.6)	-44. **
Degree	(3)	21.4(21.1)	23.9(18.5)	- 0.9(22.6)	-2.1(23.5)	3.86 (5.93)
	(4)	23.1(21.2)	25.1(18.7)	- 1.8(21.1)	0.8(23.6)	3.62 (5.92)
	(5)	20.1(22.6)	29.9(20.4)	-7.4(25.3)	-1.9(25.5)	4.15 (5.98)
	(6)	-19.3(33.0)	9.8(27.6)	-15.4(24.0)	-7.8(30.8)	-44. **
	(7)	31.7(25.4)	37.0(21.9)	2.8(25.8)	16.5(27.4)	-44. **
Mathed	(2)	-10.4(10.4)	-6.2 (9.1)	-17.6(11.6)	-11.9(20.1)	-0.59 (3.86)
	(3)	-7.4 (9.9)	-2.8 (8.8)	-11.5(10.3)	-10.6(19.9)	0.43 (3.79)
Math	(2)	-1.1(10.2)	-8.3 (9.3)	1.6(10.1)	-7.3(10.0)	0.56 (7.26)
	(3)	-0.3(10.1)	-8.9 (9.2)	3.5(12.4)	-6.2 (9.8)	-0.32 (7.00)
Mathc	(2)	-2.3 (3.0)	-1.5 (2.7)	-2.4 (2.0)	1.1 (3.1)	-0.56 (0.33)
	(3)	-1.4 (3.1)	-0.3 (2.8)	-2.0 (1.6)	-0.6 (3.2)	-0.23 (0.32)
	(4)	-1.6 (4.8)	-0.7 (4.2)	-3.6 (3.7)	-4.3 (4.9)	-0.03 (0.45)
Tech	(2)	-9.0 (4.0)	-6.9 (3.6)	-9.2 (1.8)	-7.3 (3.7)	0.07 (0.38)
Soft	(2)	1.2 (2.6)	0.8 (2.5)	2.9 (3.3)	-2.9 (2.6)	0.88 (0.27)
Train	(2)	0.9 (2.4)	0.8 (2.2)	-0.3 (2.2)	1.6 (2.3)	-0.16 (0.23)
Studnum	(2)	-6.4 (4.3)	-8.1 (3.7)	-8.1 (2.1)	-7.3 (3.9)	-0.53 (0.40)
	(3)	-1.6 (4.2)	-3.0 (3.7)	-4.0 (4.6)	-0.1 (3.9)	-0.82 (0.44)
	(4)	-2.2 (4.2)	-4.0 (3.7)	-5.1 (1.7)	-2.6 (3.9)	-0.64 (0.46)
	(5)	-12.3 (8.8)	-11.3 (7.8)	-12.1 (3.7)	-8.8 (7.5)	-2.06 (1.98)
Resourc	(2)	3.0 (2.7)	3.6 (2.5)	0.5 (1.7)	6.6 (2.8)	-0.71 (0.29)
	(3)	-1.4 (3.0)	0.2 (2.7)	-2.6 (2.4)	2.0 (3.1)	-1.06 (0.35)
	(4)	-4.6(11.7)	5.7(11.0)	6.8 (4.4)	3.5(11.6)	-0.49 (1.09)
sigma^2_sch		.177 (.035)	.123 (.024)	.138 (.038)	.086 (.023)	
sigma^2		1.0	1.0	1.0	1.0	
log L		-25204.70	-25163.94	-25033.31	-25084.35	
# params		266	267	337	323	

Appendix D
Model Parameter Estimates and SEs, 2005 California survey

D.1 Parameter estimates for MIMIC models – limited teacher data

Table D1

Reporting group estimates and SEs, 70-item scale

		MIMIC	3PL	2-guess	2-guess-prob	engaged membership	
1		-	-	-	118.5(33.1)	1.25 (0.79)	
No disability		20.8 (2.1)	26.9 (2.7)	21.7 (2.3)	10.9 (3.4)	1.29 (0.18)	
Girl		-7.2 (0.9)	-8.4 (1.2)	-7.5 (1.0)	-6.7 (1.3)	-0.30 (0.11)	
Locate	(2)	0.5 (2.1)	-0.6 (2.9)	0.2 (2.4)	-1.9 (2.6)	0.18 (0.15)	
	(3)	-1.8 (1.7)	-1.8 (2.0)	-1.8 (2.0)	-6.4 (2.3)	0.23 (0.13)	
	(4)	-8.1 (3.1)	-7.8 (3.6)	-8.6 (3.3)	-8.6 (3.6)	0.10 (0.21)	
	(5)	-14.1(13.1)	-5.3(12.6)	-16.7(14.9)	-27.9(15.1)	46	***
	(6)	4.3 (7.5)	3.9 (9.9)	5.5 (7.8)	1.0 (9.4)	0.38 (0.72)	
	(7)	-2.9 (5.8)	-3.6 (6.7)	-3.3 (6.1)	-0.9 (8.5)	-0.45 (0.58)	
	(8)	4.6 (4.9)	2.1 (6.9)	3.5 (5.3)	-2.1 (4.7)	2.38 (1.60)	
Not accom		24.2 (2.7)	32.4 (3.4)	25.0 (2.9)	21.1 (5.0)	0.95 (0.22)	
Race	(2)	-28.9 (2.1)	-32.7 (2.5)	-31.1 (2.3)	-28.3 (2.8)	-1.19 (0.21)	
	(3)	-15.4 (1.5)	-17.9 (1.9)	-16.8 (1.7)	-15.1 (2.1)	-0.69 (0.20)	
	(4)	5.1 (1.9)	5.1 (2.1)	6.7 (2.0)	4.9 (2.3)	0.27 (0.27)	
	(5)	-21.6 (5.3)	-24.3 (5.9)	-23.4 (5.3)	-15.9 (6.8)	-0.84 (0.51)	
	(6)	-3.5 (5.5)	-2.0 (7.4)	-2.4 (6.0)	1.6 (6.4)	-0.83 (0.45)	
Enrol	(2)	-3.3 (4.4)	-5.0 (5.7)	-5.3 (4.7)	-0.0 (5.1)	-0.75 (0.71)	
	(3)	-2.3 (6.7)	-3.9 (5.6)	-4.0 (4.6)	2.6 (4.9)	-0.98 (0.70)	
	(4)	-8.1 (4.2)	-11.6 (5.7)	-10.3 (4.5)	-2.1 (4.8)	-1.31 (0.70)	
Exp	(2)	2.3 (1.4)	3.0 (1.7)	2.0 (1.5)	5.3 (1.9)	-0.17 (0.13)	
	(3)	5.0 (1.5)	5.8 (1.8)	5.4 (1.6)	6.8 (1.9)	0.07 (0.15)	
	(4)	4.9 (1.8)	7.4 (2.1)	5.3 (1.9)	4.1 (2.3)	0.43 (0.21)	
Not LEP		23.2 (1.3)	29.5 (1.6)	24.8 (1.4)	22.1 (1.9)	0.86 (0.14)	
Computer		-4.9 (1.3)	-6.8 (1.6)	-6.1 (1.4)	-0.9 (1.9)	-0.41 (0.13)	
Absence	(2)	-8.6 (1.1)	-9.0 (1.3)	-9.0 (1.2)	-6.8 (1.4)	-0.47 (0.12)	
	(3)	-12.1 (1.6)	-14.2 (1.9)	-12.3 (1.7)	-14.6 (2.1)	-0.30 (0.18)	
	(4)	-11.3 (2.3)	-12.0 (2.7)	-13.1 (2.5)	-8.1 (3.1)	-0.68 (0.21)	
	(5)	-29.5 (2.6)	-33.6 (3.3)	-31.0 (2.8)	-20.5 (5.0)	-1.46 (0.24)	
Lang	(2)	1.4 (1.4)	0.5 (1.6)	1.4 (1.5)	1.2 (1.8)	0.09 (0.16)	
	(3)	7.8 (1.8)	9.1 (2.2)	7.8 (2.0)	7.2 (2.5)	0.21 (0.21)	
	(4)	7.7 (1.4)	9.4 (1.7)	7.9 (1.6)	6.4 (2.2)	0.36 (0.18)	
No Games		5.6 (1.1)	6.9 (1.3)	6.3 (1.2)	4.1 (1.4)	0.24 (0.12)	
Lunch	(2)	-8.5 (1.9)	-9.1 (2.3)	-9.9 (2.1)	-7.4 (2.4)	-0.38 (0.20)	
	(3)	-16.2 (1.4)	-18.9 (1.6)	-17.7 (1.5)	-16.8 (1.8)	-0.47 (0.15)	
	(4)	-16.1 (5.1)	-21.2 (7.0)	-18.1 (5.6)	-11.3(10.4)	-0.61 (0.40)	
	(5)	-26.3 (7.7)	-13.9 (8.7)	-8.9 (8.9)	-16.8 (8.4)	-0.64 (0.62)	
	(6)	25.7 (5.1)	27.7 (5.7)	27.1 (5.6)	26.4 (5.5)	0.68 (1.79)	
sigma^2_sch		.113 (.011)	.161 (.017)	.126 (.013)	.155 (.017)		
sigma^2		1.0	1.0	1.0	1.0		
log L		-63494.47	-63399.96	-63385.73	-63215.95		
# params		177	247	248	284		

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5, © Murray Aitkin and Irit Aitkin 2011

Table D1 (continued)

Reporting group estimates and SEs, 70-item scale

term		2-mix	2-mix-prob	easi. items membership
1/shift		96.9(21.7)	-85.8(42.3)	-2.30 (0.55)
Disable		24.1 (2.7)	12.4 (2.7)	1.25 (0.26)
Sex		-9.0 (1.3)	-20.1 (1.9)	0.75 (0.13)
Locate	(2)	0.4 (2.9)	0.9 (2.5)	-0.07 (0.16)
	(3)	-2.8 (2.2)	-4.5 (2.1)	0.03 (0.16)
	(4)	-7.4 (3.6)	-11.8 (3.6)	0.13 (0.19)
	(5)	-22.9(16.6)	6.3 (1.8)	-50 **
	(6)	-0.3 (9.5)	-5.9(10.9)	0.67 (0.55)
	(7)	4.2 (7.2)	-3.4 (5.3)	0.30 (0.52)
	(8)	1.5 (5.6)	4.1 (5.0)	-0.15 (0.41)
Accom		27.9 (3.5)	24.3 (3.9)	0.41 (0.34)
Race	(2)	-37.2 (2.7)	-30.2 (3.7)	-0.48 (0.23)
	(3)	-19.6 (2.0)	-14.8 (2.3)	-0.35 (0.16)
	(4)	7.8 (2.4)	-0.0 (2.5)	0.57 (0.19)
	(5)	-18.8 (7.0)	-16.6(10.3)	-0.28 (0.66)
	(6)	0.2 (6.1)	-6.3 (7.3)	0.16 (0.47)
Enrol	(2)	-0.9 (5.1)	2.3 (5.5)	-0.31 (0.42)
	(3)	-1.5 (5.1)	1.2 (5.9)	-0.25 (0.43)
	(4)	-8.1 (5.1)	-5.8 (5.6)	-0.25 (0.41)
Exp	(2)	1.9 (1.8)	-2.0 (2.0)	0.33 (0.13)
	(3)	5.3 (1.8)	1.9 (1.9)	0.30 (0.14)
	(4)	5.7 (2.2)	4.1 (2.2)	0.22 (0.16)
Not LEP		29.4 (1.8)	18.7 (2.3)	0.76 (0.16)
Computer		-4.7 (1.6)	-0.9 (2.1)	-0.43 (0.15)
Absence	(2)	-8.7 (1.4)	-9.2 (1.5)	-0.05 (0.11)
	(3)	-14.6 (2.0)	-8.1 (2.5)	-0.50 (0.19)
	(4)	-12.8 (2.8)	-8.3 (3.4)	-0.47 (0.25)
	(5)	-31.9 (3.5)	-24.6 (3.4)	-0.84 (0.27)
Lang	(2)	2.3 (1.8)	1.2 (1.7)	0.10 (0.13)
	(3)	9.2 (2.2)	7.0 (2.3)	0.18 (0.18)
	(4)	8.6 (1.8)	1.4 (2.1)	0.58 (0.15)
Games		6.0 (1.4)	3.1 (1.8)	0.21 (0.12)
Lunch	(2)	-10.0 (2.4)	-2.1 (2.6)	-0.73 (0.19)
	(3)	-19.8 (1.8)	-14.9 (2.2)	-0.40 (0.15)
	(4)	-19.5 (5.4)	-9.0 (8.2)	-0.81 (0.59)
	(5)	-17.2(17.4)	-17.9(14.5)	0.07 (0.77)
	(6)	25.4(11.1)	10.7 (6.4)	1.63 (0.63)
sigma^2_sch		.145 (.019)	.151 (.017)	
sigma^2		1.0	1.0	
log L		-63148.75	-63047.61	

** - "infinite" SE

Table D2

Guessing probabilities c_j for 3PL and c.d_j for 2-guess models

item	3PL	2-guess	item	3PL	2-guess
1	0.46	0.08	36	0.27	0.02
2	0	0.08	37	0.31	0.07
3	0.47	0.07	38	0	0.05
4	0.40	0.05	39	0	0.03
5	0	0	40	0	0.00
6	0.15	0.02	41	0.034	0.08
7	0.003	0	42	0.26	0.07
8	0	0	43	0.11	0.09
9	0.34	0.08	44	0.33	0.07
10	0.069	0.06	45	0	0.06
11	0	0.06	46	0	0.05
12	0.12	0.02	47	0.065	0.03
13	0.064	0.04	48	0.016	0.01
14	0	0.01	49	0.18	0.08
15	0	0.07	50	0.13	0.02
16	0.21	0.03	51	0	0.04
17	0.41	0.04	52	0.16	0.08
18	0.078	0.05	53	0.010	0
19	0	0.04	54	0	0.08
20	0.48	0.06	55	0.25	0.02
21	0.34	0.06	56	0.17	0.05
22	0.15	0.01	57	0.25	0.05
23	0.16	0.05	58	0.22	0.02
24	0.065	0.03	59	0.25	0.01
25	0.19	0.03	60	0.13	0.04
26	0	0.08	61	0.17	0.07
27	0.30	0.07	62	0.097	0.08
28	0.19	0.01	63	0.49	0.06
29	0.007	0.01	64	0.40	0.06
30	0.15	0	65	0	0.08
31	0.20	0	66	0.21	0.04
32	0	0	67	0	0.06
33	0.74	0.01	68	0.13	0
34	0.050	0.01	69	0.056	0.05
35	0.59	0	70	0	0.06

Table D3

Engaged and non-engaged item parameters for 2-guess-prob model

	Engaged		Non-engaged			Engaged		Non-engaged	
item	a2	b2	a1	Cut	item	a2	b2	a1	Cut
1	1.85	0.46	-0.03	-4.1	36	-0.49	0.62	-0.56	-0.2
2	2.41	1.12	-0.14	-2.3	37	0.57	0.87	-0.35	-1.1
3	0.98	0.37	-0.43	-3.8	38	-1.43	1.16	-51	-20
4	0.03	0.62	-0.24	-0.4	39	-2.50	1.17	-51	-20
5	-4.57	1.88	-3.32	0.7	40	-5.51	1.49	-51	-20
6	-1.44	0.73	-1.22	0.3	41	1.41	0.83	-0.69	-2.5
7	-2.74	1.55	-2.44	0.2	42	1.13	0.81	-0.33	-1.8
8	-4.44	1.05	-51	-20	43	0.70	0.65	-1.55	-3.5
9	1.46	0.62	-0.29	-2.8	44	0.40	0.45	-1.24	-3.6
10	-0.39	0.90	-1.82	-1.6	45	-0.24	1.02	-2.83	-2.5
11	-0.64	0.84	-51	-20	46	-0.27	1.17	-2.57	-2.0
12	-1.07	1.08	-1.05	0.0	47	-1.48	1.02	-1.84	-0.4
13	-0.56	0.77	-2.32	-2.3	48	-2.26	0.89	-3.96	-1.9
14	-3.55	1.12	-52	-20	49	0.87	0.71	-0.89	-2.5
15	0.26	1.20	-2.75	-2.5	50	-1.12	0.91	-1.30	-0.2
16	-0.52	0.73	-1.01	-0.7	51	-0.89	0.73	-4.56	-5.0
17	-0.17	0.45	-0.27	-0.2	52	0.55	0.76	-1.03	-2.1
18	-0.22	0.66	-2.43	-3.4	53	-2.87	1.30	-2.89	-0.0
19	-1.61	1.36	-3.04	-1.1	54	2.08	0.89	-0.46	-2.9
20	0.94	0.48	-0.15	-2.3	55	-0.24	0.89	-0.65	-0.5
21	0.94	0.77	-0.17	-1.4	56	0.27	0.94	-1.01	-1.4
22	-1.43	0.62	-1.24	0.3	57	0.31	0.91	-0.43	-0.8
23	0.42	0.78	-1.12	-2.0	58	-1.09	0.49	-1.14	-0.1
24	-0.29	1.05	-1.54	-1.2	59	-1.00	0.49	-0.84	0.3
25	-0.85	0.56	-1.56	-1.3	60	-0.38	0.96	-1.37	-1.0
26	1.09	0.64	-2.03	-4.9	61	1.55	0.83	-0.29	-2.2
27	1.04	0.77	-0.06	-1.4	62	1.93	1.00	0.10	-1.8
28	-1.13	0.46	-1.47	-0.7	63	0.47	0.47	0.03	-0.9
29	-2.32	0.91	-4.39	-2.3	64	0.74	0.85	0.14	-0.7
30	-1.81	0.48	-1.38	0.9	65	1.30	1.15	-1.05	-2.8
31	-1.39	0.48	-1.15	0.5	66	-0.66	0.78	-0.96	-0.4
32	-6.68	2.04	-52	-20	67	0.62	1.12	-1.27	-1.7
33	1.81	0.26	0.79	-3.9	68	-1.73	0.95	-1.11	0.6
34	1.00	0.91	-0.92	-2.1	69	-0.73	0.93	-2.14	-1.5
35	0.55	0.19	0.02	-4.1	70	0.00	0.68	-4.33	-6.4

D.2 Parameter estimates for MIMIC models – extensive teacher data

Table D4
Reporting group estimates and SEs, 70-item scale

term	MIMIC	3PL	2-guess	2-guess-prob	engaged membership
1				150.6(38.4)	-0.67 (1.16)
No disability	18.0 (2.4)	22.8 (3.0)	18.6 (2.6)	12.2 (4.0)	1.00 (0.23)
Girl	-6.6 (1.1)	-7.8 (1.3)	-7.1 (1.2)	-7.8 (1.4)	-0.17 (0.13)
Locate (2)	-1.3 (2.6)	-2.4 (2.8)	-2.2 (2.8)	-3.5 (2.9)	0.24 (0.20)
(3)	-2.0 (2.1)	-1.7 (2.5)	-2.2 (2.5)	-6.8 (2.6)	0.27 (0.15)
(4)	-1.9 (3.1)	-2.1 (3.3)	-1.5 (3.7)	-4.5 (3.9)	0.06 (0.28)
(6)	5.4 (7.6)	8.3 (8.0)	8.9 (8.4)	-5.6(10.2)	1.09 (0.99)
(7)	-1.3 (6.7)	-0.1 (7.4)	-2.9 (7.2)	1.5 (9.0)	-0.93 (0.61)
(8)	6.3 (5.4)	2.2 (5.8)	6.1 (5.7)	1.4 (4.9)	2.78 (2.02)
Not accom	27.1 (3.0)	35.4 (4.1)	27.8 (3.3)	22.6 (6.0)	1.23 (0.30)
Race (2)	-28.7 (2.3)	-33.2 (2.8)	-31.5 (2.7)	-28.2 (3.0)	-1.30 (0.26)
(3)	-25.3 (1.8)	-17.9 (2.0)	-16.9 (1.9)	-15.4 (2.4)	-0.63 (0.25)
(4)	4.2 (2.1)	3.8 (2.5)	3.2 (2.3)	4.4 (2.7)	0.08 (0.30)
(5)	-19.8 (7.5)	-27.3 (6.8)	-23.6 (7.1)	-11.4 (8.6)	-1.16 (0.58)
(6)	-4.2 (7.2)	-5.4 (9.7)	-4.3 (8.1)	0.2 (7.8)	-0.70 (0.60)
Enrol (2)	-1.8 (4.4)	-5.3 (5.5)	-3.6 (4.7)	1.2 (5.2)	-0.76 (0.83)
(3)	-3.5 (4.4)	-7.0 (5.5)	-6.2 (4.7)	2.0 (5.1)	-1.03 (0.81)
(4)	-10.5 (4.4)	-16.1 (5.5)	-13.2 (4.7)	-1.4 (5.0)	-1.46 (0.82)
Exp (2)	2.6 (1.7)	3.2 (2.0)	2.9 (1.9)	3.7 (2.1)	-0.14 (0.16)
(3)	4.8 (1.8)	5.0 (2.1)	5.3 (2.0)	6.4 (2.2)	-0.02 (0.18)
(4)	3.0 (2.3)	5.2 (2.6)	3.4 (2.6)	3.3 (2.7)	0.27 (0.24)
Not LEP	23.0 (1.5)	28.0 (1.8)	24.5 (1.6)	21.3 (2.1)	0.96 (0.17)
No computer	-4.4 (1.5)	-6.5 (1.8)	-5.9 (1.6)	0.5 (2.1)	-0.53 (0.15)
Absence (2)	-7.5 (1.2)	-8.0 (1.5)	-7.4 (1.3)	-6.5 (1.6)	-0.35 (0.13)
(3)	-12.3 (1.8)	-13.4 (2.1)	-12.6 (2.0)	-14.4 (2.4)	-0.20 (0.23)
(4)	-10.0 (2.6)	-10.0 (3.0)	-11.4 (3.0)	-5.5 (3.3)	-0.75 (0.24)
(5)	-29.0 (3.0)	-32.6 (3.6)	-30.7 (3.3)	-23.2 (6.0)	-1.21 (0.34)
Lang (2)	3.0 (1.5)	2.2 (1.8)	3.4 (1.7)	4.0 (2.0)	-0.03 (0.20)
(3)	9.2 (2.0)	9.8 (2.4)	10.3 (2.2)	10.5 (2.9)	0.06 (0.27)
(4)	8.3 (1.6)	7.9 (1.4)	8.8 (1.8)	9.7 (2.5)	0.20 (0.23)
No games	6.3 (1.2)	7.0 (1.5)	7.5 (1.4)	4.7 (1.6)	0.21 (0.14)
Lunch (2)	-7.8 (2.2)	-8.6 (2.6)	-8.8 (2.4)	-10.0 (2.7)	-0.14 (0.26)
(3)	-16.0 (1.6)	-18.1 (1.8)	-17.8 (1.8)	-18.6 (2.0)	-0.29 (0.18)
(4)	-15.2 (6.8)	-18.5 (6.4)	-17.9 (7.3)	-7.2 (7.6)	-0.58 (0.43)
(5)	-22.9 (9.5)	-22.1(12.1)	-21.2(10.8)	-23.1(15.8)	-1.34 (0.72)
(6)	29.9 (6.2)	33.6 (6.5)	32.8 (9.5)	25.8 (6.8)	0.81 (2.06)
Degree (4)	7.1 (1.4)	8.7 (1.7)	8.2 (1.6)	5.9 (1.8)	0.50 (0.15)
(5)	6.8 (3.3)	6.9 (3.6)	8.3 (3.6)	5.1 (4.0)	0.80 (0.43)
(6)	2.9 (9.7)	-5.7(10.4)	3.7 (9.5)	-7.1(10.2)	45 **
(7)	-6.2 (6.7)	-7.2 (6.9)	-6.9 (7.2)	3.9 (7.9)	-0.32 (0.58)
Mathed (2)	-17.7(11.3)	-27.2(14.0)	-23.2(12.1)	-9.9(16.5)	0.02 (0.98)
(3)	-12.7(11.2)	-20.1(13.8)	-17.9(12.0)	-9.4(16.4)	0.50 (1.01)
Math (2)	-1.1(13.3)	1.0(17.4)	-0.4(14.4)	-6.9(18.0)	-0.53 (0.94)
(3)	2.8(13.3)	5.9(17.5)	4.9(14.5)	-13.2(18.1)	0.15 (0.98)
Mathc (2)	1.0 (2.2)	1.2 (2.6)	1.7 (2.5)	2.6 (2.7)	-0.11 (0.24)
(3)	0.0 (2.3)	-0.1 (2.7)	0.1 (2.5)	2.8 (2.8)	-0.31 (0.25)
(4)	-1.0 (2.8)	-0.8 (3.3)	-0.7 (3.1)	-1.2 (3.5)	-0.03 (0.29)
Tech (2)	4.3 (1.8)	5.6 (2.2)	4.6 (2.1)	6.0 (2.3)	-0.16 (0.17)
Soft (2)	3.2 (1.4)	3.8 (1.6)	3.4 (1.5)	0.2 (1.9)	0.29 (0.15)
Train (2)	-4.2 (1.4)	-4.2 (1.7)	-4.0 (1.6)	-3.5 (1.9)	0.12 (0.14)
Studnum (2)	23.8 (9.0)	28.7(10.5)	22.8(10.0)	-13.4(10.9)	2.58 (1.05)
(3)	12.4 (7.6)	14.8 (8.8)	9.1 (8.3)	-12.7 (9.6)	1.59 (0.68)
(4)	2.4 (5.9)	4.8 (6.7)	-0.7 (6.5)	-22.0 (7.6)	1.07 (0.40)
(5)	9.2 (5.5)	12.9 (6.3)	8.1 (8.0)	-15.5 (7.0)	1.28 (0.35)
Resourc (2)	0.3 (1.9)	-1.1 (2.2)	0.1 (2.1)	1.3 (2.5)	-0.06 (0.19)
(3)	-0.0 (2.0)	-2.0 (2.3)	-0.4 (2.3)	0.6 (2.6)	-0.02 (0.20)
(4)	-0.6 (3.7)	-2.5 (4.5)	-1.7 (4.2)	-2.2 (4.7)	-0.25 (0.35)
sigma^2_sch	.130 (.015)	.191 (.023)	.152 (.020)	.180 (.022)	
sigma^2	1.0	1.0	1.0	1.0	
log L	-49844.81	-49769.92	-49743.69	-49610.58	
# params	199	269	270	329	

Table D4 (continued)
Reporting group estimates and SES, 70-item scale

term		2-mix	2-mix-prob	easi. items membership
1		-	-11.7(14.9)	-1.11 (1.15)
No disability		19.8(3.1)	22.0 (3.3)	-0.02 (0.25)
Girl		-8.2 (1.4)	6.4 (2.1)	-1.29 (0.18)
Locate	(2)	-3.7 (2.9)	-8.2 (3.6)	0.25 (0.21)
	(3)	-5.2 (2.4)	-7.0 (3.5)	0.12 (0.20)
	(4)	-6.0 (4.1)	-12.0 (5.0)	0.41 (0.28)
	(6)	5.6(11.0)	13.1(11.5)	-1.03 (0.68)
	(7)	0.3 (7.3)	-31.3 (8.4)	2.64 (0.90)
	(8)	1.8 (5.7)	-9.8 (5.8)	1.45 (0.64)
Not accom		30.8 (3.9)	18.2 (4.2)	1.36 (0.38)
Race	(2)	-34.8 (2.9)	-21.4 (3.8)	-1.22 (0.30)
	(3)	-18.8 (2.1)	-10.1 (3.1)	-0.79 (0.22)
	(4)	5.4 (2.5)	7.0 (3.4)	-0.10 (0.26)
	(5)	-19.6 (7.4)	8.8 (9.0)	-2.02 (0.64)
	(6)	-4.7 (6.7)	-0.1 (8.1)	-0.24 (0.54)
Enrol	(2)	-2.0 (6.6)	-3.6 (6.3)	0.40 (0.45)
	(3)	-5.5 (6.8)	1.4 (6.0)	0.15 (0.45)
	(4)	-11.8 (6.7)	-8.9 (6.0)	-0.03 (0.46)
Exp	(2)	1.8 (2.0)	5.3 (2.4)	-0.27 (0.18)
	(3)	4.8 (2.0)	6.0 (2.7)	0.05 (0.19)
	(4)	2.2 (2.6)	4.4 (3.0)	-0.10 (0.22)
Not LEP		26.8 (1.9)	18.9 (2.4)	0.84 (0.18)
No computer		-4.9 (1.7)	-8.1 (2.0)	0.10 (0.17)
Absence	(2)	-6.8 (1.4)	-5.4 (1.6)	-0.18 (0.15)
	(3)	-13.2 (2.1)	-9.0 (2.6)	-0.49 (0.22)
	(4)	-11.4 (2.9)	-20.0 (4.0)	0.48 (0.30)
	(5)	-30.8 (3.9)	-31.0 (4.7)	-0.14 (0.38)
Lang	(2)	4.6 (1.9)	2.0 (2.3)	0.18 (0.21)
	(3)	11.4 (2.4)	5.6 (3.0)	0.38 (0.25)
	(4)	10.2 (1.9)	7.9 (2.4)	0.17 (0.24)
No games		6.7 (1.4)	7.7 (1.7)	-0.02 (0.14)
Lunch	(2)	-11.6 (2.6)	-0.2 (3.3)	-0.75 (0.28)
	(3)	-19.9 (1.9)	-8.3 (2.2)	-0.79 (0.16)
	(4)	-17.3 (6.7)	-5.6 (6.1)	-0.81 (0.40)
	(5)	-31.7(13.6)	-8.1(12.1)	-0.61 (0.72)
	(6)	32.8(10.0)	31.2 (9.2)	0.12 (0.82)
Degree	(4)	8.0 (1.7)	6.4 (2.1)	0.19 (0.15)
	(5)	8.3 (4.1)	7.6 (4.2)	0.15 (0.30)
	(6)	3.7(10.3)	16.9(12.3)	-0.94 (0.87)
	(7)	-3.1 (8.0)	-7.4 (9.3)	0.26 (0.63)
Mathed	(2)	-20.4(21.1)	-42.8(16.7)	0.97 (1.25)
	(3)	-13.6(21.3)	-36.4(17.7)	1.19 (1.32)
Math	(2)	-1.2(19.3)	7.7(16.8)	-0.11 (1.15)
	(3)	2.4(19.4)	16.8(17.8)	-0.59 (1.23)
Mathc	(2)	2.5 (2.7)	3.3 (3.3)	-0.16 (0.26)
	(3)	1.1 (2.8)	4.8 (3.3)	-0.42 (0.26)
	(4)	-0.2 (3.2)	2.2 (4.0)	-0.25 (0.32)
Tech	(2)	2.3 (2.2)	5.2 (2.5)	0.08 (0.18)
Soft	(2)	4.0 (1.7)	4.2 (2.1)	-0.04 (0.16)
Train	(2)	-2.8 (1.7)	-4.5 (2.1)	0.11 (0.16)
Studnum	(2)	22.3(10.5)	26.2(11.5)	-0.88 (1.28)
	(3)	4.0 (9.0)	3.3 (9.9)	0.76 (0.67)
	(4)	-0.9 (6.5)	1.5 (7.2)	-0.27 (0.53)
	(5)	7.3 (5.9)	6.2 (6.6)	0.07 (0.46)
Resourc	(2)	-0.1 (2.2)	0.8 (2.7)	-0.10 (0.19)
	(3)	-0.3 (2.3)	1.6 (3.1)	-0.24 (0.21)
	(4)	-5.3 (4.6)	-4.3 (4.9)	-0.21 (0.34)
sigma^2_sch		.141 (.023)	.173 (.021)	
sigma^2		1.0	1.0	
log L		-49569.81	-49443.71	
# params		340	399	

Table D5

Guessing probabilities for 3PL and 2-guess models

item	3PL	2-guess	item	3PL	2-guess
1	0.44	0.10	36	0.28	0.02
2	0	0.10	37	0.35	0.08
3	0.47	0.09	38	0	0.05
4	0.41	0.06	39	0	0.03
5	0	0.00	40	0	0
6	0.16	0.02	41	0	0.10
7	0	0	42	0.24	0.09
8	0	0	43	0.11	0.09
9	0.39	0.10	44	0.34	0.07
10	0.079	0.07	45	0	0.09
11	0	0.07	46	0	0.06
12	0.11	0.04	47	0.075	0.03
13	0.069	0.04	48	0.017	0.02
14	0	0.02	49	0.15	0.10
15	0	0.08	50	0.13	0.03
16	0.22	0.05	51	0	0.04
17	0.41	0.05	52	0.18	0.10
18	0.063	0.07	53	0.011	0
19	0	0.06	54	0	0.10
20	0.48	0.07	55	0.26	0.02
21	0.32	0.08	56	0.16	0.07
22	0.14	0.01	57	0.21	0.06
23	0.14	0.06	58	0.23	0.02
24	0.050	0.05	59	0.26	0.01
25	0.19	0.04	60	0.13	0.05
26	0	0.02	61	0.16	0.09
27	0.35	0.08	62	0.11	0.09
28	0.20	0.02	63	0.51	0.08
29	0.013	0.02	64	0.39	0.09
30	0.16	0.01	65	0	0.10
31	0.21	0.01	66	0.22	0.05
32	0	0.00	67	0	0.08
33	0.75	0.10	68	0.13	0
34	0.026	0.10	69	0.063	0.06
35	0.58	0.08	70	0	0.06

Table D6

Engaged and non-engaged item parameters for 2-guess-prob model

	Engaged		Non-engaged			Engaged		Non-engaged	
item	a2	b2	a1	Cut	item	a2	b2	a1	Cut
1	4.28	0.55	0.01	-7.8	36	1.45	0.68	-0.42	-2.8
2	3.10	1.10	-0.23	-3.0	37	2.14	0.80	-0.27	-3.0
3	3.76	0.44	-0.38	-9.5	38	-0.95	1.13	-51	-20
4	2.37	0.59	-0.21	-16	39	-2.46	1.24	-5.22	-2.2
5	-6.73	1.89	-3.19	1.9	40	-7.08	1.63	-51	-20
6	0.66	0.66	-1.16	-2.8	41	2.83	0.86	-0.68	-4.1
7	-3.72	1.56	-2.25	0.9	42	2.82	0.79	-0.36	-4.1
8	-3.96	1.09	-51	-20	43	2.85	0.62	-1.54	-7.1
9	3.34	0.67	-0.04	-5.0	44	3.33	0.40	-1.38	-11
10	3.89	0.83	-1.90	-7.0	45	1.12	0.88	-6.99	-9.2
11	0.57	0.90	-51	-20	46	-0.15	1.23	-2.60	-2.0
12	-0.56	1.11	-1.01	-0.4	47	-0.31	0.96	-1.87	-1.6
13	1.04	0.78	-2.31	-4.3	48	-0.97	0.90	-3.60	-2.9
14	-2.55	1.03	-52	-20	49	2.80	0.71	-1.10	-5.5
15	0.83	1.11	-3.37	-3.8	50	0.25	0.89	-1.28	-1.7
16	1.28	0.74	-0.88	-2.9	51	0.96	0.73	-4.20	-7.1
17	2.41	0.49	-0.14	-5.2	52	2.24	0.78	-0.90	-4.0
18	1.67	0.70	-2.22	-5.5	53	-3.09	1.33	-2.90	0.1
19	-2.08	1.39	-2.72	-0.5	54	2.96	0.98	-0.26	-3.3
20	3.68	0.46	-0.26	-8.6	55	1.08	0.88	-0.50	-1.8
21	2.37	0.84	-0.04	-2.9	56	1.62	0.88	-1.13	-3.1
22	0.52	0.66	-1.20	-2.6	57	1.61	0.89	-0.56	-2.4
23	2.35	0.72	-1.54	-5.4	58	1.54	0.51	-1.11	-5.2
24	1.01	0.93	-1.93	-3.2	59	1.40	0.54	-0.63	-3.8
25	1.54	0.58	-1.43	-5.1	60	0.55	0.99	-1.40	-2.0
26	2.87	0.72	-1.78	-6.4	61	2.88	0.88	-0.21	-3.5
27	2.84	0.76	0.15	-3.5	62	2.31	1.13	0.28	-1.8
28	1.53	0.48	-1.27	-5.8	63	3.17	0.47	0.17	-6.4
29	-0.56	0.82	-5.11	-5.6	64	2.43	0.78	0.07	-3.0
30	0.99	0.44	-1.21	-5.0	65	1.41	1.24	-0.76	-1.8
31	1.61	0.41	-1.27	-7.0	66	1.19	0.76	-0.91	-2.8
32	-8.98	2.00	-52	-20	67	1.02	1.16	-1.19	-1.9
33	5.38	0.22	0.76	-21	68	-0.62	0.94	-1.07	-0.5
34	2.12	0.92	-0.86	-3.2	69	0.84	0.84	-2.21	-3.6
35	4.36	0.14	0.06	-30	70	1.95	0.70	-51	-20

References

1. Adams, R. J., M. R. Wilson and M. L. Wu (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics* 22, 46–75.
2. Aitkin, Murray (1988). Preliminary Research on NAEP (60 Month). P/J 093-57, Princeton: Educational Testing Service.
3. Aitkin, Murray (2008). Applications of the Bayesian bootstrap in finite population inference. *Journal of Official Statistics* 24, 21–51.
4. Aitkin, Murray (2010). *Statistical Inference: An Integrated Bayesian/Likelihood Approach*. Boca Raton FL: Chapman and Hall/CRC Press.
5. itkin, Murray, Dorothy A. Anderson and John P. Hinde (1981). Statistical modelling of data on teaching styles (with Discussion). *Journal of the Royal Statistical Society A* 144, 419–461.
6. Aitkin, Murray, Neville Bennett and Jane Hesketh (1981). Teaching styles and pupil progress: a reanalysis. *British Journal of Educational Psychology* 51, 170–186.
7. Aitkin, Murray, Brian Francis and John P. Hinde (2005). *Statistical Modelling in GLIM4*. Oxford: Clarendon Press.
8. Aitkin, Murray, Brian J. Francis, John P. Hinde and Ross E. Darnell (2009). *Statistical Modelling in R*. Oxford: Clarendon Press.
9. Aitkin, Murray and Nicholas T. Longford (1986). Statistical modelling issues in school effectiveness studies (with Discussion). *Journal of the Royal Statistical Society A* 149, 1–43.
10. Aitkin, Murray and Ruth Zuzovsky (1994). Multilevel interaction models and their use in the analysis of large-scale school effectiveness studies. *School Effectiveness and School Improvement* 5, 45–73.
11. Andersen, Erling B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society B* 34, 42–54.
12. Beaton, Albert E. et al. (1986). *Expanding the New Design: the NAEP 1985-86 Technical Report*. Princeton NJ: Educational Testing Service.
13. Birnbaum, Allan (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick *Statistical Theories of Mental Test Scores*, 397–479. Reading, MA: MIT Press.
14. Bock, R. Darrell and Murray Aitkin (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika* 46, 443–459.
15. De Boeck, Paul and Mark Wilson (eds) (2004). *Explanatory Item Response Models: A Generalized Linear and Non-linear Approach*. New York: Springer.
16. Boughton, Keith A. and Kentaro Yamamoto (2007). A HYBRID model for test speededness. In Matthias von Davier and C.H. Carstensen (eds.), *Multivariate and Mixture Distribution Rasch Models*, 147–156. New York: Springer.
17. Chambers, Ray (1998). Discussion. In D. Pfefferman, C. J. Skinner, D.J. Holmes, H. Goldstein and J. Rasbash, Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B* 60, 41–43.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,*157
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5, © Murray Aitkin and Irit Aitkin 2011

18. Dossey, John A., Ina V. S. Mullis, Mary M. Lindquist and Donald L. Chambers (1988). *The Mathematics Report Card, Trends and Achievement Based on the 1986 National Assessment*. Princeton NJ: Educational Testing Service.
19. DuMouchel, William H. and Greg J. Duncan (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association* 78, 535–543.
20. Garthwaite, Paul H., Jay B. Kadane, and Anthony O'Hagan (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* 100, 680–701.
21. Goldstein, Harvey (2003). *Multilevel Statistical Models* (3rd ed.). London: Edward Arnold.
22. Hutchinson, T.P. (1991). *Ability, Partial Information, Guessing: Statistical Modelling Applied to Multiple-Choice Tests*. Sydney: Rumsby Scientific Publishing.
23. Johnson, Eugene and Rebecca Zwick (1988). *Focusing the New Design: The NAEP 1988 Technical Report*. Princeton NJ: Educational Testing Service.
24. Jöreskog, Karl G. and Arthur S. Goldberger (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* 70, 631–639.
25. Klein, L.R. and James N. Morgan (1951). Results of alternative statistical treatments of sample survey data. *Journal of the American Statistical Association* 46, 442–460.
26. Lindsey, J. K. (1996). *Parametric Statistical Inference*. Oxford: Clarendon Press.
27. Little, Roderick J. and Donald B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
28. Little, Roderick J. and M.D. Schluchter (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 72, 497–512.
29. Lohr, S.L. (1999) *Sampling: Design and Analysis*. Pacific Grove: Duxbury.
30. Longford, Nicholas T. (1993). *Random Coefficient Models*. Oxford: Clarendon Press.
31. Longford, Nicholas T. (1995). *Models for Uncertainty in Educational Testing*. New York: Springer-Verlag.
32. Lord, Frederic M. (1952). A theory of test scores. *Psychometric Monographs* 7.
33. Lord, Frederic M. and Melvin R. Novick (1968). *Statistical Theories of Mental Test Scores* (with contributions by Allan Birnbaum). Reading, MA: Addison-Wesley.
34. Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174.
35. Pfefferman, Danny (1993). The role of sampling weights when modelling survey data. *International Statistical Review* 61, 317–337.
36. Pfefferman, Danny, Chris J. Skinner, D.J. Holmes, Harvey Goldstein and Jon Rasbash (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B* 60, 23–40.
37. Rabe-Hesketh, Sophia and Anders Skrondal (2008). *Multilevel and Longitudinal Modeling using Stata* (2nd ed.). College Station, TX: StataCorp.
38. Rasch, Georg (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielsen and Lydiche.
39. Raudenbush, Steven W. and Anthony S. Bryk (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.
40. Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
41. San Martin, E., G. del Pino and P. De Boeck (2006). IRT models for ability-based guessing. *Applied Psychological Measurement* 30, 183–203.
42. Särndal, Carl-Erik, Bengt Swensson and Jan Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.
43. Schafer, Joseph (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton FL: Chapman and Hall/CRC Press.
44. Skinner, Chris J. (1989) Domain means, regression and multivariate analysis. In C.J. Skinner, D. Holt and T.M.F. Smith (eds.), *Analysis of Complex Surveys*, 59–87. Chichester: Wiley.
45. Skrondal, Anders and Sophia Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Relations Models*. Boca Raton, FL: Chapman and Hall/CRC Press.

46. Tanner, Martin and Wing Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82, 528–550.
47. von Davier, Matthias and Claus H. Carstensen (2007) (eds.) *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*. New York: Springer.
48. Welsh, Alan (1996). *Aspects of Statistical Inference*. New York: John Wiley.
49. Yamamoto, Kentaro (1987). A model that combines IRT and latent class models. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.
50. Yamamoto, Kentaro (1989). *HYBRID Model of IRT and Latent Class Models* (ETS Research Report no. RR-89-11). Princeton, NJ: Educational Testing Service.
51. Yamamoto, Kentaro (1995). *Estimating the Effects of Test Length and Test Time on Parameter Estimation Using the HYBRID Model* (TOEFL Technical Report no. TOEFL-TR-10). Princeton, NJ: Educational Testing Service.
52. Yamamoto, Kentaro and Howard T. Everson (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost and R. Langeheine (eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences*, 89–98. Münster: Waxmann.

Author Index

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,*161
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5, © Murray Aitkin and Irit Aitkin 2011