

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

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 i th jackknife pair, and which is an estimate of the statistic of interest t based on an altered sample. Specifically, the i th 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 i th 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**ulti**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 $i = 1, \dots, n$, $j = 1, \dots, J$. The MIMIC model is

$$p_{ij} \mid \theta_i = \exp[a_j(\theta_i - b_j)] / \{1 + \exp[a_j(\theta_i - b_j)]\},$$

$$\theta_i \sim N(\beta' \mathbf{x}_i, 1),$$

which is equivalent to

$$\text{logit } p_{ij} = \log[p_{ij}/(1 - p_{ij})] \mid \theta_i = a_j(\theta_i - b_j),$$

$$\theta_i \sim N(\beta' \mathbf{x}_i, 1).$$

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 $-a_j b_j = \alpha_j$, 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 official 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

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

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

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

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

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

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:

$$p_{ij} | \theta_i = c_j + (1 - c_j) \exp[\mathbf{a}'_j(\theta_i - \mathbf{b}_j)] / \{1 + \exp[\mathbf{a}'_j(\theta_i - \mathbf{b}_j)]\},$$

$$\theta_i \sim N(\Gamma' \mathbf{x}_i, \Sigma).$$

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 $E(\theta_{ci}) = \mathbf{w}'\Gamma'\mathbf{x}_i$ and the variance will be $\text{Var}(\theta_{ci}) = \mathbf{w}'\Sigma\mathbf{w}$. 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 & O (H-L)	N & O (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

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

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 contribution 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 $\beta'x$, in which the vector 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 z (of around 300) of these principal variables is used instead in a regression $\gamma'z$, giving an estimated $\hat{\gamma}$ with estimated covariance matrix \hat{A} . In large samples, it may reasonably be assumed that $\hat{\gamma} \sim N(\gamma, \hat{A})$.

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 $\gamma'z = 0$.
- 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 $\hat{\gamma}'z$ 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 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 $\pi(\theta_q | \mathbf{y}_{ij})$ 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:

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

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.

Statistical Modeling of the National Assessment of
Educational Progress

Aitkin, M.; Aitkin, I.

2011, XII, 164 p., Hardcover

ISBN: 978-1-4419-9936-8