18

Multilevel Analysis of Assessment Data

Jee-Seon Kim
University of Wisconsin-Madison

Carolyn J. Anderson
University of Illinois at Urbana-Champaign

Bryan Keller
Teachers College, Columbia University

CONTENTS

Introduction

Large-scale assessment data often exhibit a multilevel structure as a result of either sampling procedures, such as stratified sampling, or contextual factors, such as school settings where students are nested within schools, or cross-cultural settings where individuals are nested within countries. Observations within a cluster are likely to be correlated with one another and their dependency should be accounted for in the data analysis to permit valid statistical inferences. Moreover, relationships among variables may vary within clusters allowing for more detailed and informative study of contextual effects and their correlates.

The notion of clustered data and related issues of dependence are not new and have been studied for a long time. For example, the concept of a *design effect* was already used in 1965 by Kish as a required adjustment to account for correlations among observations within clusters in randomized trials. However, major advances in multilevel analysis have been made during the past couple of decades, and more recently, the methodology has become increasingly popular across disciplines owing to increased computing power and the subsequent emergence of user-friendly software. There now exist several software packages developed specifically for multilevel models, including HLM, MLwiN, and SuperMix (intermixture of MIXOR, MIXREG, MIXNO, and MIXPREG), as well as modules available for multilevel analysis in several general-purpose software packages, such as lme4 and nlme in R, MIXED, NLMIXED, and GLIMMIX in SAS, and xtreg and GLLAMM in STATA. Links to multilevel modeling software are available at the *Centre for Multilevel Modelling* website (www.bristol.ac.uk/cmm/learning/mmsoftware/). Other general statistical programs that enable some form of multilevel modeling include WinBUGS, Latent Gold, and Mplus. Although software for multilevel analysis may not be difficult to use for most researchers and practitioners with basic computer skills and fundamental knowledge of regression, the implementation of the methodology with real data is not always straightforward due to the intrinsic complexity of multilevel models and the complicated structures often present in multilevel data. Many national or international datasets are also observational or quasi-experimental, have sampling weights for primary and secondary sampling units, and/or may include missing data across levels.

This chapter demonstrates a multilevel analysis of large-scale assessment data using reading achievement scores and related variables from the 2006 Progress in International Reading Literacy Study (PIRLS) within the United States (Martin et al., 2007). Often a goal of research using such datasets is to study factors that potentially impact reading literacy and related issues. Initially, we examine descriptive statistics at each level. We then consider a multilevel model specification for the data, along with associated assumptions, diagnostics, and interpretation. We also address the presence of

sampling weights across levels and the need to impute missing data at different levels.

The next section explains the PIRLS 2006 data used for this analysis and describes the variables of interest. Such examinations reflect *exploratory data analysis* (EDA) and can be conducted in various ways, for example, by inspecting distributions and/or patterns across variables and visualizing relationships among them. This step is critical to understand the data and also to check whether the planned modeling technique appears appropriate. Subsequent sections discuss steps in applying and evaluating multilevel models with large-scale data. The chapter concludes by presenting some advanced topics and remaining methodological issues in multilevel modeling, suggesting further readings, and cross-referencing other chapters in this volume for specific topics such as nonlinear multilevel models for discrete data, imputing multilevel missing data, and making causal inferences based on large-scale assessment data.

Data

PIRLS

PIRLS is an international comparative study of the reading literacy of young students. It focuses on the reading achievement and reading behaviors and attitudes of fourth-grade students in the United States and students in the equivalent of fourth grade across other participating countries. PIRLS was first administered in 2001 and included 35 countries, but expanded to 40 countries in 2006. The assessment includes a written test of reading comprehension and a series of questionnaires focusing on the factors associated with the development of reading literacy. PIRLS is coordinated by the International Association for the Evaluation of Educational Achievement.

In PIRLS, two types of reading are assumed that account for most of the reading young students do: reading for literary experience and reading to acquire and use information. In the assessment, narrative fiction is used to assess students' abilities to read for literary experience, while a variety of informational texts are used to assess students' abilities to acquire and use information while reading. The PIRLS assessment contains an equal proportion of texts related to each purpose.

The multilevel analysis in this chapter is illustrated using PIRLS 2006 data only from the United States. The data are from a two-stage cluster sample with sampling probability weights. In 2006, a nationally representative sample of fourth-grade students was selected resulting in 5190 students from 183 schools. Schools were randomly selected first, and then one or two classrooms were randomly selected within each school. Although some schools have students from two classrooms, most schools have only one classroom

in the study, and we therefore treated the data as consisting of two levels: students within schools. The number of students per school ranges from 8 to 68 with an average equal to 28. Data from both the fourth-grade student questionnaire and the school questionnaire are used.

Plausible Value Outcomes

In PIRLS 2006, students were asked to engage in a full repertoire of reading skills and strategies, including retrieving and focusing on specific ideas, making simple to more complex inferences, and examining and evaluating text features. To provide good coverage of each skill domain, the test items required a total of 5 h and 20 min of testing time. However, testing time was kept to 1 h and 20 min for each student by clustering items in blocks and randomly rotating the blocks of items through the nine student test booklets. As a consequence, no student received all items but each item was answered by a randomly equivalent sample of students.

The matrix-sampling design used in PIRLS 2006 solicits relatively few responses from each sampled student while maintaining a wide range of content representation when responses are aggregated across all students. In such designs, the advantage of estimating population characteristics efficiently comes at the cost of an inability to make precise statements about individuals. More significantly, the amount of uncertainty may be different for different individuals. The uncertainty associated with individual proficiency estimates becomes too large to be ignored and a simple aggregation of individual student scores can lead to seriously biased estimates of population characteristics (Wingersky et al. 1987).

To address this issue, a *plausible values methodology* by Mislevy (1991) was implemented in the current analysis to represent the measured reading proficiency of each student. In short, the method uses all available data to estimate the characteristics of student populations and subpopulations and uses multiply imputed scores, called *plausible values*, to account for the uncertainty. The PIRLS 2006 data consists of five plausible values representing an overall reading score for each student. We refer to them as `readingPV1` to `readingPV5` in this chapter. All five plausible values were used as outcomes in our analysis. We explain in the section "Multiple Imputation" how results from the five analyses are combined. The distributions of the five plausible values are shown in Figure 18.1. The five empirical distributions suggest that the normality assumption appears reasonable. The figure also shows that, although plausible values fluctuate within persons (e.g., five values for a subject were 525.22, 603.41, 591.00, 621.33, and 583.37), the means and standard deviations of the five distributions across persons are very close to each other, as well as their general distributional form, as one would expect given their generation from the same distributions.

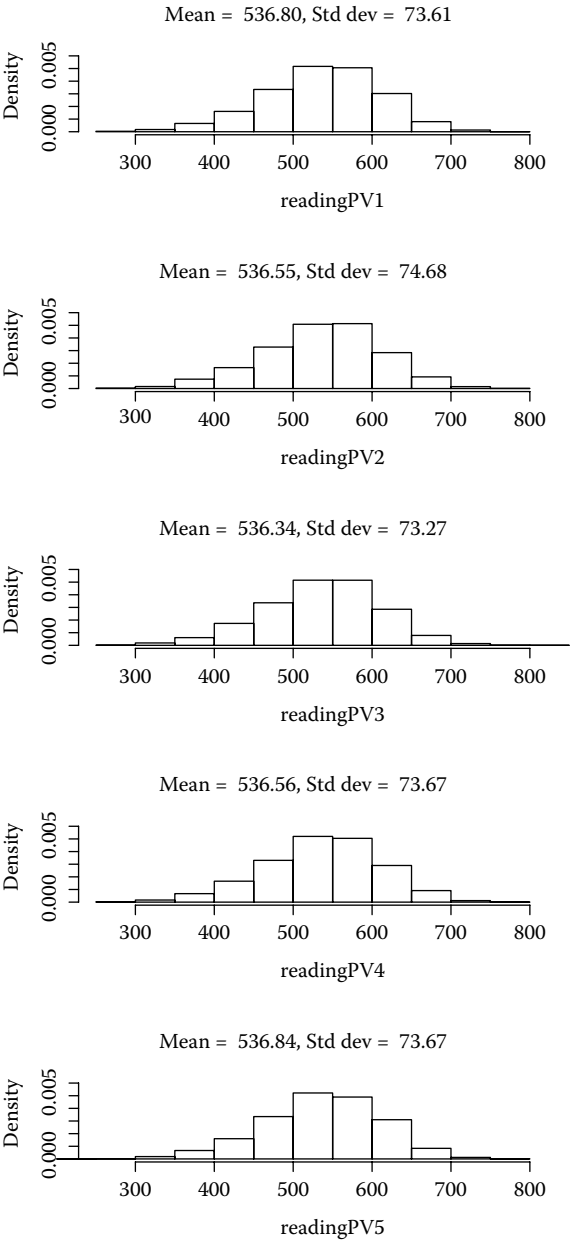

FIGURE 18.1
Histograms of plausible values 1 through 5 with sample means and standard deviations for the five reading plausible value scores.

Predictors

Background questionnaires were administered to collect information about student home and school experiences in learning to read. A student questionnaire addressed student attitudes toward reading and their reading habits. In addition, questionnaires were given to students' teachers and school principals to gather information about school experiences in developing reading literacy. In countries other than the United States, a parent questionnaire was also administered.

Along with the outcome variables, a number of explanatory variables were selected from both the student and school levels, as they are hypothesized to affect or be related to reading performance. Basic summary statistics are shown in Table 18.1. In the table, the means and standard deviations for subgroups (e.g., female vs. male) were calculated for `readingPV1` to `readingPV5` separately and then averaged. The differences of the statistics across the five plausible values are not reported here as the outcome distributions are very similar across plausible values, as shown in Figure 18.1, and the group means and standard deviations were also stable across the five outcome variables. For example, the means of `readingPV1` to `readingPV5` for females ($n = 2582$) were 541.42, 541.99, 541.54, 541.29, and 542.45. If the descriptive statistics were substantially different across the plausible values, we would have reported the variability across the plausible values.

Exploratory Data Analysis

EDA is an important part of data analysis. In this section, we describe and illustrate various graphical methods that are especially useful for multilevel analysis. The particular figures drawn depend on the nature of the predictor variables.

We start by examining the relationship between reading scores and possible microlevel predictor variables. If predictors are numerical (continuous), a plot of student data where each cell of the panel contains a plot of the response variable by a predictor variable can be very informative. Linear (alternatively quadratic, cubic, or spline) regression curves fit to each school's data can be overlaid in each cell of the panel.* Features to look for in such plots are the nature of the relationship between the predictor and response (e.g., Does a linear relationship seem reasonable? Is the relationship positive, negative, or nonexistent? Is the relationship basically the same for all schools?); the variability of the data for each school; and anything anomalous (e.g., Is a predictor that was thought to be numerical actually discrete? Is there a school that is very different from all the others? Are there outliers in the response or predictor variable?). Information about similarities and differences between schools can often be better seen by overlaying the school regression lines in a single figure.

* If there are a large number of clusters, a random sample of groups can be used.

TABLE 18.1

Information about Explanatory Variables

Variable	Levels	Percentage (%)	Mean	Std Dev
Level-1 Variables (N = 5190 Students)				
female	Female	49.78	541.74	70.59
	Male		531.51	76.46
enjoyreading	Disagree a lot	11.50	504.42	67.86
	Disagree a little	9.72	516.51	68.79
	Agree a little	24.95	536.77	68.53
	Agree a lot	53.83	549.30	74.26
bornUS	Student born in US	92.12	540.71	72.29
	Student not born in US		507.24	75.04
mombornUS	Mother born in US	67.57	546.42	71.65
	Mother not born in US	21.55	523.46	72.92
	Don't know	10.88	515.62	72.54
dadbornUS	Father born in US	62.67	547.84	71.41
	Father not born in US	21.82	524.17	72.59
	Don't know	15.51	518.24	73.14
home25books	25 or more books at home	64.50	552.01	72.09
	Less than 25 books at home		517.22	67.46
watchTV5H	Watch TV 5+ h on normal school day	28.37	513.14	70.16
	Watch TV <5 h on normal school day		549.51	71.67
comgames5H	Play computer/video games 5+ h/day	30.29	513.12	69.39
	Play computer/video games <5 h/day		550.79	71.49
Level-2 Variables (M = 183 Schools)				
econDisadv	>50% economically disadvantaged	41.67	511.99	71.88
	26–50% economically disadvantaged	23.33	542.29	72.07
	11–25% economically disadvantaged	10.00	557.84	68.39
	<10% economically disadvantaged	25.00	567.05	65.61
schoolLIB	School has a library	98.29	537.89	73.60
	School does not have library		504.45	56.65

Note: Values are based on nonmissing data. Mean and standard deviation are based on the averages across the reading plausible values.

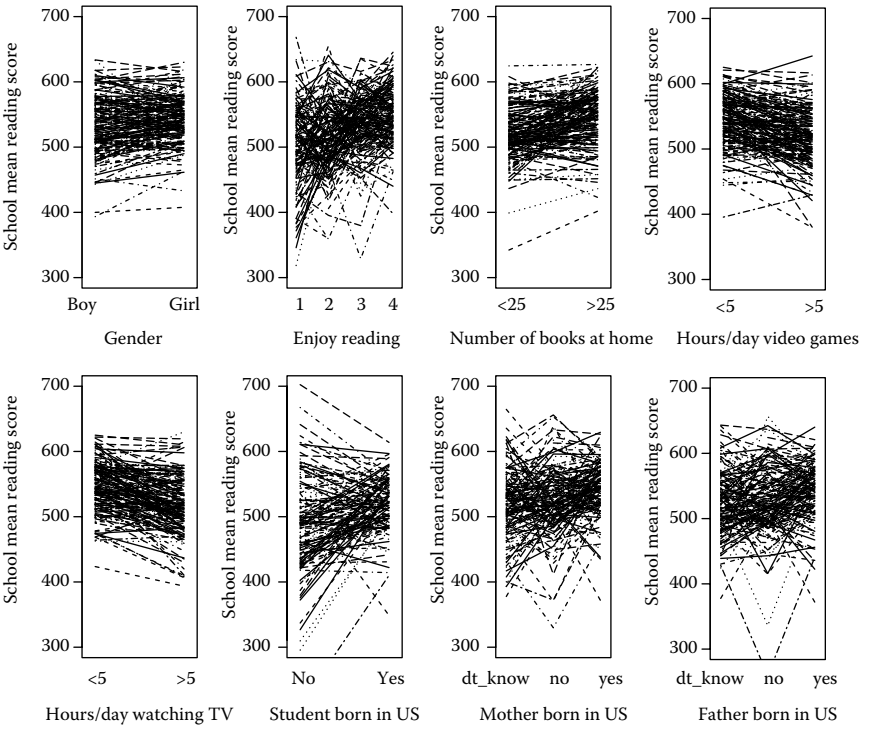

FIGURE 18.2
Panel plot of school mean reading scores plotted by each predictor variable where lines are connecting school means.

In our PIRLS example, all the predictors are discrete; therefore, we did not plot individual student data separately for each school. In Figure 18.2, we plot the school mean reading scores for each level of a predictor variable and connect the means within school by lines. The same pattern is found for each of the five plausible reading scores; therefore, we only present the figure for one of the plausible values, `readingPV1`. With the exception of the multicategory variables (`enjoyreading`, `mombornUS`, and `dadbornUS`), these lines correspond to straight lines. What is apparent in the figure is that there is considerable vertical spread among schools and this suggests that school regressions will have different intercepts. Also noticeable is the fact that some predictors have schools with vastly different slopes. For example, the slopes for `bornUS` are both negative and positive, and they show great variability among schools. Also important to note is that most students were born in the United States. The regression coefficient for `bornUS` is likely to be unstable because of the small proportion of students who were not born in the United States. The slopes for hours per day playing video games (i.e., `compgames5H`) also show considerable variability; whereas, the lines

for gender are basically flat, which suggests that gender may not be a good predictor of reading scores. Since the means for both levels of `compgames5` (more or less than 5 h per day) are based on a relatively large number of students, the model fit to the data may require different slopes for `compgames5H` for each school.

Another feature of the data that can be seen in Figure 18.2 is that all the predictors should be treated as discrete even if they have three or more categories. For example, the relationships between reading scores and `mombornUS`, `dadbornUS`, and `enjoyreading` are not linear. Also from Figure 18.2, we can get a sense of the general direction of the relationship. Reading scores on average are shown to be higher for girls, larger values of enjoyment of reading, more books in the home, fewer hours playing video games, fewer hours watching TV, students born in the United States, and students whose mothers and fathers were born in the United States. To see the direction of the relationship more clearly, in Figure 18.3, the means taken over all schools and plausible values are plotted by each predictor variable, also shown in Table 18.1. Although the data are clustered, these means are unbiased estimates of

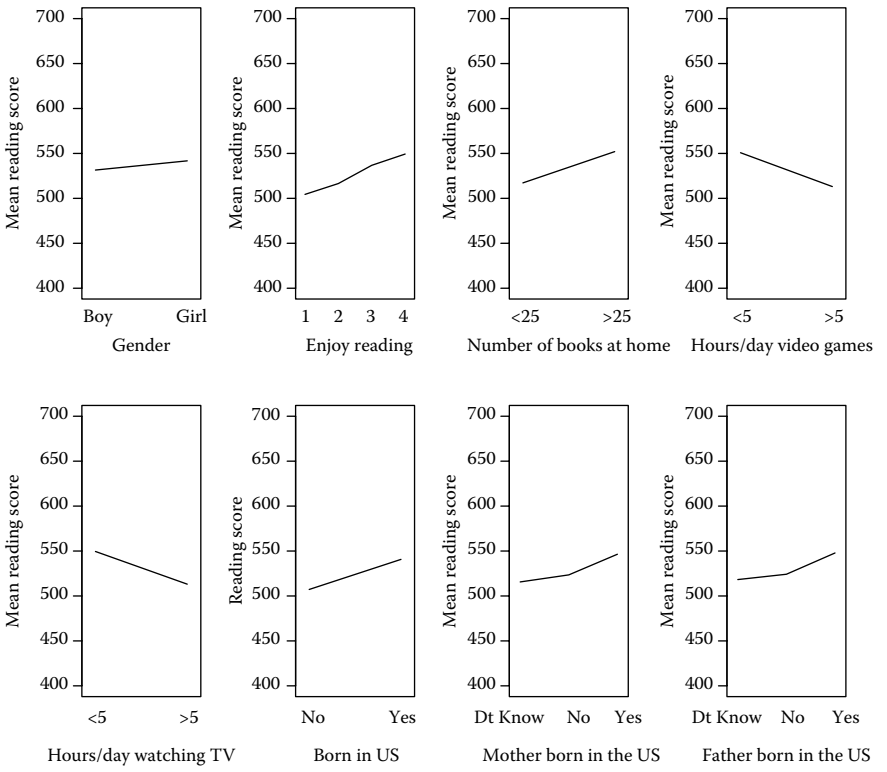

FIGURE 18.3
Panel plot of marginal mean reading scores by each predictor variable.

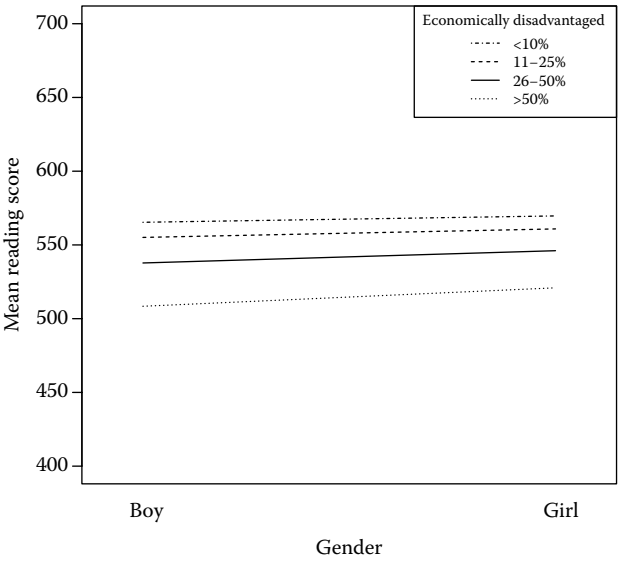

FIGURE 18.4
Mean reading scores plotted by gender with different lines (means) for different school percentages of students from economically disadvantaged families.

the effects of each predictor on the response (Diggle et al. 1994; Verbeke and Molenberghs 2000). In the models, the (fixed) regression parameters that are the same for all schools (and the average school) are expected to be in the same direction as the means illustrated in the figure.

Figures 18.1 and 18.2 focused only on effects at the level of the student. In the next figure, we illustrate how effects at the school level can be examined. In Figure 18.4, the mean reading scores are plotted for boys and girls at each level of `econDisadv`.* The lines for `econDisadv` are parallel and suggest there is no interaction between gender and `econDisadv`; however, the vertical spread between the level of economically disadvantaged students at a school suggests that this predictor may help account for differences between school intercepts. The schools with the lowest percentage of economically disadvantaged families (top curve, <10%) have the highest mean reading scores, followed by levels of 11–25% and 26–50%, while the lowest reading scores are for schools where >50% of families are disadvantaged.

As a final form of EDA, we fit normal least squares regression models with only the student (micro)-level predictor variables to each school's data. Since the intercept and regression coefficients are school specific, these regressions represent the best possible model fit for the data where goodness of fit can

* If a school-level predictor is continuous, a figure similar to Figure 18.4 can be created by discretizing the predictor and computing means for the artificial levels. However, when the time to model the data comes, the predictor should be treated as a continuous variable.

be measured by multiple R^2. Since schools with smaller numbers of students will tend to be fit better by the model than schools with larger numbers of students, the multiple R^2s for each school are plotted against school sample size. For our PIRLS data, these are plotted in Figure 18.5. Most of the schools' multiple R^2s are between .30 and .90, indicating that this set of student-level predictors could lead to a good level 1 or student-specific model for the data.

As a global summary statistic, we also computed meta-R^2 as

$$\text{meta-}R^2 = \frac{\sum_{j=1}^{M} (\text{SSTOT}_j - \text{SSE}_j)}{\sum_{j=1}^{M} \text{SSTOT}_j}$$

where SSTOT_j and SSE_j are the total sum of squares and the error sum of squares, respectively, for school j (Verbeke and Molenberghs 2000). Meta-R^2 measures the proportion of total within-school variability that can be explained by linear regression. For our PIRLS data, meta-R^2 equals .49. Both the R^2 and meta-R^2 can be used to compare plausible, student-specific models for the data.

Other exploratory analyses are possible that focus on the random aspects of the data and measure how well alternative models may fit the data (e.g.,

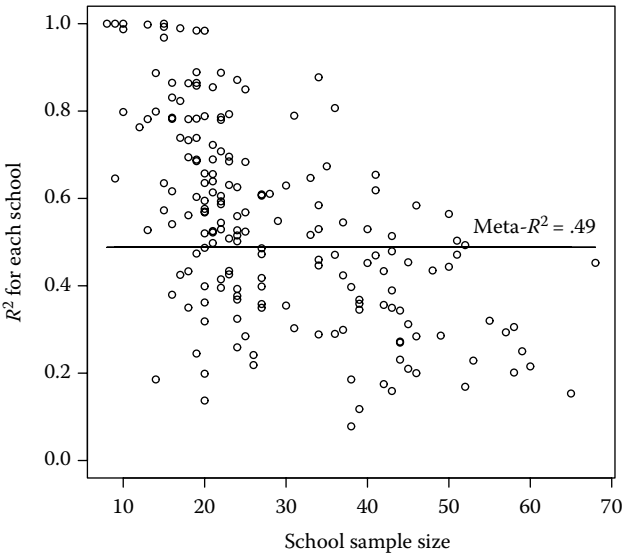

FIGURE 18.5
Multiple R^2s from linear regressions fit to each school's data plotted against school size with meta-R^2 indicated by the horizontal line.

see Diggle et al. 1994; Verbeke and Molenberghs 2000); however, a better understanding of multilevel modeling is required to explain such analyses. Therefore, we forgo these in favor of presenting more about the basic models and their properties. While conducting our EDA, we noticed that there are missing values on both student and school predictors. In the next section, we discuss how we dealt with missing data.

Missing Data

Although the percentages of missingness are not alarmingly high for many variables in PIRLS, most variables have missing values, as is often the case in large-scale assessment data. The patterns and percentages of missingness at the student and school levels are summarized in Tables 18.2 and 18.3. Each row represents a pattern of missing data where an "X" indicates that the column variable does not have missing values and "." indicates that the variable has missing values. The last two columns give the frequency and percentages with which each of the row patterns occur. The variables are ordered from those with the least missingness to those with the most missingness. For example, `female` has the least missing with three missing values (0.06%) while `comgames5H` has the most missing with 265 missing values (5.11%) at the student level. Among the 5190 students and 183 schools, 4570 (88.05%) students and 173 (94.54%) schools have no missing values. All of the variables in Table 18.1 as well as the five reading plausible values were used for imputing missing predictor values.

Multiple Imputation

Missing data are a major problem in most large-scale datasets and the clustered structure of multilevel data adds another layer of complexity to the treatment of missing observations. Simply removing students or whole schools due to missing data is not only a waste of information but also can result in biased parameter estimates (Allison 2002; Enders 2010; Schafer 1997; Van Buuren 2011, 2012). The program Mplus (Muthén and Muthén 2010) can fit models using maximum likelihood estimation (MLE) with missing values on the response variable, but MLE cannot handle missing data for the predictor or explanatory variables. A multiple imputation approach can impute missing response and predictor variables. The missing data mechanism we assume is *missing at random (MAR)*, which implies that conditional on the observed data, the missing values are independent of the unobserved data or, in other words, the missingness itself contains no information about the unobserved values that is not also contained in the observed data (Rubin 1976; Snijders and Bosker 2012).

Plausible values are provided for reading proficiency and thus no imputation is needed for outcomes, but we have missing data at both student and school levels with discrete predictors. Relatively little research has been conducted regarding missing clustered data. Most of the work pertains to either normally distributed variables (e.g., Chapter 20 of this volume by Shin; Shin

TABLE 18.2

Missing Data on Student Questionnaire

Pattern	female	enjoyreading	dadbornUS	mombornUS	bornUS	home25books	watchTV5H	comgames5H	Frequency	Percentage
1	X	X	X	X	X	X	X	X	4570	88.05
2	.	X	X	X	X	X	X	X	2	0.04
3	X	.	X	X	X	X	X	X	62	1.19
4	X	X	X	X	.	X	X	X	6	0.12
5	X	X	X	.	X	X	X	X	1	0.02
6	X	X	.	X	X	X	X	X	1	0.02
7	X	X	X	X	X	.	X	X	116	2.24
8	X	X	X	X	X	X	.	X	13	0.25
9	X	X	X	X	X	X	X	.	34	0.66
10	X	X	.	.	X	X	X	X	2	0.04
11	X	.	X	X	X	.	X	X	7	0.13
12	X	X	X	X	.	.	X	X	1	0.02
13	X	.	X	X	X	X	.	X	17	0.33
14	X	X	X	X	X	.	.	X	2	0.04
15	X	.	X	X	X	X	X	.	9	0.17
16	X	X	X	X	X	.	X	.	2	0.04
17	X	X	X	X	X	X	.	.	171	3.29
18	X	X	.	.	.	X	X	X	38	0.73
19	X	.	X	.	X	X	X	.	1	0.02
20	X	.	X	X	X	.	X	.	2	0.04

continued

TABLE 18.2 (continued)

Missing Data on Student Questionnaire

Pattern	female	enjoyreading	dadbornUS	mombornUS	bornUS	home25books	watchTV5H	comgames5H	Frequency	Percentage
21	X	.	X	X	X	X	.	.	8	0.15
22	X	X	X	.	X	X	.	.	1	0.02
23	X	X	X	X	X	.	.	.	15	0.29
24	X	.	.	.	.	X	X	X	1	0.02
25	X	X	.	.	.	.	X	X	57	1.10
26	X	.	X	X	X	.	.	.	1	0.02
27	.	X	.	.	.	.	X	X	1	0.02
28	X	.	.	.	.	.	X	X	25	0.48
29	X	.	.	.	.	X	.	X	2	0.04
30	X	X	.	.	.	.	.	X	1	0.02
31	X	X	.	.	.	X	.	.	4	0.08
32	X	X	.	.	.	.	.	.	4	0.08
33	X	.	.	.	.	.	.	.	13	0.25
Frequency	3	148	149	151	153	247	252	265	5190	
Percentage	0.06	2.85	2.87	2.91	2.95	4.76	4.86	5.11		

Note: "X" indicates a nonmissing value and "." indicates a missing value.

TABLE 18.3

Missing School-Level Data

Pattern	econDisadvN	schoolLIB	Frequency	Percentage
1	X	X	173	94.54
2	.	X	2	1.09
3	X	.	7	3.83
4	.	.	1	0.55
Frequency	3	8	183	
Percentage	1.64	4.37		

Note: "X" indicates a nonmissing value and "." indicates a missing value.

and Raudenbush 2007, 2010) or uses simple cases without missing level-2 variables (Van Buuren 2011, 2012). Neither of these two solutions works for our data. Some proposals have been put forth that include dummy variables for each cluster (Reiter et al. 2006); however, this presupposes that clusters only differ in terms of their intercepts. In our case, we do not want to make this assumption and due to the large number of clusters, this method is not pursued here. Also rare are proposals for incorporating sampling weights into the imputation model for missing data. An exception is Amer (2009), but this example only deals with two clusters.

There are two general approaches to impute multivariate missing data: joint modeling (JM) and fully conditional specification (FCS) (Van Buuren 2012). JM imputes missing values in multiple variables simultaneously based on the assumption that the data can be described by a multivariate distribution. The multivariate normal distribution is the mostly widely applied, although any multivariate distribution can be assumed in theory. On the other hand, FCS imputes data on a variable-by-variable basis. FCS requires an imputation model for each incomplete variable as the outcome, and creates imputations per variable in an iterative fashion. As FCS directly specifies the conditional distributions from which draws should be made, the method does not require specification of a multivariate model for multiple variables. An overview of similarities and differences between JM and FCS is provided in Van Buuren (2007).

For missing values among binary and ordinal predictors that do not likely follow a multivariate normal distribution, we used FCS as implemented in the package mice in R (Van Buuren 2011). We used two datasets for imputation; the school data and student data. The student data ($N = 5190$) consist of all level-1 and level-2 variables, where the level-2 values are constant for students within the same schools. The school data ($M = 183$) consists of the same number of variables, where the means of the student-level variables are used as the corresponding school-level variables. Although we acknowledge that our approach of imputing missing values at each level separately is not optimal, imputing missing values while preserving the multilevel structure is not simple; more complex approaches that may be

superior are described in Swoboda (2011) and Kim and Swoboda (2012). However, we at least use information across levels to improve the quality of the multiple imputation. For example, the five plausible values at the student level are used to impute the missing values of the school-level predictors, and the school-level predictors are used to impute missing values of the student-level predictors.

There appears to be no consensus on the appropriate number of imputed datasets (Allison 2002; Enders 2010; Schafer 1997; Van Buuren 2012), and the answer may depend on the missing patterns and frequencies. We imputed five datasets for the predictors and crossed them with the five plausible value outcomes, resulting in 25 datasets for further analysis. The different models were all fit to each of the 25 imputed datasets, and we thus obtained 25 different outcomes for each model.

To synthesize results from the multiple datasets for each model, we used Little and Rubin's procedures for combining multiple results, which appears to be the generally accepted standard in the imputation literature (Little and Rubin 2002; Snijders and Bosker 2012). The Little and Rubin procedure accounts for two sources of imprecision in the parameter estimates: *the within-dataset uncertainty* and *the imputation uncertainty.* The within-dataset uncertainty is accounted for by the standard error of estimate, and the imputation uncertainty is accounted for by the variance of the estimates across datasets. This procedure can be presented as follows: Let θ be a parameter of interest. The average estimate is obtained across M_{imp} imputed datasets:

$$\bar{\theta} = \frac{1}{M_{\text{imp}}} \sum_{m=1}^{M_{\text{imp}}} \hat{\theta}_m. \tag{18.1}$$

The within-dataset uncertainty is the average of the squared standard errors,

$$\bar{W} = \frac{1}{M_{\text{imp}}} \sum_{m=1}^{M_{\text{imp}}} \text{SE}\left(\hat{\theta}_m\right)^2,$$

and the between-imputation uncertainty is the variance of the estimate across the multiple datasets:

$$B = \frac{1}{M_{\text{imp}} - 1} \sum_{m=1}^{M_{\text{imp}}} \left(\hat{\theta}_m - \bar{\theta}\right)^2.$$

These two sources of uncertainty are used to compute the standard errors of the average estimate by taking the square root of a weighted sum of the two sources as follows:

$$\mathrm{SE}(\bar{\theta}) = \sqrt{\overline{W} + \left(1 + \frac{1}{M_{\mathrm{imp}}}\right)B}. \tag{18.2}$$

When we later fit multilevel models, Equations 18.1 and 18.2 are used to report results in the tables.

Multilevel Modeling

Multilevel modeling is a statistical methodology with many alternative names, including hierarchical linear modeling, mixed modeling, random-effects modeling, nested modeling, random coefficient modeling, and variance component analysis. Not surprisingly, the notation for multilevel models varies across the literature and software used in various disciplines. While we attempt to use notation that is largely consistent with those of the other chapters in the volume, readers should expect some differences across other resources for multilevel analysis.

Despite subtle or sometimes not-so-subtle differences in notation and terminology across disciplines, there are a number of core concepts that are consistent, such as intraclass correlation (ICC), fixed and random effects, random intercepts and random slopes (a.k.a. random coefficients), centering, reliability of aggregated variables, within- and between-group covariability, cross-level interactions, and slopes as outcomes models. Owing to space limits, we provide a list of multilevel modeling textbooks at the end of this chapter and refer readers to these books for definitions and examples of these concepts.

Presentation of the Model

A single-level regression model can be generalized to a multilevel model by allowing a regression coefficient to be random over clusters. Let the outcome variable (e.g., reading proficiency) for an individual i in cluster (e.g., school) j be Y_{ij}, where $i = 1,\ldots,n_j$ and $j = 1,\ldots,M$. The distribution of Y_{ij} within a cluster is assumed to be normal. The level-1 cluster-specific model can be written as

$$Y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} x_{pij} + \varepsilon_{ij}, \tag{18.3}$$

where β_{0j} is the intercept for cluster j, β_{pj} is the regression coefficient of cluster j for predictor variable x_{pij}, $p = 1,\ldots,P$, and ε_{ij} is the student-level random effect or error term, usually assumed to be independently and normally distributed with a zero mean and an unknown variance σ^2 (i.e., $\varepsilon_{ij} \sim N(0, \sigma^2)$ *i.i.d.*).

While the level-1 model accounts for variability at the student level, the level-2 model describes variability between clusters (schools in our analysis). The level-2 models are linear models for each of the level-1 regression coefficients:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + U_{0j} \\
\beta_{1j} &= \gamma_{10} + \sum_{q=1}^{Q} \gamma_{1q} z_{qj} + U_{1j} \\
&\vdots \\
\beta_{Pj} &= \gamma_{P0} + \sum_{q=1}^{Q} \gamma_{Pq} z_{qj} + U_{Pj},
\end{aligned} \tag{18.4}$$

where the z_{qj}'s are predictors or explanatory variables that model systematic differences between clusters, $q = 1,\ldots,Q$, the γ_{p0}'s are level-2 intercepts, the other γ's are fixed regression coefficients for level-2 predictors, and the U_{pj}'s are unobserved random effects (residuals).

The unexplained, random, or stochastic between-cluster differences are modeled as random effects. The distributional assumption for the U_{pj}'s in Equation 18.4 is

$$\boldsymbol{U}_j = \begin{pmatrix} U_{0j} \\ U_{1j} \\ \vdots \\ U_{Pj} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}\ \tau_{10} \cdots \tau_{p0} \\ \tau_{10}\ \tau_{11} \cdots \tau_{p1} \\ \vdots\ \ \vdots\ \ \ddots\ \ \vdots \\ \tau_{p0}\ \tau_{p1} \ldots \tau_{pp} \end{pmatrix} \right) i.i.d., \tag{18.5}$$

where *MVN* stands for multivariate normal and *i.i.d.* is for independent and identically distributed. For short, $\boldsymbol{U}_j \sim MVN(\boldsymbol{0}, \boldsymbol{T})$ *i.i.d.*

Substituting the level-2 equations into the regression coefficients of the level-1 model yields a linear mixed-effects model

$$Y_{ij} = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} x_{pij} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + \sum_{p=1}^{P} \sum_{q=1}^{Q} (\gamma_{pq} z_{qj} + U_{pj}) x_{pij} + U_{0j} + \varepsilon_{ij}. \tag{18.6}$$

The mixed-effects model shows that γ_{00} is the overall intercept, the γ_{p0}'s are fixed regression coefficients for level-1 effects, the γ_{0q}'s are fixed regression coefficients for level-2 effects, and the γ_{pq}'s are fixed effect regression coefficients for cross-level interaction terms. The first subscript on γ_{pq} corresponds to the predictor in the level-1 model and the second subscript corresponds

to the predictor in the level-2 model. For example, γ_{32} is the regression coefficient for the interaction between x_{3ij} and z_{2j}. Random effects consist of the random intercept U_{0j}, random slopes U_{pj}, and the error term ε_{ij}.

Empty Model

Whereas multilevel analysis investigates relationships between predictors and outcomes in the context of hierarchical data, it is often useful to consider *the empty model* before entering predictors into the model. The empty model consists of the outcome variable (e.g., reading proficiency) and the group membership identification variable (e.g., school ID#) as the only variables in the model, and is also known as *the null model* or *the unconditional means model.* The empty model can be presented as

$$Y_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij}. \tag{18.7}$$

Equation 18.7 represents the simplest multilevel model in the two-level case, and is identical to a one-way random effects ANOVA model. In addition to the grand mean of the outcome value, γ_{00}, the model reflects the variability of the outcome variable as the sum of the between- and within-group variance components:

$$\text{var}(Y_{ij}) = \text{var}(U_{0j}) + \text{var}(\varepsilon_{ij}) = \tau_{00} + \sigma^2. \tag{18.8}$$

The covariance between two observations within the same cluster j is equal to the variance of U_{0j}, and can be viewed as the amount of variance shared by the two units (e.g., two students within the same school) as a result of context:

$$\text{cov}(Y_{ij}, Y_{i'j}) = \text{var}(U_{0j}) = \tau_{00}. \tag{18.9}$$

Equation 18.8 shows that the model assumes a constant variance within each cluster. This assumption is not required for multilevel models with random slopes (see the section "Model Specification and Comparison"). Equation 18.9 shows there exists dependency within clusters, unlike the independence assumption in single-level models. It also demonstrates that the covariance between observations within each cluster is equal to the between-group variance.

Based on the between- and within-group variance components, we can calculate the correlation between two observations from the same cluster:

$$\text{corr}(Y_{ij,}Y_{i'j}) = \frac{\tau_{00}}{\tau_{00} + \sigma^2}, \tag{18.10}$$

which is known as the *intraclass correlation coefficient* and denoted as $\rho_I(Y)$. The intraclass correlation (ICC) can be interpreted as the correlation between two observations within a cluster (as defined in Equation 18.10) and also as the proportion of the variance accounted for by groups. In other words, the ICC indicates the degree of association among observations within clusters. The value can be as small as zero (when $\tau_{00} = 0$) and as large as one (when $\sigma^2 = 0$). A large ICC reflects a high degree of resemblance within clusters and strong dependence among observations. The ICC can also be calculated after entering predictors to the model, in which case it is referred to as a *residual* ICC.

Note that although $\rho_I(Y)$ is often simplified to ρ_I, the ICC depends on the variable of interest and different ICCs would be calculated for different variables. For example, an ICC of mathematics scores can be higher than an ICC of history scores, indicating a stronger correlation among mathematics scores within the same classes or schools than history scores. In the PIRLS 2006 US data, we fit empty models using `readingPV1` to `readingPV5` separately and combined the results (see the section "Multiple Imputation"). The average ICC for reading proficiency was 1244.05/(1244.05 + 4282.11) = .225 with values of .223 to .230 across the five plausible values. The estimated ICC suggests a total variance of 5526.16, about 22.5% of which is attributable to differences between schools.

Modeling Strategy

Statistical modeling is a process that is guided by substantive theory, the results of exploratory analysis, and results from fitting various models to the data. The two most common approaches advocated in the literature on linear mixed models are "step-up" and "top-down" methods (Ryoo 2011). The step-up method starts with a simple level-1 model including a random intercept to which fixed effects are successively added followed by random effects (Pinheiro and Bates 2000; Raudenbush and Bryk 2002). The top-down approach starts with the most complex polynomial model representing (level 1) effects and a random intercept, where the first step is to determine the correct order of the polynomial and the second step is to build the random effects part of the model (Diggle et al. 1994; Verbeke and Molenberghs 2000). The top-down approach lends itself better than the step-up approach to longitudinal data where change is not always linear. The little research that exists on the subject of modeling approaches with linear mixed models has found that the step-up approach tends to identify the true model in simulations more effectively than the top-down method (Ryoo 2011). In this section, a version of the step-up approach is used where we start with simple models and work toward more complex ones. Besides the work by Ryoo (2011), a reason for preferring this approach is that a model with all potentially interesting fixed and random effects suggested by the exploratory analysis may fail as a starting model because complex models are often not supported by the data; that is, such models frequently fail to converge or yield improper solutions (such as an improper covariance matrix $\boldsymbol{T}$).

Model building in multilevel analysis consists of selecting fixed effects for the mean structure and random effects for the covariance structure. The two structures are not independent. The mean structure aims to explain the systematic part of the variability of the outcome (e.g., reading proficiency in our example) and the covariance structure helps to account for the random or unexplained variance in the data. Whereas traditional single-level regression such as ANCOVA models focus mainly on the mean structure, it is critical to specify the appropriate covariance structure in multilevel models because an underparameterized covariance structure invalidates statistical inferences and an overparameterized covariance structure leads to inefficient estimation and poor standard errors (Verbeke and Molenberghs 2000). Therefore, both appropriate mean and covariance structures are essential for making valid conclusions and proper predictions.

Although no strategy guarantees the optimal model or a model that satisfies all assumptions, a combination of general modeling guidelines and EDA can help find a good model with appropriate mean and covariance structures. We propose an eight-step procedure for model building as follows. First, examine the data using EDA, including the strategies in the section "Exploratory Data Analysis." Second, specify systematic (fixed effects) that are of theoretical importance. Third, build a random-effects structure (possibly based on results of EDA). Fourth, revisit and revise fixed effects based on EDA and conduct tests for the fixed effects. Fifth, retest random effects and possibly revise. Sixth, repeat/cycle through the fourth and fifth steps. Seventh, conduct model diagnostics and residual analysis. Eighth, and finally, interpret results. We followed this eight-step strategy for our analysis in this chapter.

Model Specification and Comparison

After the examination of the empty model, we included a set of theoretically important fixed effects (Model 1), examined potential random effects (Models 2 and 3), and considered other predictors that might also be important (Model 4). The parameter estimates and model fit indexes of Models 1 through 4 are shown in Table 18.4. The fixed effect estimates are the means of the estimates taken across the 25 imputed datasets using Equation 18.1, and the corresponding standard errors are calculated using Equation 18.2 in the section "Multiple Imputation".

One can use either model-based or empirical standard errors to compute test statistics for the fixed effects. The model-based standard errors are obtained under the assumption that the covariance matrix of the observations is specified correctly, and the empirical standard errors are based on the residuals. The latter are also referred to as Huber–White's robust sandwich standard errors (Huber 1967; White 1980). It is known that model-based standard errors tend to be underestimated when a model is misspecified, whereas the empirical standard errors are relatively robust to model misspecification (Diggle et al. 1994; Raudenbush and Bryk 2002). Kim and Frees

TABLE 18.4

Model Coefficients, Standard Errors, and Fit Indexes, Averaged across 25 (5 Plausible Outcome Values × 5 Independent Variable Imputations) Datasets

Model	Model 1	Model 2	Model 3	Model 4
Fixed Effects		**Coefficient (SE)**		
Intercept	544.93 (6.08)	545.29 (6.14)	544.19 (6.15)	523.72 (11.95)
female	2.05 (2.20)	2.10 (2.19)	2.08 (2.19)	−3.00 (3.64)
enjoyreading (agree a little)	−9.32 (2.29)	−9.31 (2.29)	−9.18 (2.29)	−9.13 (2.30)
enjoyreading (disagree a little)	−22.78 (3.70)	−22.90 (3.70)	−22.56 (3.67)	−22.28 (3.66)
enjoyreading (disagree a lot)	−28.53 (3.60)	−28.57 (3.61)	−28.65 (3.59)	−28.73 (3.58)
bornUS	18.30 (4.10)	18.00 (4.15)	18.01 (4.17)	17.27 (4.25)
dadbornUS (don't know)	−8.40 (3.43)	−8.26 (3.45)	−8.21 (3.46)	−5.55 (4.31)
dadbornUS (yes)	6.84 (3.13)	6.92 (3.13)	7.04 (3.12)	5.32 (3.49)
home25books	16.12 (2.22)	15.99 (2.20)	16.01 (2.21)	15.93 (2.22)
watchTV5H	−15.15 (2.52)	−15.22 (2.53)	−15.09 (2.54)	−14.93 (2.55)
comgames5H	−16.05 (2.79)	−16.14 (2.80)	−16.02 (2.85)	−15.75 (2.86)
econDisadv (>50%)	−42.15 (5.09)	−42.45 (5.11)	−40.80 (5.05)	−45.68 (6.27)
econDisadv (26–50%)	−14.62 (5.18)	−14.53 (5.23)	−14.00 (5.22)	−17.19 (6.24)
econDisadv (11–25%)	−6.11 (5.93)	−6.27 (5.96)	−5.82 (5.71)	−8.17 (7.77)
mombornUS (don't know)				−5.48 (5.12)
mombornUS (yes)				2.83 (3.51)
schoolLIB				23.82 (9.92)
female × econDisadv (>50%)				8.13 (4.67)
female × econDisadv (26–50%)				5.61 (5.80)
female × econDisadv (11–25%)				3.47 (7.26)
Variance Components		**Estimate (SE)**		
σ^2	3834.80 (117.33)	3804.26 (116.431)	3773.05 (120.57)	3768.94 (119.92)
Intercept τ_{00}	513.84 (76.12)	520.11 (85.64)	712.09 (131.62)	700.66 (133.84)
Slope (comgames5H) τ_{11}		152.87 (91.31)	165.53 (110.43)	162.12 (110.82)
Slope (female) τ_{22}			125.33 (83.38)	114.59 (79.55)
Covariance τ_{01}		−35.50 (72.43)	−82.73 (103.61)	−75.37 (105.34)
Covariance τ_{02}			−201.15 (86.79)	−197.11 (88.31)
Covariance τ_{12}			14.43 (71.42)	2.52 (72.54)
Fit Indexes		**Mean (SD)**		
−2 × log-likelihood	57831.47 (83.13)	57826.33 (83.04)	57814.53 (81.98)	57803.22 (83.10)
AIC	57863.47 (83.13)	57862.33 (83.04)	57856.53 (81.98)	57857.22 (83.10)
BIC	57968.35 (83.13)	57980.32 (83.04)	57994.18 (81.98)	58034.19 (83.10)

Note: Between-imputation standard deviations for the fit indexes are also presented.

(2006, 2007) also showed through simulations that model-based standard errors are particularly sensitive to misspecification due to omitted variables. We used empirical standard errors in our analysis.

In addition to statistical tests for individual regression coefficients using *t*-test statistics, we compared the models by means of likelihood values, Akaike Information Criterion (AIC = –2 log-likelihood + 2 × the number of parameters), and Bayesian Information Criterion (BIC = –2 log-likelihood + 2 × the number of parameters × log(*N*)). We obtained these model fit indexes for each model in the 25 datasets and reported the means and standard deviations of the three indexes across the imputed datasets. As only the likelihood function varies across the imputed datasets while the number of parameters and sample size stay constant, the standard deviations are the same for the three fit indexes for each model.

With respect to evaluating the random effects, standard errors of variance components are known to be unreliable in multilevel models, and it is advised not to use standard errors to evaluate the significance of variance components. Some multilevel software (e.g., package lme4 in R) do not provide standard errors for variance components. Instead, a likelihood ratio test statistic based on the full and reduced models is used to compare two models with and without a random effect. Moreover, the likelihood ratio test for comparing fixed effect parameters needs to be modified to test the significance of a random effect, as testing variance components deals with a null hypothesis at the boundary value (e.g., H_0: $\tau_{11} = 0$) and a directional alternative hypothesis (e.g., H_1: $\tau_{11} > 0$).

A simple modification of the χ^2 critical value can account for the otherwise overly conservative likelihood ratio test for variance components. Specifically, results from simulation studies (Self and Liang 1987; Stram and Lee 1994, 1995) have shown that when testing H_0: $\tau_{00} = 0$ versus H_1: $\tau_{00} > 0$, the likelihood ratio test statistic asymptotically follows a distribution of $\frac{1}{2}\chi_1^2$ rather than χ_1^2. The asymptotic distribution of the likelihood ratio test statistic for testing an additional random effect (e.g., random slope) is shown to be a mixture of χ_1^2 and χ_2^2 distributions. This implies that for testing the significance of the first random slope by comparing –2log-likelihood of Model 1 and Model 2 (which has two more parameters, τ_{11} and τ_{01}, than Model 1) at the type I error rate of 0.05, the proper critical value is $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2 = \frac{1}{2}(3.84 + 5.99) = 4.92$, which is smaller than the standard likelihood ratio test critical value with two degrees of freedom. For testing the second random slope by comparing Model 2 and Model 3 (with three more parameters τ_{22}, τ_{02}, and τ_{12}), we can use the critical value of $\frac{1}{2}\chi_2^2 + \frac{1}{2}\chi_3^2 = \frac{1}{2}(5.99 + 7.81) = 6.90$. This downward adjustment in the critical values using mixture distributions makes the likelihood ratio test for random effects more accurate.

Table 18.4 shows that AIC and BIC disagree in regard to the best fitting model for this example. While Model 3 (random slopes of `comgames5H` and female) has the smallest AIC, Model 1 (random intercept only) has the smallest BIC. At the same time, the four models are nested within each other and we can also directly compare them using the likelihood ratio test. To compare

nested models with different fixed effects, such as Model 3 versus Model 4, it is important to estimate models using the full information maximum likelihood (FIML) method, not the residual maximum likelihood (REML) method. Models 1 through 3 are different only in random effects, and thus both FIML and REML methods can be used in comparing them via likelihood values. All models in this chapter were estimated by the FIML method.

In the comparison of Model 1 versus Model 2 with the random slope of `comgames5H`, the difference in –2log-likelihood is 5.14, which is greater than the critical value of $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2 = 4.92$, suggesting the effects of comgames5H vary across schools. In the comparison of Model 2 versus Model 3 with the additional random slope of female, the difference in –2log-likelihood is 11.80 and is greater than the critical value of $\frac{1}{2}\chi^2_2 + \frac{1}{2}\chi^2_3 = 6.90$, suggesting the effect of gender is also different across schools. Finally, in comparing Model 3 to Model 4 with the additional predictors of `mombornUS`, `schoolLIB`, and the interaction `female` × `econDisadv`, the standard critical value for the likelihood ratio test was used because the difference between the two models involves only fixed effects. The difference in –2log-likelihood is 11.31 between Model 3 and Model 4 with six additional parameters, and this difference is smaller than the critical value of $\chi^2_6 = 14.07$. Therefore, based on the series of likelihood ratio tests, we chose Model 3 as the best-fitting model in our analysis.

Although the summary statistics in Table 18.1 suggest that whether a student's mother was born in the United States is important with respect to that student's reading performance, the effect was not significant in Model 4. We suspect this is due to the strong association between `dadbornUS` and `mombornUS`. Among 5038 students who answered each question as "yes," "no," or "don't know," 84% chose the same category for the two questions. Therefore, there is little unique information in `mombornUS` when `dadbornUS` is already in the model, and the two highly correlated variables lead to multicollinearity in estimation. For a similar reason, we speculate whether the school has a library or not seems important based on the mean reading performance, but the effect of `schoolLIB` is not significant when *econDisadv* is already in the model. Those schools without a library all had high percentages of economically disadvantaged students. We also examined a number of interactions, including the cross-level interaction `female` × `econDisadv` but did not observe a significant interaction effect.

R^2-Type Measures for Explained Variance

In standard regression models, the R^2 index represents the percentage of variance in the outcome variable that is accounted for by the explanatory variables. In multilevel models, there are complications with the use of R^2 to assess prediction. One complication is that there are models at multiple levels that represent different sources of variance in the data; within-cluster and between-cluster models. Therefore, different R^2-type measures can be calculated at each level, often referred to as R^2_1 for a level-1 model and R^2_2 for a level-2 model.

In single-level linear regression, R^2 has a number of different interpretations, including the squared correlation between observed and predicted values of the response, the proportional decrease in variance of the response given the predictor variables, and the proportional reduction of prediction error variance. The definition of the proportional reduction in explained variance is problematic in multilevel models because it is theoretically and empirically possible that a residual variance at some level of the model will increase after adding an explanatory variable, and thus interpreting R^2 as a simple percentage of variance accounted for (analogous to single-level regression analysis) is not recommended.

Snijders and Bosker (1994, 1999) proposed some alternative ways to define explained variation that provide less problematic interpretations based on the proportional decrease in prediction error variance (i.e., the mean squared error of prediction). Specifically, consider a two-level random intercept model with variance components at levels 1 and 2, denoted σ^2 and τ_{00}, respectively. We will obtain different estimates of the two variance components under an empty unconditional model (Equation 18.7) and a conditional model with explanatory variables. Denote the estimates from the unconditional model as $(\hat{\sigma}^2$ and $\hat{\tau}_{00})_{\text{unconditional}}$ and the conditional model as $(\hat{\sigma}^2$ and $\hat{\tau}_{00})_{\text{conditional}}$. Snijders and Bosker (1999) defined their alternative R^2 indices as follows. The first index is the proportional reduction of error for predicting a level-1 outcome, and can be computed as

$$R_1^2 = 1 - \frac{(\hat{\sigma}^2 + \hat{\tau}_{00})_{\text{conditional}}}{(\hat{\sigma}^2 + \hat{\tau}_{00})_{\text{unconditional}}}. \tag{18.11}$$

The second index is the proportional reduction of error for predicting a group mean, which can be computed as

$$R_2^2 = 1 - \frac{(\hat{\sigma}^2/\tilde{n} + \hat{\tau}_{00})_{\text{conditional}}}{(\hat{\sigma}^2/\tilde{n} + \hat{\tau}_{00})_{\text{unconditional}}}, \tag{18.12}$$

where $\tilde{n}$ is the expected number of level-1 units per level-2 unit. When the number of level-1 units, n_j, varies greatly across level-2 units, $\tilde{n}$ can be substituted by the harmonic mean

$$\tilde{n} = \frac{M}{\sum_{j=1}^{M}(1/n_j)}.$$

Relative to the arithmetic mean, the harmonic mean gives less weight to clusters with much larger sample sizes and thus is more appropriate as the expected number of level-1 units per level-2 unit in the population. In the

PIRLS 2006 example, the harmonic mean for the number of students per school is 23.39 (as opposed to the arithmetic mean of 28.36).

Note that Equations 18.11 and 18.12 represent R_1^2 and R_2^2 for a random intercept model as the conditional model. However, our final model (Model 3 in Table 18.4) consists of two random slopes. Although R_1^2 and R_2^2 can be calculated for random slope models, the process is considerably more complicated than for random intercept models, because level-2 variances are not constant but functions of explanatory variables. Recchia (2010) developed a SAS macro to calculate R_1^2 and R_2^2 for multilevel models. Alternatively, Snijders and Bosker (1999) suggested using approximate R_1^2 and R_2^2 values based on a random intercept model, because the values of R_1^2 and R_2^2 for random intercept and random slope models are similar when they have the same fixed effects specification.

Following the suggestion by Snijders and Bosker (1999), we calculated approximate R_1^2 and R_2^2 in our example by comparing the empty unconditional model to the random intercept conditional model, Model 1 in Table 18.4, which is identical to our final model except for the two random slopes. We obtained R_1^2 and R_2^2 for each of the 25 datasets and the values are reasonably stable across the imputed datasets. At level 1, R_1^2 varies from .206 to .225, with a mean of .214 (SD = .005), implying that there is a 21.4% reduction in the mean square prediction error within schools. At level 2, R_2^2 varies from .509 to .544, with a mean of .525 (SD = .009), implying that we reduced predictive error 52.5% for the school means by including explanatory variables.

Incorporating Design Weights

Often with complex, large-scale surveys, the probability of selection of clusters and observations within cluster are unequal. Typically, weights are given by the organization conducting the survey that reflect the design of the study. When weights are excluded from the modeling, parameter estimates may be biased but efficient; including weights leads to unbiased estimates but less efficient ones. The decision to include design weights or not should be based on whether the weights are likely to have an impact or are *informative*, and whether the probability of selection (the sampling model) is related to the probability model for the data (i.e., the HLM). To determine whether weights are informative, we need to compute them and examine their distribution.

We adopted recommendations by Rutkowski et al. (2010) for computing weights in surveys such as the PIRLS. Weights should be computed for each level of sampling. Besides having unequal selection probability, all of the selected units may not respond. The weights for students (level 1 or secondary units) will be computed as the product of student and class weights:

$$w_{1ij\ell} = \underbrace{\left(WF_{ij\ell} \quad WA_{ij\ell}\right)}_{\text{student } i}\underbrace{\left(WF_{j\ell} \quad WA_{j\ell}\right)}_{\text{class } \ell}, \tag{18.13}$$

where $WF_{ij\ell}$ and $WF_{j\ell}$ are the inverses of the probabilities for the selection of students and classes from school j, respectively, and $WA_{ij\ell}$ and $WA_{j\ell}$ are the weight adjustments for nonresponse for student i and class ℓ from school j, respectively. The weight adjustments are for those students and classes that were selected but did not choose to participate (Rutkowski et al. 2010). The school weights used are

$$w_{2j} = WF_j \quad WA_j, \tag{18.14}$$

where WF_j is the inverse of the probability of selecting school j, and WA_j is the weight adjustment for school j.

One way to assess the informativeness of the weights is to examine their distribution. If the selection probabilities are equal and all selected units respond, then the weights would all be equal and their variance would be zero. In this case, the sampling would be the same as simple random sampling and the weights could be set to one. In the PIRLS 2006 data, the mean of the weights for schools equals 305.42 and their standard deviation is 218.64, which suggests that the level 2 weights are informative. For the student level, the means of students within schools are mostly equal to one and 74% of the schools have standard deviations that equal zero. Of the remaining schools, 21% have standard deviations that are less than 0.05 and the other 5% have standard deviations less than 0.13. These standard deviations suggest that the level-1 weights are not informative and will likely have no impact on the analysis. Additional methods to assess the informativeness of weights are given in Chapter 21 by Anderson, Kim, and Keller.

The weights are typically scaled according to one of two methods that were discussed in Pfeffermann et al. (1998). The method used here is to scale the weights so that their sum equals the sample size; that is, $\sum_j w_{2j} = 183$ and $\Sigma_i \Sigma_\ell w_{1ij\ell} = n_j$.

Weights are incorporated into the model during estimation. The log-likelihoods for the lowest-level units (i.e., students) are multiplied by their respective weights and summed over the values within a cluster. Cluster-specific values are found by integrating out the random effects (i.e., U_{pj}'s). The cluster-specific values are summed after being multiplied by their respective weights w_{2j} to yield a function of all the data (Grilli and Pratesi 2004; Rabe-Hesketh and Skronkal 2006). The parameters that maximize the function are maximum pseudolikelihood estimates. They are not maximum likelihood estimates because the log-likelihoods are multiplied by weights. For more details, see Chapter 21 by Anderson, Kim, and Keller (see also Grilli and Pratesi 2004; Rabe-Hesketh and Skronkal 2006).

To illustrate the effect of including weights, we fit Model 3 (our best model without weights) with weights. The results are summarized in Table 18.5. Since we are using pseudolikelihood estimation, the sandwich or robust estimates of the standard errors are presented. The lower efficiency resulting from including weights is evident by comparing the standard errors of the

TABLE 18.5

Model 3 with and without Weights

	Model 3 without Weights			Model 3 with Weights		
Fixed Effects	**Coef**	**SE**	***t***	**Coef**	**SE**	***t***
Intercept	544.19	6.15	88.51	535.07	7.09	75.53
female	2.08	2.19	0.95	2.89	2.84	1.02
enjoyreading (agree a little)	−9.18	2.29	−4.01	−9.71	3.21	−3.02
enjoyreading (disagree a little)	−22.56	3.67	−6.15	−29.15	4.76	−6.13
enjoyreading (disagree a lot)	−28.65	3.59	−7.97	−29.23	4.37	−6.69
bornUS	18.01	4.17	4.32	21.00	5.29	3.97
dadbornUS (don't know)	−8.21	3.46	−2.38	−8.53	4.17	−2.04
dadbornUS (yes)	7.04	3.12	2.26	7.75	3.53	2.19
home25books	16.01	2.21	7.25	17.49	2.59	6.75
watchTV5H	−15.09	2.54	−5.94	−13.30	2.86	−4.66
comgames5H	−16.02	2.85	−5.61	−16.02	3.08	−5.20
econDisadv (>50%)	−40.80	5.05	−8.08	−37.85	6.80	−5.57
econDisadv (26–50%)	−14.00	5.22	−2.68	−4.86	6.64	−0.73
econDisadv (11–25%)	−5.82	5.71	−1.02	2.10	6.35	0.33
Variance Components	**Estimate**	**SE**		**Estimate**	**SE**	
σ^2	3773.05	120.57		3719.26	136.67	
Intercept τ_{00}	712.09	131.62		771.03	225.17	
Slope (comgames5H) τ_{11}	165.53	110.43		151.06	121.20	
Slope (female) τ_{22}	125.33	83.38		166.10	112.42	
Covariance τ_{01}	−82.73	103.61		−71.53	116.70	
Covariance τ_{02}	−201.15	86.79		−227.32	129.01	
Covariance τ_{12}	14.43	71.42		45.86	83.41	

Note: Model coefficients, standard errors, and fit indexes, averaged across 25 (5 plausible outcome values × 5 independent variable imputations) datasets are shown.

parameter estimates. The estimates of the standard errors with the weights are larger than those without weights. One of the school-level variable effects, `econDisadv (26–50%)` has very different regression coefficient estimates (–14.00 vs. –4.86) and is not significant in the model with weights.

Model Diagnostics

Before turning to interpretation, we consider an often neglected part of multilevel modeling: investigating whether there are potential violations of models assumptions or systematic misfit of the model to data. Snijders and Bosker (2012) cover diagnostics for multilevel models in great detail. We will restrict our attention to the examination of residuals and the assumption of normally distributed random effects and illustrate some of the possible methods using Model 3 without weights.

Figure 18.6 includes various model diagnostic plots. In the top row of plots, we examine the cluster-specific or conditional Pearson residuals. These residuals are based on predictions that include both the fixed effects from Table 18.4 and empirical Bayes estimates of the random effects (i.e., U_{0j}, U_{1j}, and U_{2j}). The analogous plots for marginal predictions (i.e., those that excluded estimates of the random effects) were similar to those in Figure 18.6. The plots for each imputed dataset were very similar to each other; therefore, we averaged over the 25 imputed datasets and present a single set of plots.

In Figure 18.6a, conditional Pearson residuals are plotted against predicted reading scores. If a model is a reasonable one, we should see a random collection of points, approximately equal residual variances across levels of the predicted variables, and no apparent patterns. This is the case in our example. A pattern will often be seen when a response variable is subject to either a floor or ceiling effect (i.e., scores were bounded either from below or above). In such a case, we would likely see a linear relationship. The analogous plot for the marginal effects shows a pattern in that there is a maximum pos-

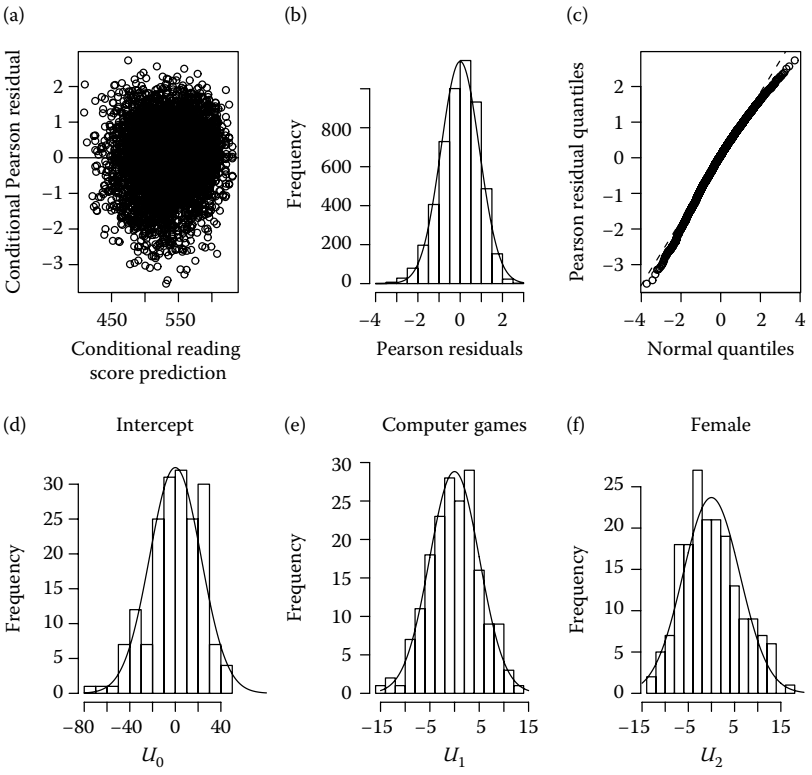

FIGURE 18.6
Various model diagnostics for Model 3 that include empirical Bayes estimates for random effects.

sible value for the predicted reading scores (i.e., a vertical line of points at predicted reading scores equal to around 575).

Figures 18.6b and 18.6c are designed to examine the normality assumption. Figure 18.6b is a histogram of the conditional Pearson residuals with the normal distribution overlaid and Figure 18.6c is a Q–Q plot. From the Q–Q plot, it appears that the tails of the distribution of residuals are a bit heavy, but overall normality seems reasonable.

Figures 18.6d through 18.6f are histograms of the estimated random effects with normal distributions overlaid. If the assumption that the random effects are multivariate normal is satisfied, then the distribution of estimated U_{pj}'s should be normal. Although the estimated U_{pj}'s look roughly normal, this only means that the assumption is tenable. It is possible that the true underlying distribution of random effects is not normal but the estimates of the U_{pj}'s may appear normal (Verbeke and Molenberghs 2000). If the distribution of $\hat{U}_{pj}$ is not normal, then the assumption is not valid.

Interpretation of Results

The final model, Model 3, has 14 regression coefficients and seven random effect parameters (variance σ^2 at level 1 and a 3×3 variance–covariance matrix at level 2) with one random intercept and two random slopes. The intercept of 544.21 is the expected reading proficiency score when all predictors are zero, implying male students who enjoy reading a lot, were not born in the United States, have a father who is also not born in the United States, have less than 25 books at home, watch TV or play computer games less than 5 h per day, and attend a school where the percentage of economically disadvantaged students is less than 10%. Female students who were born in the United States, have a father born in the United States, and/or have more books at home on average would have better reading performance. On the other hand, those who do not enjoy reading as much, watch TV or play computer games 5 h or more per day, and/or attend a school with a higher percentage of economically disadvantaged students on average would be expected to have lower reading proficiency scores.

The random intercept represents heterogeneity across schools in terms of their residual reading performance even after accounting for explainable differences by the fixed effects. Model 3 has two random slopes, both corresponding to predictors that are binary variables. The random slopes suggest that the effects of gender as well as excessive hours on computer games vary across schools. Since the two variables are both binary, we can consider four subgroups whose variances are different. For male students who spend less than 5 h on computer games (i.e., `female` = 0, `comgames5H` = 0), the variance of reading proficiency is estimated as $\hat{\tau}_{00} + \hat{\sigma}^2 = 4485.88$. For male students who spend 5 or more hours on computer games, the variance is $\hat{\tau}_{00} + 2\hat{\tau}_{02} + \hat{\tau}_{22} + \hat{\sigma}^2 = 4485.65$. For female students who spend less than 5 h on computer games, the variance is $\hat{\tau}_{00} + 2\hat{\tau}_{01} + \hat{\tau}_{11} + \hat{\sigma}^2 = 4208.72$.

Finally, for female students who spend 5 or more hours on computer games, the variance is estimated as $\hat{\tau}_{00} + 2\hat{\tau}_{01} + \hat{\tau}_{11} + 2\hat{\tau}_{02} + 2\hat{\tau}_{12} + \hat{\tau}_{22} + \hat{\sigma}^2 = 4235.31$. Therefore, the heterogeneity is the greatest among males who spend less than 5 h on computer games. Females who spend less than 5 h on computer games (the highest performance group among the four) are the most homogeneous group of students.

Discussion

Other Multilevel Models

Although the two-level models in this chapter where students are nested within schools are common in large-scale assessment data, other multilevel models are also widely used and are gaining greater interest. Three in particular are worth considering in the analysis of international assessment data: three-level models, repeated measures models, and nonlinear models for discrete outcomes. For three-level models, countries or other primary sampling units, such as states, districts, or counties, can be entered as the highest level. Depending on the number of level-3 units, distributional assumptions, and omitted factors, the highest level can be treated as *fixed* or *random*. There exists an extensive literature on this topic in econometrics (Hausman 1978; Hausman and Taylor 1981) and more recently in psychometrics (Kim and Frees, 2006, 2007). For example, we can extend the current analysis by considering another country or two and modeling results for multiple countries simultaneously. In such analyses, it should be clarified that the meanings of the variables and levels are comparable across countries. However, many variables in such analyses need to be analyzed in context and their meanings may not be exchangeable across countries (e.g., born in the country, parents born in the country, first language, economically disadvantaged, and eligible for free or reduced-price meal).

Another important type of multilevel analysis is repeated measures analysis, which includes but is not limited to longitudinal data analysis. As large-scale assessment data often consist of scores on multiple subjects or multiple aspects of a subject (e.g., reading achievement/behavior/attitude or different domains of mathematics), several related outcomes can be available for each student. In a longitudinal design, students' performances and attitudes may be recorded over time to study intraindividual change as well as interindividual differences. When multiple dependent variables are nested within students, these repeated measures can be treated as level-1 units in multilevel models. In such models, students are often entered as level-2 units, and schools are considered as level-3 units. Dependency among the level-1 units is usually very high in repeated measures models because the values belong to the same students. A flexible variance–covariance structure may be

needed to account for the complex dependency among repeated measures. Multivariate multilevel models or latent growth curve models are commonly used for multiple subjects with longitudinal assessment data.

Finally, large-scale assessment data almost always include survey questionnaires which require respondents to select one of a given set of response options, thus yielding discrete outcomes. Discrete outcomes data call for the use of alternative models designed to handle their special nature. In the context of large-scale assessment studies that use complex sampling designs, complications arise because the data are typically clustered data and the probability of selection is not equal for all units. Chapter 21 by Anderson, Kim, and Keller in this volume is devoted to this issue. Multilevel models for discrete outcomes are also referred to as *generalized linear mixed models*.

Current and Future Research in Multilevel Analysis

There was a time when multilevel analysis was considered a highly advanced statistical method and only used by a limited number of trained methodologists. In the recent years, however, multilevel modeling has become available to many substantive researchers and practitioners, in our opinion largely owing to the publication of accessible textbooks, the availability of user-friendly software, and regularly offered courses and workshops. Several textbooks appear to be widely used across the disciplines, including Raudenbush and Bryk (2002), Snijders and Bosker (2012), Hox (2010), and Goldstein (2010).

Multilevel analysis has proven its importance in the social sciences, and the methodology has great potential for advancing the design and utilization of large-scale assessment data. In educational research, despite immense effort and investment, results from educational effectiveness studies remain inconsistent (Ehrenberg et al. 2001; Goldhaber and Brewer 1997; Ludwig and Bassi 1999). There are mixed findings concerning which district, school, teacher, neighborhood, family, and student variables make significant differences on educational outcomes such as improving students' academic achievement or reducing dropout rates, let alone the size of their effects. More recently, the Programme for International Student Assessment (PISA) and other international assessments have found counterintuitive relationships between achievement scores and attitudes at the country level. For example, in the PISA 2006 data, students with higher science achievement scores indicated more positive attitudes toward science as expected, when the data are analyzed within countries. However, the relationship between achievement and attitudes was reversed at the country level and it was found that high science achievement was associated with more negative attitudes based on between-country correlations (Lu and Bolt, 2012).

Puzzling or inconsistent findings in educational research are due to multiple factors including inherent difficulties in isolating the effects of variables

involved in complex educational processes, omitted confounding factors, nonrandom missing data and attrition, selection bias related to school and program choice, and different dynamics among variables across clusters (e.g., countries or schools). Yet, it is often infeasible to design comprehensive experimental studies in school settings for apparent ethical and/or practical reasons. It is also impractical to expect no missing data or missing completely at random, or to collect all relevant variables to account for selection mechanisms and cultural differences.

Consequently, there is always a danger in education as well as other areas in the social sciences that decision making and policy implementation can be based on inaccurate information or in some cases misleading findings. Although we attempted to be careful in our model specification and diagnostics, the purpose of our data analysis is for demonstrating the application of multilevel analysis to large-scale assessment data rather than making practical recommendations or drawing substantive conclusions, as further investigation is necessary to provide stronger claims of causal inference between predictors and reading outcomes. Chapter 22 by Robinson in this handbook deals with making causal inferences using large-scale assessment data.

Despite many advances during the past several decades, there remain important theoretical and technical issues to be addressed in multilevel analysis. For example, although there exists a large body of literature with regard to imputation approaches for missing data, there is a lack of methodological development for effectively imputing missing values while preserving yet utilizing the multilevel structure, especially for discrete outcomes (for continuous outcomes, see Chapter 20 by Shin in this volume). Also, whereas various matching designs and strategies have been proposed in studies of educational effectiveness with observational data, little has been done on matching nonequivalent groups in the context of multilevel data. This delayed progress is largely due to the complexity of adapting statistical techniques for imputation and matching strategies into the multilevel framework, as the necessary modifications are not straightforward (Steiner et al. 2012). Nonetheless, considering the needs of these techniques and active research on these topics, we expect promising methodological developments as well as increasing applications of multilevel analysis in the coming years.

Acknowledgment

This work was supported in part by grant R305D120005 from the Institute of Education Sciences, the U.S. Department of Education awarded to the first author.

References

Allison, P. D. 2002. *Missing Data.* Sage, Newbury Park, CA.

Amer, S. R. 2009. Neural network imputation in complex survey design. *International Journal of Electrical and Electronics Engineering*, 3:52–57.

Diggle, P. J., Liang, J.-Y., and Zeger, S. L. 1994. *Analysis of Longitudinal Data*. Oxford University Press, NY.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., and Willms, J. D. 2001. Class size and student achievement. *Psychological Science in the Public Interest*, 2:1–30.

Enders, C. K. 2010. *Applied Missing Data Analysis.* Guilford Press, New York, NY.

Goldhaber, D. D. and Brewer, D. J. 1997. Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources*, 32, 505–523.

Goldstein, H. 2010. *Multilevel Statistical Models.* 4th Edition, John Wiley and Sons, Ltd, Chichester, UK.

Grilli, L. and Pratesi, M. 2004. Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology,* 30:93–103.

Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica*, 46:1251–1272.

Hausman, J. A. and Taylor, W. E. 1981. Panel data and unobservable individual effects. *Econometrica*, 49:1377–1398.

Hox, J. J. 2010. *Multilevel Analysis: Techniques and Applications.* 2nd edition, Routledge, New York, NY.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–231.

Kim, J.-S. and Frees, E. W. 2006. Omitted variables in multilevel models. *Psychometrika*, 71:659–690.

Kim, J.-S. and Frees, E. W. 2007. Multilevel modeling with correlated effects. *Psychometrika*, 72:505–533.

Kim, J.-S. and Swoboda, C. 2012. Strategies for imputing missing values in hierarchical data: Multilevel multiple imputation. Paper presented at the *Annual Meeting of American Educational Research Association,* Vancouver, CA.

Kish, L. 1965. *Survey Sampling.* John Wiley & Sons, Inc., New York, NY.

Little, R. J. A. and Rubin, D. B. 2002. 2nd edition, *Statistical Analysis with Missing Data.* John Wiley, New York.

Lu, Y. and Bolt, D. 2012. Application of a MIRT model to address response style differences in cross-national assessment of educational attitudes. Paper presented at the *Annual Meeting of National Council on Measurement in Education,* Vancouver, CA.

Ludwig, J. and Bassi, L. 1999. The puzzling case of school resources and student achievement. *Educational Evaluation and Policy Analysis*, 21:385–403.

Martin, M. O., Mullis, I. V. S., and Kennedy, A. M. 2007. PIRLS2006 technical report. Technical report, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mislevy, R. J. 1991. Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56:177–196.

Muthén, L. K. and Muthén, B. 1998–2010. *Mplus User's Guide.* 6th edition, Muthén & Muthén, Los Angeles, CA.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., and Rasbash, J. 1998. Weighting for unequal selection probabilities in multilevel models (with discussion). *Journal of the Royal Statistical Society, Series B*, 60:23–56.
Pinheiro, J. and Bates, D. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer Verlag, New York, NY.
Rabe-Hesketh, S. and Skronkal, A. 2006. Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series A*, 169:805–827.
Raudenbush, S. W. and Bryk, A. S. 2002. *Hierarchical Linear Models.* Sage, Thousand Oaks, CA.
Recchia, A. 2010. R-squared measures for two-level hierarchical linear models using SAS. *Journal of Statistical Software*, 32:1–9.
Reiter, J., Raghunathan, T., and Kinney, S. 2006. The importance of modeling smapling design in multiple imputation for missing data. *Survey Methodology*, 32:143–149.
Rubin, D. B. 1976. Inference and missing data. *Biometrika*, 63:581–592.
Rutkowski, L., Gonzalez, E., Joncas, M., and von Davier, M. 2010. International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39:142–151.
Ryoo, J. H. 2011. Model selection with the linear mixed model for longitudinal data. *Multivariate Behavioral Research*, 46:598–624.
Schafer, J. L. 1997. *Analysis of Incomplete Missing Data.* Chapman & Hall, London.
Self, S. G. and Liang, K. Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610.
Shin, Y. and Raudenbush, S. W. 2007. Just-identified versus overidentified two-level hierarchical linear models with missing data. *Biometrics*, 63:1262–1268.
Shin, Y. and Raudenbush, S. W. 2010. A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, 35:26–53.
Snijders, T. A. B. and Bosker, R. J. 1994. Modeled variance in two-level models. *Sociological Methods Research*, 22:342–363.
Snijders, T. A. B. and Bosker, R. J. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 1st edition, Sage, London.
Snijders, T. A. B. and Bosker, R. J. 2012. *Multilevel Analysis*. 2nd edition, Sage, Thousand Oaks, CA.
Steiner, P., Kim, J.-S., and Thoemmes, F. 2012. Matching strategies for observational multilevel data. In *Joint Statistical Meeting Proceedings, Social Statistics Section*, San Diego, CA. American Statistical Association.
Stram, D. O. and Lee, J. W. 1994. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50:1171–1177.
Stram, D. O. and Lee, J. W. 1995. Correction to: Variance components testing in the longitudinal mixed effects model. *Biometrics*, 51:1196.
Swoboda, C. M. 2011. *A New Method for Multilevel Multiple Imputation.* Unpublished doctoral dissertation, University of Wisconsin–Madison.
Van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242.
Van Buuren, S. 2011. Multiple imputation of multilevel data. In Hox, J. and Roberts, K., eds, *Handbook of Advanced Multilevel Analysis.* Taylor & Francis, New York, NY.
Van Buuren, S. 2012. *Flexible Imputation of Missing Data.* Chapman Hall/CRC, Boca Raton, FL.

Verbeke, G. and Molenberghs, G. 2000. *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York, NY.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838.

Wingersky, M., Kaplan, K. B., and Beaton, A. E. 1987. Joint estimation procedures. In Beaton, A. E., ed., *Implementing the New Design: The NAEP 1983–1984 Technical Report*, pages 285–292. Educational Testing Services, Princeton, NJ.