CHAPTER 11

REPORTING RESULTS FROM MULTILEVEL ANALYSES

John M. Ferron, Kristin Y. Hogarty, Robert F. Dedrick, Melinda R. Hess, John D. Niles, and Jeffrey D. Kromrey

OVERVIEW

In recent years there have been dramatic advances in the field of multilevel modeling. These advances, coupled with the addition of new features and options for statistical output in current multilevel software programs such as HLM, SAS PROC MIXED and MLwiN (Roberts & McLeod, 2008), have posed challenges to researchers attempting to communicate the results of these models to audiences with varying levels of statistical and research expertise. Unlike the area of structural equation modeling for which recommendations and guidelines have been presented to enhance the communication value of the results (Boomsma, 2000; Hoyle & Panter, 1995; McDonald & Ho, 2002; Raykov, Tomer, & Nesselroade, 1991), the field of multilevel modeling has provided few guidelines for conveying research findings.

A recent review of the reporting practices of articles from education and related journals (Ferron et al., 2006) supports the need for guidelines. In this review, Ferron et al. analyzed 98 multilevel modeling articles from 19

Multilevel Modeling of Educational Data, pages 391–426

journals with an educational or related focus (e.g., *American Educational Research Journal*, *Child Development*) to determine how authors addressed the following issues: (a) model development and specification; (b) data considerations including distributional assumptions, outliers, measurement error, power, and missing data; (c) estimation procedures; and (d) hypothesis testing and statistical inference including inferences about variance parameters and fixed effects. Overall, the results indicated that in many cases not enough information was presented to allow readers to fully interpret the results or replicate the analyses. Additionally, the use of different terminology such as variance estimates versus random effects and the use of terms (e.g., standardized) left undefined by authors present challenges to understanding the results of multilevel analyses.

This chapter offers suggestions for *what* to present when reporting results of multilevel analyses and options for *how* to present these results using text, tables, and figures. The assumption underlying these guidelines is that the organization and presentation of multilevel models and their results have the potential to critically impact the utility and understanding of multilevel research. These guidelines reflect the realities that are present in most current publishing opportunities (e.g., space restrictions in paper journals); although, with the advent of online publication, issues such as length of articles and the number of illustrations and tables may be less critical. Because a single chapter cannot include guidelines for every type of multilevel model, it is important to clarify that the focus of this chapter is primarily on what might be termed "traditional" multilevel models. This chapter considers linear models of continuous outcomes where the random effects are assumed normally distributed. This allows consideration of two-level applications where individuals are nested in contexts, such as students nested in schools, and applications where observations are nested within individuals, such as growth curve models. Models in which the outcome is represented by binary, count, or ordinal data are not considered (see O'Connell, Goldstein, Rogers, & Peng, 2008, and Raudenbush & Bryk, 2002, for discussions of these types of applications), nor are multilevel structural equation models (SEM; Muthén & Muthén, 1998–2004) or multilevel item response models (Kamata, 2001; Kamata, Bauer, & Miyazaki, 2008).

As with any set of guidelines, flexibility is needed to take into account the requirements of the publication outlet and the intended audience. For example, a novice reader of multilevel studies may be able to interpret graphs of a growth curve model more easily than a complex equation with coefficients. Reactions to these guidelines by journal editors and researchers experienced in multilevel modeling can be used to further refine criteria for reporting multilevel results.

The organization of this chapter parallels the sections and subsections of many journal articles: (a) research questions, (b) literature review, (c) meth-

od, (d) results, and (e) discussion. Each section of the journal article can play a role in enhancing the interpretability and value of results from multilevel studies. The first section of this chapter presents some common research questions addressed through the use of two-level models. Following this section, we discuss how the literature review might be used to provide a rationale for multilevel analyses, including advantages and disadvantages of this approach. The method section consists of five subsections and offers suggestions for communicating information about: (a) participants, including the number at each level of analysis, sampling procedures, and missing data; (b) type and limitations of the research design; (c) variables, including how the variables were coded and procedures used to address measurement quality; (d) models, including the use of equations for model specification, the centering of predictors, the process for defining the model, and the approach used to evaluate model integrity; and (e) estimation and inference, including technical details of the algorithms for parameter estimation and approaches used for making inferences about variance parameters, fixed effects, and level-one coefficients. The results section consists of two subsections and offers guidelines for presenting: (a) preliminary results on data quality and (b) results directly tied to the research questions. The discussion section presents the core elements that should be part of the discussion in any research study and identifies some elements that are unique to multilevel modeling. Finally, a list of questions that generally should be answerable by the reader of a well-written report of a multilevel modeling application is provided in the form of a checklist. This checklist summarizes the guidelines and suggestions presented in this chapter.

RESEARCH QUESTIONS

As a starting point for communicating the purpose of the study and the appropriateness of using a multilevel approach, the researcher needs to clearly state the questions under investigation. Once these questions have been stated, the statistical models that are aligned with these questions, along with their corresponding assumptions, can be specified. Results linked to these models and ultimately to the research questions then can be presented.

There are a variety of multilevel designs focusing on different types of research questions. With these different research questions come correspondingly different types of results, including preliminary results checking assumptions and those focused directly on the research questions, as well as multiple formats for presenting results.

For example, multilevel designs in which individuals are measured within some larger unit, such as a classroom, often address questions related to

how much of the variability in an outcome is associated with within- and between-group differences, as well as the extent to which various within- and between-group factors account for this variability. In contrast, multilevel designs in which individuals are measured repeatedly over time often address questions related to the form of change (e.g., linear, nonlinear), variation in growth parameters (e.g., intercept and slope), and factors associated with the variation in the growth parameters (e.g., gender). Table 11.1 presents examples of some research questions addressed in two-level multilevel studies and the types of data structures associated with these questions.

LITERATURE REVIEW

In research reports of multilevel analyses, the literature review should describe how the multilevel nature of the research problem under investigation has been addressed in the past. For example, has past research dealt with the unit of analysis issue by ignoring the independence assumption or by aggregating nested data within units? To provide a connection with the current application of multilevel modeling, relevant methodological issues addressed in prior research should be discussed. Through this discussion, the rationale of using a multilevel approach to address the specific questions under investigation can be provided along with the advantages and disadvantages of the multilevel approach. Controversies that are being discussed in the multilevel literature that are relevant to the current investigation can be presented (e.g., use of pseudo R^2 values, use of Akaike's Information Criterion [AIC] and the Bayes Information Criterion [BIC] for model selection). The author of the literature review also can clarify whether the current multilevel application is a replication of a previous study, an extension of prior research, or a new line of inquiry.

The literature review also should foreshadow some of the methodological decisions that are made in the multilevel modeling phase of the study. For example, if some or all of the predictors in the models were selected based on *a priori* considerations (i.e., theory or previous research versus exploratory analyses and tests of significance), the connection with the previous research should be explicit. Similarly, if past research and/or theory were used to justify decisions about other modeling issues such as the variance-covariance structures or centering of predictors, these links need to be made clear. Boote and Beile (2005) have provided additional criteria in developing the literature review for research studies in general; these include providing a rationale for what previous literature to include or exclude and a discussion of the practical and theoretical significance of the research problem.

TABLE 11.1 Examples of Research Questions Addressed by Two-Level Multilevel Designs

General question	Applied question
Individuals nested within units	
1. How much of the variation in an outcome is there within- and between-groups?	How much of the variation in eighth-grade mathematics achievement is within schools? How much is between schools?
2. What is the proportional reduction in the within-group variance when a within-group predictor is added to the model?	What proportion of the within-group variation in eighth-grade mathematics achievement is associated with students' seventh-grade mathematics achievement?
3. What is the relationship between a selected within-group factor and an outcome?	What is the relationship between students' seventh-grade mathematics achievement and students' eighth-grade mathematics achievement?
4. Does the relationship between a selected within-group factor and an outcome vary across the level two units?	Does the relationship between students' seventh-grade mathematics achievement and students' eighth-grade mathematics achievement vary across schools?
5. What is the proportional reduction in the between-group variance in a level two parameter (i.e., intercept) when a between-group predictor is added to the model?	What proportion of the variability in average eighth-grade mathematics achievement is associated with school SES?
6. What is the relationship between a selected between-group factor and an outcome?	What is the effect of a school mathematics instructional program on average school mathematics achievement for eighth graders?
7. To what extent is the relationship between a selected within-group factor and an outcome moderated by a selected between-group factor?	To what extent is the relationship between students' seventh-grade mathematics achievement and students' eighth-grade mathematics achievement moderated by the school's mathematics program?
Observations nested within individuals	
8. Is the functional form of individual change linear, quadratic, or cubic?	What is the functional form of individual change in reading achievement from grades 1 to 5?
9. To what extent do individuals vary in initial status on an outcome?	To what extent do first graders differ in their initial status in reading achievement?
10. To what extent do individuals vary in their rate of change on an outcome?	To what extent does the rate of change in reading achievement of elementary students vary across individuals?
11. What is the relationship between selected individual characteristics and initial status?	To what extent do boys and girls differ in their initial reading achievement?
12. What is the relationship between selected individual characteristics and rate of change?	To what extent do boys and girls differ in their rate of reading achievement change?
13. What is the relationship between individuals' initial status and their rate of change?	What is the relationship between initial reading achievement and the rate of change in reading achievement?

METHOD

Participants

Issues of sample size, sample characteristics, sampling procedures, and power are more complex in multilevel models because of the multiple units of analysis. For a multilevel design in which individuals, such as students, are nested within some larger units, such as schools, simply reporting the total number of students or the total number of schools is not sufficient because the distribution of students across schools can impact model specification and the precision of the parameter estimates. In addition, communicating information about sample sizes requires presenting a rationale for the number of units selected at each level. This rationale may rely on statistical power analyses that include considerations of expected effect sizes, alpha levels, and anticipated attrition and missing data rates (Mok, 1995; Raudenbush, 1997; Raudenbush & Liu, 2000; Spybrook, 2008).

Table 11.2 provides one approach to communicating sample sizes at each level of analysis for a multilevel unbalanced design (unequal sample sizes across units) involving 600 students from 86 schools. If the dataset is large, it may not be practical to provide a table like Table 11.2. In this case, researchers could present descriptive information, including the average number of level-one units per level-two unit, as well as the minimum and maximum number of level-one units. For example, the information in Table 11.2 could be summarized by indicating that there were 86 schools, with the number of students ranging from 1 to 10 per school, with an average of about 7 students per school.

TABLE 11.2 Example of Table for Summarizing Sample Sizes for Students Nested Within Schools in a Two-Level Design

Number of students per school	Number of schools with specified number of students	Cumulative frequency of schools	Cumulative frequency of students
1	11	11	11
2	3	14	17
3	6	20	35
4	2	22	43
5	3	25	58
6	2	27	70
7	10	37	140
8	10	47	220
9	10	57	310
10	29	86	600

TABLE 11.3 Example of Table for Summarizing Sample Sizes for Two-Level Growth Curve Design

Number of time points observed	Number of individuals	% of individuals	Cumulative frequency of individuals
1	10	7.1	10
2	20	14.2	30
3	16	11.4	46
4	21	15.0	67
5	73	52.1	140

For a multilevel design in which individuals are measured repeatedly over time, the distribution of the number of observed time points should be specified. For example, reporting the number of individuals with two data points, three data points, etc. allows readers to evaluate the possibility of identifying nonlinear models and the precision of the parameter estimates from these more complex models (see Table 11.3).

Investigators also should describe the type of sampling procedures that were implemented and discuss if the same sampling procedures were employed at different levels. For example, schools may be selected randomly, and students within those schools may be selected randomly, providing a probability sample at each level. A mixed sampling approach, employing probability sampling at one level and nonprobability methods at another level, also may occur. For example, schools may be selected randomly, but the sample of students at each school may come from teachers who were willing to participate. In studies using existing data, the original database may have been collected using complex sampling methods (Stapleton & Thomas, 2008). In these circumstances, researchers should communicate the type of sampling such as cluster, stratified, or disproportionate, and the implications for the use of sampling weights. For datasets that make available multiple sets of sampling weights, it should be clear what sampling weights were used in the analysis.

Whatever the sampling approach, it is important to describe the final dataset in sufficient detail to allow other researchers to be able to critique or replicate the study. Part of this description should be a discussion of missing data at each level, the degree to which missingness is related to the variables being studied, the method used to handle missing data, and the corresponding consequences, such as introduction of bias and reduction in power (for additional discussion on missing data see Collins, Schafer, & Kam, 2001; Little & Rubin, 1987; Roy & Lin, 2002). Finally, as part of the description of the participants, authors may acknowledge that they complied with all

applicable federal, state, and local regulations and standards related to the ethical treatment of human subjects.

Research Design

The research design and procedures of the study should be reported in sufficient detail to allow readers to replicate the study, to judge whether human subjects were treated ethically, and to critically interpret the results. A challenge in communicating information about the research design is that there is a lack of universally accepted terminology (Maciejewski, Diehr, Smith, & Hebert, 2002). For this reason, researchers need to describe the essential characteristics of the design (e.g., use of experimental manipulation of variables, use of longitudinal data collection) as well as the limitations of the design. Attention should be drawn to how extraneous variables were controlled through methods such as randomization, matching, or statistical adjustments at one or more levels of the analysis. Because these methods can be implemented in a variety of ways, actual implementation procedures need to be detailed.

Description of the design also may involve defining terms that might be used differently across disciplines (for example, omitted variable versus unmeasured confounder). For details on design issues in multilevel studies, see Murray (1998) and Murray, Varnell, and Blitstein (2004). By clearly communicating the design and its limitations, researchers will help readers to judiciously interpret the results of the multilevel anaylsis.

Variables

Clear descriptions and definitions of the variables under investigation are essential in communicating information about the research design. Issues that should be addressed include how the variables were coded (e.g., dummy/effect coding), procedures used to form composite variables (e.g., items used to form a subscale), procedures used to form aggregate level-two variables (e.g., average SES of all students at the school, versus average SES of the students at the school who participated in the study), and at which level(s) the variables were measured in the multilevel models. One way to convey these details efficiently is through the inclusion of a codebook in an appendix that provides information about the variables and their measurement (Lee & Loeb, 2000; Marks, 2000).

Measurement quality of the variables in terms of reliability and validity is also of critical importance. Most measures in educational studies contain error, and these errors, if not accounted for, can bias estimates of variance

parameters, variance ratios, fixed effects, and the standard errors of fixed effects (Woodhouse, Yang, Goldstein, & Rasbash, 1996). Consequently, researchers need to provide psychometric information on the variables used in the multilevel analysis. Reliance on estimates of reliability or validity provided in technical manuals or previously reported research is typically not sufficient because such estimates are sample specific (Thompson & Vacha-Haase, 2000). In situations in which the measurement error is substantial, researchers may consider analytical methods for specifying and adjusting for the measurement error (Longford, 1993; Woodhouse et al., 1996).

It is helpful to divide the study variables by the level at which they are measured (e.g., level one, level two, etc.). Researchers investigating a variable that is measured at different levels (e.g., student SES vs. SES of the school) need to present psychometric information about the variable at each level and discuss how the variable may have different meanings at different levels. In addition to presenting the variables at each level of analysis, the role(s) played by the variables in the study should be specified (e.g., outcome, predictor, covariate). The role delineation becomes important as the researcher attempts to communicate the multilevel models under investigation. For example, in a study examining the question of whether the relationship between student SES and student mathematics achievement (i.e., slope) varies across different types of schools, the researcher might identify student SES as a predictor and specify the β coefficient representing this relationship as random. In another study examining the effects of an instructional program on mathematics achievement, student SES may be used as a control variable, and therefore, the researcher might fix the variance of the β coefficient to zero.

Models

In view of the complexity of multilevel models, researchers need to address multiple issues in their descriptions of their models. First, the statistical models need to be specified clearly and fully. Second, the method used to center/scale each variable in the model should be provided. Third, the process used to derive the models should be communicated to help the reader understand the degree to which the analyses were exploratory or confirmatory in nature. Finally, the methods used to examine the integrity of the model should be detailed to help the reader evaluate the resulting inferences.

Specification

In multilevel modeling, the statistical models need to be presented in an understandable manner so that readers can gauge the appropriateness of

the models for addressing the research questions as well as for replicating the analyses. Although it is possible to communicate a multilevel model in words, verbal descriptions are often ambiguous or incomplete, and thus may not be an efficient way to communicate the model. A more effective strategy to specify the multilevel model is through the use of one or more equations for each level of the model. Some software programs, such as HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004), generate the equations that specify the model, facilitating insertion into a manuscript.

Consider a researcher who is studying students nested in schools. The researcher may be interested in the effects of student seventh-grade mathematics achievement (Math7) and student SES (SES) on student eighth-grade mathematics achievement (Math8). A level-one model could be developed to describe Math8 as a function of Math7 and SES within a specific school:

$$\text{Math8}_{ij} = \beta_{0j} + \beta_{1j}\,\text{Math7}_{ij} + \beta_{2j}\,\text{SES}_{ij} + r_{ij}, \tag{11.1}$$

where Math8_{ij} is the eighth-grade mathematics achievement score for the ith student in the jth school, β_{0j} is the intercept of the regression equation predicting Math8 in the jth school, β_{1j} is the regression coefficient indexing the strength of the association of Math7 with Math8 in the jth school, β_{2j} is the regression coefficient indexing the strength of the association of student SES with Math8 in the jth school, and r_{ij} is the error, which is assumed to be normally distributed with a covariance of Σ.

When students are nested in schools, as in this example, Σ commonly is assumed to be $\sigma^2\mathbf{I}$, where σ^2 is the variance and $\mathbf{I}$ is a $n \times n$ Identity Matrix, where n is the number of level-one units. This implies that the errors are modeled as if they were sampled independently from a normal distribution with variance, σ^2. If multilevel models are used for longitudinal data in which repeated measures are nested within individuals, one may want to relax this assumption to allow for the correlation among errors that are close together in time. A variety of alternative structures including first-order autoregressive have been discussed and presented in the methodological literature (Wolfinger, 1993). Note that one step in communicating the model is to be clear about the assumed structure of Σ.

After specifying the level-one model, the level-two model is specified. Returning to the example, the level-two model could be used to consider the effects of school context on Math8. Assume the researcher believes, either through theory or previous research, that the level of Math8 in the school depends on whether the school is using an experimental mathematics instructional program (Program) and the school SES (SchoolSES), and that Program also moderates the effects of Math7 on Math8. The level-two

model would use Program and School SES as predictors of some of the coefficients of the level-one model. One possible specification could be:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,\text{Program}_j + \gamma_{02}\,\text{SchoolSES}_j + u_{0j} \tag{11.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\,\text{Program}_j + u_{1j} \tag{11.3}$$

$$\beta_{2j} = \gamma_{20}, \tag{11.4}$$

where Program_j is coded 0 if school j is a control school that does not use the experimental mathematics program and coded 1 if school j is using the experimental program; SchoolSES_j is the measure of school-level SES at school j; and u_{0j} and u_{1j} are level-two errors, which are assumed to be normally distributed with a covariance of **T**. In this example, **T**, could be specified in several ways. One way is as a 2×2 unstructured covariance matrix,

$$\mathbf{T} = \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}, \tag{11.5}$$

which would imply that there was random variability in the intercepts (τ_{00}) and in the regression coefficients associated with Math7 (τ_{11}) and that the errors associated with the intercepts and Math7 coefficients may covary with each other (τ_{10}). One may find that the data support constraining a variance to zero, thus reducing the number of elements estimated in **T**. Alternatively, one might define the covariance structure so that a greater number of variance components are estimated. For example, the researcher also may allow the coefficients associated with student-level SES to vary randomly; in this case, an error term, u_{2j}, would be added to Equation 11.4, and **T** would become a 3-by-3 matrix. Part of communicating the model involves letting the reader know what structure was assumed for **T**.

Although it is common in the educational literature to see multilevel models communicated using regression equations for each level of the model, it is also possible to combine the regression equations into a single equation. By substituting the level-two model for β_{0j}, β_{1j}, and β_{2j} in the level-one model, the following combined model would be obtained:

$$\begin{aligned}\text{Math8}_{ij} = {} & \gamma_{00} + \gamma_{01}\,\text{Program}_j + \gamma_{02}\,\text{SchoolSES}_j + \gamma_{10}\,\text{Math7}_{ij} + \\ & \gamma_{11}\,\text{Program}_j * \text{Math7}_{ij} + \gamma_{20}\,\text{SES}_{ij} + u_{0j} + u_{1j}\,\text{Math7}_{ij} + r_{ij},\end{aligned} \tag{11.6}$$

which has the same form as the mixed linear model,

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}v + \varepsilon, \tag{11.7}$$

where **y** is a vector of outcome data, β is a vector of fixed effects, **X** and **Z** are known model matrices, ν is a vector of random effects, and ε is a vector of errors (Henderson, 1975). Again, the structure of the covariance matrices needs to be made explicit. Using mixed model notation, one typically refers to the covariance matrix of the level-one errors as **R** and to the covariance matrix of the level-two errors as **G**. For the above example, one could indicate that the blocks of **R** were specified as $\sigma^2\mathbf{I}$ and that the blocks of **G** were specified as 2 by 2 and unstructured,

$$\mathbf{G} = \begin{bmatrix} g_{11} & \\ g_{21} & g_{22} \end{bmatrix}, \tag{11.8}$$

where g_{11} is the random variance in the intercepts, g_{22} is the random variance in the regression coefficients associated with Math7, and g_{21} is the covariance between the errors associated with the intercepts and the Math7 coefficients.

The choice of using equations for each level or a single equation should be based on the judgment of which method will communicate most easily to the intended audience. Information for making this decision can be gleaned from consulting previous issues of the target journal to determine how multilevel models typically are communicated. If it is judged that the equations provide too much technical detail for the typical reader, an appendix could be included (for an example, see Marsh, Köller, & Baumert, 2001).

Centering of Predictors

Centering of the level-one and level-two predictors has implications for interpreting the results of multilevel models (Kreft & de Leeuw, 1998; Kreft, de Leeuw, & Aiken, 1995; Morrell, Pearson, & Brant, 1997; Raudenbush & Bryk, 2002) and, therefore, is an important consideration in reporting the results. In the example specified in Equations 11.1–11.5, suppose the seventh-grade mathematics achievement (Math7) was measured on a scale ranging from 200 to 800, student SES was dummy coded (0 = eligible for free or reduced lunch, 1 = not eligible), school SES was defined as the proportion of students in the school not eligible for free or reduced lunch, and the mathematics program variable was dummy coded (0 = control school, 1 = mathematics program school). If Math7 was kept in its natural metric, γ_{00} would be the predicted eighth-grade mathematics achievement (Math8) for a student in a control school with 0% of the students not eligible for free or reduced lunch, who is individually eligible for free or reduced lunch, and who has a Math7 score of zero. Since a Math7 score of zero is not possible, this coefficient is difficult to interpret in a substantively meaningful way. The effect of the instructional program in this model, γ_{01}, would be interpreted as the difference in the effectiveness of the two programs

when Math7 was zero (again, a value that is not particularly informative). Difficulties also would arise in interpreting the variance components. For example, the random variance in the intercepts, τ_{00}, would be the between-school variation in predicted Math8 scores for students who are eligible for free or reduced lunch and who have Math7 scores of zero. Centering or rescaling prior mathematics achievement makes the interpretation of the coefficients and variance components more meaningful.

One approach to scaling predictor variables is to subtract the grand mean of the predictor variable from each score $(x_{ij} - \bar{x}_{..})$; this can be done for variables at level one or at level two. Using grand-mean centering of Math7 and school SES in our example, γ_{00} is the predicted Math8 score for students in a control school with sample average school SES, who are individually eligible for free or reduced lunch, and who have a sample average Math7 score. Similarly, the effect of instructional program, γ_{01}, is interpreted as the difference in the effectiveness of the two programs for students having the sample average of seventh-grade mathematics achievement and the same individual and school SES.

A second approach to scaling the predictor variable is to subtract the level-two unit mean of the predictor variable from each score $(x_{ij} - \bar{x}_{.j})$; this centering process can only be done for level-one predictors. Using group-mean centering of Math7 and grand-mean centering of school SES in our example, γ_{00} is interpreted as the predicted Math8 score for a student in a control school with sample average school SES, who is individually eligible for free or reduced lunch, and whose Math7 score was at the sample average for his or her school. The effect of the experimental program, γ_{01}, is interpreted as the difference in the effectiveness of the two programs for students who are at their school's sample average level of mathematics achievement and have the same individual SES.

A third approach to scaling a predictor variable is to subtract a theoretically meaningful value (k) from each score $(x_{ij} - k)$. This approach is similar to grand-mean centering in that a constant is subtracted from each score. The β_{0j} is interpreted as the expected outcome for individuals at the specific value that has been set by the researcher. For example, in a growth curve model examining change in mathematics achievement from grades 1 through 8, a researcher may center the grade predictor at grade 8. In this case, β_{0j} is interpreted as the expected value of the outcome for a student in eighth grade.

The differences in the substantive interpretation of these regression coefficients (fixed effects) illustrate the importance of clearly delineating the type of centering that has been employed. In addition, centering has consequences for interpreting the variance components. For example, the variance in the intercepts will depend on how the intercepts are defined, which in turn depends on the centering. Vague statements that "all predictors were centered" or that "mean centering was employed to facilitate inter-

pretation of the models" are not sufficient to insure proper interpretation of the results. If model estimates are presented in tables, the researcher should use a table note to describe the type of centering used so that interpretation of parameter estimates readily follows.

Process for Defining the Model

In some situations researchers are able to use theory and past research to define the multilevel model(s) prior to examining their data. In these situations the data are used as a check to verify the reasonableness of the model but not as a means for building the model. Consequently, hypothesis testing for key parameters and the construction of confidence intervals around an effect of interest are relatively straightforward. When researchers rely on the data to help define the model, the research is more exploratory and strong inference becomes considerably more difficult. To critically examine the inferences made, the reader needs to fully understand the degree to which the data were used to develop the model.

Consider, for example, the model specified in Equations 11.1–11.5. Suppose the researcher had made a strong argument supporting the details of the model specification and that the only decision based on the data was to allow the errors in the level-two equations to covary. A reader concerned that this decision may have been incorrect could think through the potential consequences of estimating a covariance parameter that has a value of zero in the population. This type of misspecification can negatively affect the precision in estimating other parameters in the model (Verbeke, 1997) and sometimes leads to estimation difficulties (Van den Noortgate & Onghena, 2003; Verbeke, 1997). The reader may conclude that the potential misspecification has negligible consequences for interpretation if estimation difficulties were not encountered and a reasonable level of precision was obtained for the parameter estimates.

Alternatively (again considering the model specified in Equations 11.1–11.5), suppose the researcher arrived at this model after considering 12 potential predictors of variability in the intercepts and regression coefficients. The presented model contains only the predictors that were statistically significant. Again the reader may wish to consider the consequences of possible misspecifications. Relevant variables may have been omitted from the model (a possible consequence of insufficient power), which might lead to substantial biasing of the effect estimates of the included predictors. In this case, readers may judge the potential misspecifications to have substantial enough ramifications to alter the way they evaluate the results.

Evaluation of Model Integrity

A variety of statistical tools may be employed to obtain information about the integrity and trustworthiness of a model. Researchers may ex-

amine fit indices, the degree to which data are consistent with modeling assumptions, and the sensitivity of parameter estimates to outliers and changes in model specification. Such examinations may provide support for, or indicate appropriate caveats related to, the fidelity of model estimates. The clear explication of the results of investigations of model integrity, including what approaches were taken, what results were obtained, and what these results suggest about the model reported, is important in interpreting the study results.

Fit indices may be used to guide selection among alternative models. The fit indices most commonly used are the deviance statistic (Raudenbush et al., 2004), AIC (Akaike, 1974), and BIC (Schwartz, 1978). More details about model fit indices are provided in Chapter 7 of this volume (McCoach & Black, 2008). It is also important to note, however, that not all multilevel software packages provide all these estimates of model fit and that not all researchers use the same indices. Consequently, it is important to be specific about how model fit was assessed.

Distributional assumptions (normality and equal variance) are made about the errors at each level in the model. Violations can be suggestive of specification errors and can lead to biases in the standard errors at both levels of the model (Raudenbush & Bryk, 2002). The multilevel modeling results also can be influenced by outliers. There are multiple methods available to screen data for violations of assumptions (Jiang, 2001; Raudenbush & Bryk, 2002; Teuscher, Herrendorfer, & Guiard, 1994) and the presence of outliers (Longford, 2001). Given the variety of methods available, researchers need to not only communicate that data were screened for violations of assumptions and outliers but to note the specific methods used.

Also available are approaches for assessing the *impact* of outliers, assumption violations, and alternative specification decisions. Bayesian techniques, such as the Gibbs sampling methods as well as other strategies and algorithms, can be used to examine the impact of extreme observations at either level one or level two of the model (Seltzer, Novak, Choi, & Lim, 2002). Models can be estimated with and without a transformation of a nonnormal outcome variable to examine the impact of nonnormality on the results (for an example, see Kochenderfer-Ladd & Wardrop, 2001). Models also can be estimated under multiple plausible covariance structure specifications to examine the impact of specification decisions on inferences (Ferron, Dailey, & Yi, 2002). Because multiple methods are available to assess the degree to which inferences are sensitive to modeling decisions, researchers should communicate the specifics of any methods utilized.

Collectively, techniques employed to provide evidence of model robustness and sensitivity of parameter estimates to changes in model specification will serve to enhance the trustworthiness of an estimated model. For readers to critically evaluate the results presented and the inferences made,

they need to know the particulars of the methods used to evaluate model integrity.

Estimation and Inference

Technical details about estimation of the multilevel model and approaches to statistical inference allow readers to evaluate strengths and weaknesses of the methods selected and to permit replication. As such, these technical details should be viewed as an integral part of reporting the results. A variety of issues are subsumed under this topic, including estimation algorithms and the inferential methods used to conduct hypothesis tests and construct confidence intervals.

Estimation

A variety of methods are available for the estimation of parameters, each with its own strengths and weaknesses. As such, estimation methods and algorithms should be identified explicitly in the discussion of parameter estimation. In addition, identification of the specific software program and version used for estimation is helpful for readers interested in technical details about the analysis. In discussing the technical details, it also should be communicated whether estimation problems were encountered (e.g., improper variance estimates) and, if they were, how they were addressed.

Common methods of estimation for multilevel models include maximum likelihood (ML), restricted maximum likelihood (REML), and Bayesian (Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002). These methods of estimation can be carried out using many different algorithms, thus underscoring the need for definitive information regarding estimation methods and algorithms employed. For example, ML estimation may be accomplished using the expectation-maximization (EM) algorithm, the Newton-Raphson algorithm, the Fisher scoring algorithm, or iterative generalized least squares (IGLS), while Bayesian estimation may be accomplished using the Gibbs sampler. These algorithms have been programmed into many different software programs. Thus, one researcher may accomplish REML estimation using the EM algorithm programmed into HLM (Raudenbush et al., 2004), another may accomplish REML estimation using restricted iterative generalized least squares (RIGLS) using MLwiN (Rasbash, Steele, Browne, & Prosser, 2004), while a third may accomplish REML using the Newton-Raphson algorithm programmed in SAS PROC MIXED (SAS Institute Inc., 2000).

Reporting of the estimation method, estimation algorithm, software program, and whether estimation problems were encountered can be communicated effectively in a single sentence in the description of the data analy-

sis, a footnote, or a technical appendix. The use of less common estimation approaches, such as bootstrapping, robust ML, and robust REML methods (Carpenter, Goldstein, & Rasbash, 1999; Meijer, Van der Leeden, & Busing, 1995; Richardson & Welsh, 1995), may require more explication, possibly in an appendix.

Estimation methods typically will produce point estimates of each parameter in the multilevel model and these estimates are often valuable in addressing particular research questions. Additional information about the parameter estimates often is provided to aid the researcher in making inferences, possibly taking the form of hypothesis tests and/or confidence intervals for parameters of interest. Clear communication of the types of estimates calculated and details about the approach employed are important for valid interpretation of such inferential statistics. When considering the options available, it becomes important to distinguish between inferences made about variance parameters (elements in Σ and $\mathbf{T}$), fixed effects (γ's), and random level-one coefficients (e.g., β_{0j}).

Inferences about Variance Parameters

The simplest approach to creating a confidence interval (CI) for a variance parameter is to use the standard error of the variance parameter estimate, computed from the inverse of the information matrix. By adding and subtracting 1.96 times the standard error of the parameter estimate, one can create a 95% CI, assuming a normal sampling distribution. This approach, however, has limitations, especially when the sample size is small or the variance parameter is near zero (Littell, Milliken, Stroup, & Wolfinger, 1996; Raudenbush & Bryk, 2002). For such data, researchers may consider other options, including the Satterthwaite approach (Littell et al., 1996), bootstrapping (Carpenter et al., 1999; Meijer et al., 1995), a method based on local asymptotic approximations (Stern & Welsh, 2000), and, if the data are balanced, an approach based on a set of quadratic forms (Yu & Burdick, 1995). These alternative methods can lead to different results. If other researchers are to critically evaluate or replicate the analysis, they need to know the specific methods used. Consequently, this is another technical detail that should be reported.

For researchers wishing to test hypotheses regarding variance parameters, a similar variety of choices is available. The simplest approach would be to conduct a *z*-test by dividing the estimate by its standard-error. Although this approach is asymptotically valid, like the standard error based CIs noted previously, it becomes questionable when the sampling distribution cannot be assumed normal. Alternative approaches include a likelihood ratio χ^2 (Littell et al., 1996), an approximate χ^2 test described by Raudenbush and Bryk (2002), bootstrapping (Carpenter et al., 1999; Meijer et al., 1995), and a likelihood ratio test based on the local asymptotic approximation (Stern

& Welsh, 2000). Again, different choices can lead to different results and thus the method should be reported.

Inferences about Fixed Effects

Inferences about fixed effects may be obtained from confidence intervals for the effects of interest. For example, a 95% CI could be constructed around the point estimate by adding and subtracting 1.96 times the standard error. This approach assumes a normal sampling distribution, which can be demonstrated asymptotically, but which becomes questionable for smaller samples. Consequently, one may utilize a critical t-value with ν degrees of freedom. Several methods for defining the degrees of freedom have been given (Giesbrecht & Burns, 1985; Kenward & Roger, 1997), and some software packages allow for different definitions to be specified. An alternative to assuming an approximate t-distribution is to turn to bootstrapping to construct the confidence intervals.

Hypothesis tests also can be conducted using t- or F-tests with approximate degrees of freedom. Again, different approximations have been suggested, and thus, researchers need to be clear about the method used for obtaining the degrees of freedom for these tests. Several alternatives to these approximate tests have been discussed. These include a test based on a Bartlett-corrected likelihood ratio statistic (Zucker, Lieberman, & Manor, 2000), a permutation test (Reboussin & DeMets, 1996), and bootstrapping. Researchers using one of these methods should specify the approach that was used and the rationale.

Inferences about Level-One Coefficients

Researchers also may be interested in estimating the random level-one coefficients and making inferences about these coefficients. For example, a researcher who is interested in estimating the effects of seventh-grade mathematics achievement on eighth-grade mathematics achievement may wish to obtain a separate effect estimate for each school. Again, there are multiple choices for estimation and inference, and it is important for the researcher to convey the choices made.

One approach would be to estimate the level-one model separately for each school using ordinary least squares (OLS) estimation methods, in which case standard methods are available for constructing confidence intervals and testing hypotheses about coefficients. With this approach the estimate for a specific school is based only on information from that school, which may be just a few observations. By failing to use the information from the other schools, the obtained estimate is not as precise as it could be.

An alternative is to obtain Empirical Bayes estimates, which consider all available information. Empirical Bayes estimates tend to pull each school's

effect estimate toward a value predicted by the model, with the amount of adjustment depending upon the uncertainty in the effect estimate being considered and the variability in the effect estimates. This process biases the estimates but provides values that tend to be closer to the parameter values than those based on OLS estimation, resulting in a smaller expected mean square error (Raudenbush & Bryk, 2002). For Empirical Bayes estimates, the standard errors can be computed and used for the creation of confidence intervals or z-tests of statistical significance.

RESULTS

Researchers may consider reporting at least two types of results: (a) preliminary results that address the properties and quality of the data (e.g., measures of central tendency, reliability of outcomes, predictors, and level-one coefficients such as intercepts and slopes), missing data patterns and relationships of missing data to relevant variables, model assumptions including normality and homogeneity of variance, and model building steps; and (b) primary results directly addressing the research questions. Two examples are used to illustrate various approaches to reporting results. The first considers a two-level model examining mathematics achievement of students nested in schools; the second involves a two-level growth curve model of reading achievement.

Preliminary Results

Tables presenting descriptive univariate information about the variables under investigation (e.g., mean, standard deviation, range, skewness, kurtosis) and correlations among variables are common in published research. With a few format changes in these tables, important information about the variables in the multilevel models can be communicated efficiently. Examples of this type of information are illustrated based on data for the two-level mathematics achievement example in Table 11.4 (univariate statistics) and Table 11.5 (correlations). Dividing the study variables by the level at which they are measured provides information about the potential variables available for model building at each level and their distributional properties. Inclusion of sample sizes for each variable provides information about missing data, with implications for issues related both to statistical power and to potential convergence and estimation problems in model development.

TABLE 11.4 Example of Table for Presenting Descriptive Data for Variables in Two-Level Model with Students Nested Within Schools

Variable	*N*	*M*	*SD*	Skewness	Kurtosis	Outliers
			Level one			
Math8	2000	303.25	31.06	0.20	0.45	None
Math7	1967	303.15	51.12	–0.01	–0.12	None
Student SES	2000	0.25				
			Level two			
School SES	40	0.25	0.16	–0.21	0.28	None
Program	40	0.50				

Note: Math8 is eighth-grade mathematics achievement; Math7 is seventh-grade mathematics achievement; Student SES is coded 0 if eligible for free or reduced lunch and 1 if not eligible; School SES equals the proportion of students in the study at a school that are not eligible for free or reduced lunch; Program is coded 0 for a control school and 1 for a program school; and an outlier was defined as an observation exceeding 1.5 interquartile ranges beyond the 1st or 3rd quartile.

TABLE 11.5 Example of Table for Presenting Pearson Product Moment Correlations for Variables in Two-Level Model with Students Nested Within Schools

	Level one (*N* = 1967)		
	Math8	**Math7**	**Student SES**
Math8	1.00		
Math7	0.59	1.00	
Student SES	0.18	0.14	1.00
	Level two (*J* = 40)		
	School SES	**Program**	
School SES	1.00		
Program	–0.02	1.00	

Note: Math8 is eighth-grade mathematics achievement; Math7 is seventh-grade mathematics achievement; Student SES is coded 0 if eligible for free or reduced lunch and 1 if not eligible; School SES equals the proportion of students in the study at a school that are not eligible for free or reduced lunch; Program is coded 0 for a control school and 1 for a program school; the *N* of 1967 is based on listwise deletion.

As part of the presentation of descriptive information, it is important to distinguish what outcome variables are being examined in the research questions and then to present descriptive information about these outcomes. The

potential for confusion on this issue can be illustrated with the growth curve modeling example for reading achievement in which a researcher was interested in examining changes in reading achievement from grades 1 to 5 and the factors associated with these changes. The researcher may identify reading achievement as the outcome variable and then only present descriptive statistics and psychometric information, such as reliability estimates, for reading achievement at each grade level. Given that the researcher's focus is on *changes* in reading achievement, the outcome variable is technically the slope parameter estimate and, therefore, descriptive information (minimum, maximum, mean, standard deviation, skewness, kurtosis) both for EB and OLS slope estimates along with reliability estimates should be presented. Similar information should be presented if intercepts (e.g., initial status in a growth curve) are the focus of the research questions (see Table 11.6). The reliability estimates for the slope and intercept parameters, which are calculated in some software programs, can be used to make decisions about whether these coefficients should be specified as fixed or random and also provide information about the extent to which relationships between predictors and the coefficients may be attenuated.

An alternative way to communicate information about the distribution of intercepts and slopes efficiently is to provide a graphical display of the reading trajectories. If the number of level-two units is too large for a clear visual display of *all* units, the researcher could provide a visual display based on a random sample of the level-two units (see Figure 11.1). In addition to

TABLE 11.6 Example of Table for Summarizing Reading Achievement for Two-Level Growth Curve Model

Outcome	*N*	Min	Max	*M*	*SD*	Skewness	Kurtosis	Outliers
OLS								
Intercept (initial status)	100	168.5	273.5	220.0	24.5	0.01	–0.48	None
Slope (yearly change)	100	–30.4	150.6	53.9	37.1	0.23	–0.05	None
EB								
Intercept (initial status)	100	186.6	247.7	220.0	13.7	–0.23	–0.07	None
Slope (yearly change)	100	–33.2	148.5	53.9	35.9	0.19	–0.08	None

Note: OLS is Ordinary Least Squares; EB is Empirical Bayes; the time variable was scaled in yearly increments from grades 1 to 5 with zero corresponding to the beginning of the study (grade 1); an outlier was defined as an observation exceeding 1.5 interquartile ranges beyond the 1st or 3rd quartile; and reliability of OLS regression coefficient estimates for intercepts and slopes were .80 and .45, respectively.

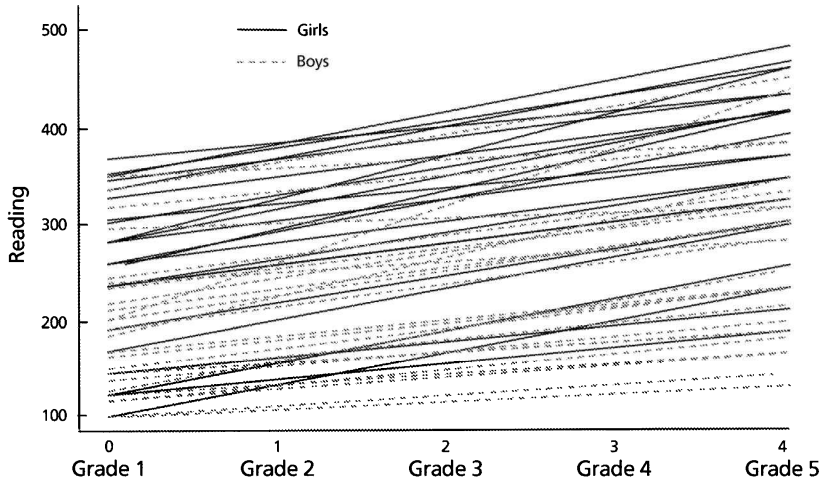

Figure 11.1 Example of fitted OLS linear regressions of reading achievement on grade level for a random sample of 50 students.

Stem	Leaves
1.0	
.9	23
.8	125679
.7	1245677899
.6	01446678899
.5	012233446788999
.4	0001223334445678899
.3	00112233444455566
.2	0111223458
.1	144789
.0	3689

Figure 11.2 Example of a stem-and-leaf plot presenting R^2 values for fitted OLS regressions of reading achievement on grade level for 100 students.

summarizing the slope and intercept distributions, it is also useful to summarize R^2 values (see Figure 11.2) for the individual OLS regression models. Displays summarizing the R^2 values for different growth models (linear or quadratic) then could be used to support decisions about the choice of model while at the same time providing a scaffold for the multilevel results addressing the research questions. Another way of summarizing the level-one regression would be to include a table listing the R^2 values along with the OLS level-one equations.

As part of these preliminary analyses, the researcher should communicate details about the data that may impact modeling in the primary analyses. It is important to discuss issues of nonnormality, heteroscedastic-

ity, multicollinearity, and outliers. Potential violations of the underlying assumptions need to be examined thoroughly and communicated clearly to the reader. In addition to reporting any anomalies that are found during data screening, researchers should document analysis decisions that are made in light of the data. For example, a researcher may transform a variable to improve normality or use an alternative covariance structure to address heteroscedasticity. In situations where how to proceed is somewhat ambiguous, researchers also should provide information on the degree to which the results are sensitive to the data anomalies or alternative modeling decisions. Given the space requirements of many journals, the results of data screening activities will need to be summarized concisely, and some of the technical details may need to be handled through footnotes, or an appendix.

Preliminary analyses should include computing the intraclass correlation coefficient (ICC). The ICC, derived from an unconditional model with no within- and between-group predictors (also called the one-way random effects ANOVA model, or the empty model), provides baseline information for evaluating the relative contributions of within- and between-group predictors.

Primary Results

In presenting the primary results from multilevel analyses, researchers should provide a listing of all estimated parameters for each model that is interpreted, while also striving to focus the reader's attention on the specific estimates and results that address the research questions. This can be challenging because the links between the questions, models, and statistical results are not always as apparent as they would be in applications using simpler statistical models. Focus can be achieved by adding visual cues such as bold-faced type in tables (see Wainer, 1997), by including statements interpreting the key parameter estimates in the narrative, and by illustrating effects using graphical displays.

Example 1: Students Nested within Schools

Consider again the example where eighth-grade mathematics achievement is being predicted based on seventh-grade mathematics achievement, student SES, school SES, and whether or not the school had used the mathematics program. Suppose the primary purpose of the research is to estimate the effects of the mathematics program on eighth-grade mathematics achievement and the degree to which the program's effect depends on prior achievement (seventh-grade mathematics achievement) of the students. The researcher may wish to start by pointing the reader to a table with

the complete listing of the parameter estimates and an indication of the precision of these estimates (e.g., standard errors or confidence intervals). There are several ways to structure such a table. One possibility is to use the format shown in Table 11.7, where predictors are listed in rows, and columns are used for different models. This format parallels a relatively standard way of reporting and comparing multiple regression models, which may facilitate a reader's understanding of the results. This table includes symbols commonly used to refer to fixed effects (e.g., γ_{00}, γ_{10}) and variance estimates (e.g., σ^2, τ_{00}) along with brief descriptors. The included symbols match those used when the model was specified (Equations 11.1–11.5) to facilitate the connection of the estimates in the table to the parameters in the model. An alternative method for tabular representation of multilevel analysis results can be found in Ethington (1997).

Since the primary focus of this analysis is on estimating the effect of the mathematics instructional program, the narrative should provide an interpretation of the estimated program effect ($\hat{\gamma}_{01}$), which as noted previously would depend on how the variables were scaled or centered. For example, assuming grand-mean centering of Math7, the effect estimate, $\hat{\gamma}_{01}$, would be interpreted in a statement such as: "students with a sample average level of seventh-grade mathematics achievement who are in a school with the mathematics instructional program are predicted to have an eighth-grade mathematics achievement score that is $\hat{\gamma}_{01}$ points higher than similar students in a control school."

Attention should also be drawn to the cross-level interaction effect (γ_{11}), which suggests that the difference in expected eighth-grade mathematics achievement between programs is not constant across seventh-grade achievement levels. A graphical display of predicted eighth-grade mathematics achievement as a function of seventh-grade mathematics achievement and program (see Figure 11.3) could be constructed using the equations with estimated parameter values. This graph helps to communicate the degree to which the program effect differs for students of varying levels of seventh-grade mathematics achievement.

An alternative display could be constructed by graphing the program effect as a function of seventh-grade mathematics achievement, where the program effect is defined as the difference in expected eighth-grade mathematics achievement between comparable program and control students at a specified level of seventh-grade mathematics achievement. Confidence interval bands then could be added (Tate, 2004), and the range of seventh-grade mathematics achievement scores for which the difference between programs is statistically significant would become apparent. An example using 95% confidence interval bands is provided in Figure 11.4.

TABLE 11.7 Example of Table Summarizing REML Parameter Estimates for Two-Level Model of Eighth-Grade Mathematics Achievement

	Unconditional model			Full model		
Parameter	Parameter estimate	SE	95% CI	Parameter estimate	SE	95% CI
Fixed effects						
Intercept (γ_{00})	303.10	0.85	301.43 to 304.77	299.76	1.12	297.57 to 301.95
Math7 (γ_{10})	—	—	—	0.44	0.05	0.34 to 0.53
Student SES (γ_{20})	—	—	—	5.87	1.06	3.79 to 7.95
School SES (γ_{02})	—	—	—	6.62	3.61	–0.45 to 13.69
Program (γ_{01})	—	—	—	1.90	0.85	0.23 to 3.57
Program*Math7 (γ_{11})	—	—	—	–0.24	0.01	–0.26 to –0.22
Variance estimates						
Level-one variance (σ^2)	784.53	24.93	735.67 to 833.39	358.43	11.46	335.97 to 380.89
Intercept variance (τ_{00})	2.77	3.45	0 to 9.53	5.31	2.89	0 to 10.97
Slope variance (τ_{11})	—	—	—	0.05	0.02	0.01 to 0.09
Error covariance (τ_{10})	—	—	—	0.36	0.18	0.01 to 0.71

Note: Math7 is seventh-grade mathematics achievement grand-mean centered; Student SES is coded 0 if eligible for free or reduced lunch and 1 if not eligible; School SES is the grand-mean centered proportion of students in the study at a school that are not eligible for free or reduced lunch; Program is coded 0 for a control school and 1 for a program school; CIs constructed using 1.96*SE; level-one sample size equals 1967; level-two sample size equals 40; and the intra-class correlation (ICC) derived from the unconditional model equals .0035.

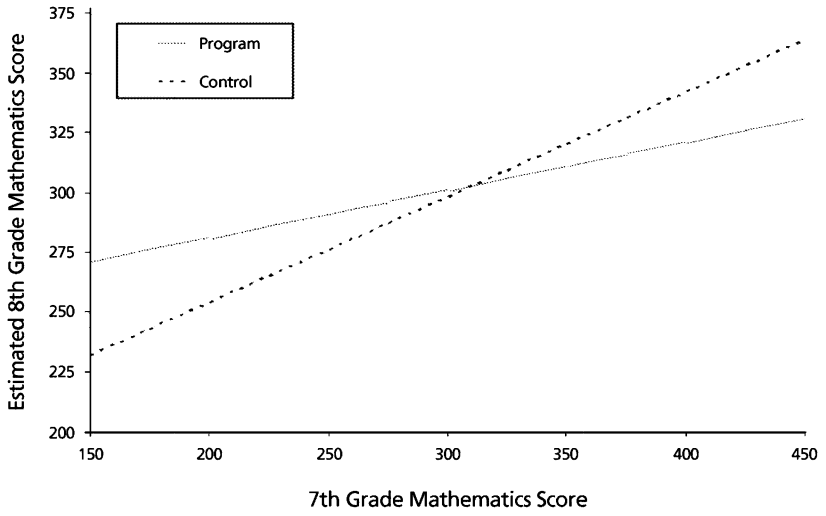

Figure 11.3 Graphical illustration of the effect of the mathematics program on eighth-grade mathematics achievement as a function of seventh-grade mathematics achievement for low SES students from a school with average SES.

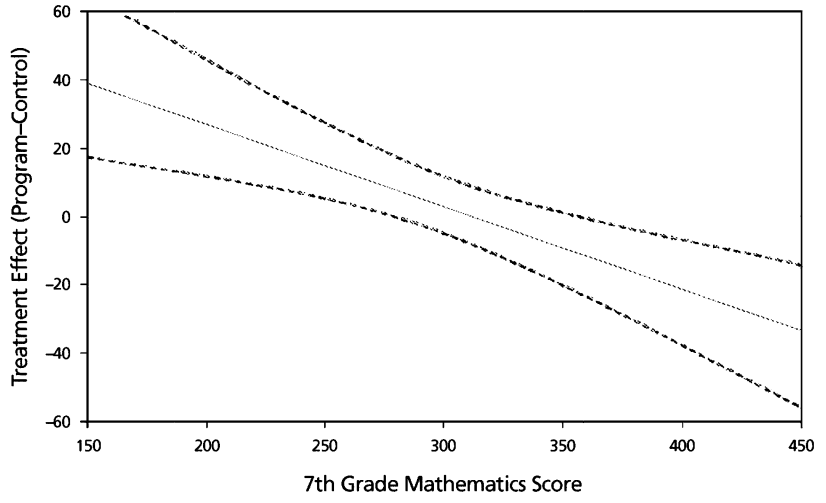

Figure 11.4 Graphical illustration of the mathematics program effect (solid line) and 95% confidence interval (dotted lines) as a function of seventh-grade mathematics achievement.

In addition to graphical displays designed to illustrate effects of interest, researchers sometimes compute pseudo R^2 values, giving the proportion of the variance at a particular level associated with the effect of interest (Kreft & de Leeuw, 1998; McCoach & Black, 2008; Snijders & Bosker, 1994).

Given that alternative calculations have been debated in the literature, it is important to specify the particular method used to estimate the pseudo R^2 value. In addition, it is important to be aware that the use of such indices is controversial and thus researchers providing these indices should do so carefully and in a manner that recognizes the limitations of the methods. For applications illustrating the use of pseudo R^2 values, see McCoach and Black, 2008, and Singer and Willett (2003).

Some researchers have provided standardized effect estimates by standardizing the regression coefficients from the multilevel model (Purcell-Gates, Degener, Jacobson, & Soler, 2002). Different methods can be used to standardize (e.g., across total sample versus within level-two units), and thus, it is important to communicate the details of the standardization process. As with pseudo R^2 values, the standardization of regression coefficients is controversial. For example, Willett, Singer, and Martin (1998) concluded that standardized regression coefficients may be misleading and caution against their use. Others have argued that for experimental studies, such as cluster-randomized trials, standardized effect sizes should be calculated and presented (Spybrook, 2008). Before presenting this type of information, researchers should critically evaluate whether it adds to their results and, if so, to present the information along with a discussion of the limitations of standardization.

Example 2: Growth Curve Model

As a second example, consider a longitudinal study of changes in reading achievement over the elementary years, where the research questions focus on the form of change (e.g., linear, nonlinear), the variation in growth parameters (e.g., intercept and slope), and gender differences in initial status and changes in reading achievement. After preliminary results have been presented, a table listing variance and covariance estimates, fixed effects, and fit indices for models where different growth trajectory forms were assumed can be useful to summarize information pertinent to questions of trajectory form and variability. Assume the researcher considered three models (an intercept only model, a linear growth trajectory model, and a quadratic growth trajectory model) and that for each, $\boldsymbol{\Sigma}$ was assumed to be $\sigma^2\mathbf{I}$ and $\mathbf{T}$ was assumed to be unstructured. One way of presenting the results for comparison of the models is provided in Table 11.8.

After identifying an appropriate form for the growth trajectories and determining the variability in the growth parameters, the researcher could address the question of the degree to which the growth trajectories differ for boys and girls. Assume the quadratic growth curve model best fit the data based on the AIC and BIC and that there was sufficient variation in the intercepts, linear, and quadratic terms, to use each of the growth parameters as an outcome in the examination of gender differences. The

TABLE 11.8 Example of Table Summarizing REML Parameter Estimates for Two-Level Growth Curve Models of Reading Achievement

	Intercept only model		Linear model		Quadratic model	
Parameter	**Parameter estimate**	**95% CI**	**Parameter estimate**	**95% CI**	**Parameter estimate**	**95% CI**
Fixed Effects						
Intercept (γ_{00})	327.8 (6.12)	315 to 339	219.9 (2.30)	215 to 224	210.2 (1.53)	207 to 213
Time: Linear (γ_{10})			53.9 (3.80)	46.5 to 61.4	73.4 (1.30)	70.8 to 75.9
Time^2: Quadratic (γ_{20})					−4.86 (0.90)	−6.63 to −3.09
Variance Estimate						
Level one (σ^2)	11076 (799.0)	9510 to 12642	524.0 (43.62)	438 to 609	75.3 (7.65)	60 to 90
Intercept (τ_{00})	1497 (561.5)	396 to 2597	275.9 (89.54)	100 to 451	198.7 (39.03)	122 to 275
Linear (τ_{11})			1326.0 (200.0)	934 to 1718	69.6 (25.51)	19 to 119
Intercept, Linear (τ_{10})			−493.3 (111.2)	−711 to −275	88.8 (22.45)	45 to 133
Quadratic (τ_{22})					73.3 (11.73)	50 to 96
Intercept, Quadratic(τ_{20})					−13.3 (15.05)	−43 to 16
Linear, Quadratic (τ_{21})					16.1 (11.73)	−7 to 39
	AIC	BIC	AIC	BIC	AIC	BIC
Fit Indices	6124.4	6129.6	4990.1	5000.5	4444.3	4462.5

Note: Time is scaled in years and is centered so that zero corresponds to the beginning of first grade; standard errors (SE) follow parameter estimates in parentheses; for variance estimates τ_{00}, τ_{11}, and τ_{22} are residual variances, while τ_{10}, τ_{20}, and τ_{21} are residual covariances; CIs were constructed using 1.96*SE; AIC = Akaike's Information Criterion; BIC = Bayes Information Criterion; estimates based on 100 students, all with five observations.

TABLE 11.9 Example of Table Summarizing REML Parameter Estimates for the Model Relating Gender to Reading Growth Curves

		Intercepts (π_{0i})	Linear terms (π_{1i})	Quadratic terms (π_{2i})
Fixed effects				
Intercept (β_{p0})	Estimate	201.44	66.99	–5.06
	SE	1.95	1.57	1.26
	95% CI	197.6 to 205.3	63.9 to 70.1	–7.6 to –2.6
Gender (β_{p1})	Estimate	17.48	12.79	0.39
	SE	2.76	2.21	1.78
	95% CI	12.0 to 22.9	8.4 to 17.1	–3.1 to 3.9
Variances	Estimate	123.5	29.5	74.0
	SE	27.9	19.9	11.4

Note: Time is scaled in years and is centered so that zero corresponds to the beginning of first grade; Gender is dummy coded (0 = Male, 1 = Female); residual level-one variance, σ^2, is 75.3, and the error covariances between intercept and linear, intercept and quadratic, and linear and quadratic terms are $\tau_{10} = 32.1$, $\tau_{20} = -15.1$, and $\tau_{21} = 14.8$; CIs were constructed using degrees of freedom estimated through the containment method; estimates based on 100 students, each with five observations.

multilevel model using gender as a predictor of each growth parameter could be arranged using a format that parallels Table 11.7 or Table 11.8, or alternatively, it could be arranged so that the columns corresponded to the growth parameters (intercept, linear term, quadratic term) and the rows correspond to the variables used to predict each growth parameter. This type of arrangement is provided in Table 11.9.

The interpretation of the coefficient describing the effect of a predictor such as gender on an intercept parameter is relatively straightforward once the type of centering has been specified, and it parallels the interpretation of an effect for a predictor variable in a multiple regression model. However, the interpretation of a coefficient describing the effect of a predictor such as gender on either the linear or quadratic parameter estimate is more complex and, in fact, addresses the question of a cross-level interaction (i.e., does gender moderate the relationship between time and reading achievement?). Given this complexity, it is suggested that a graphical display of this cross-level interaction be constructed using the equations with estimated parameter values and then presented to aid interpretation (see Figure 11.5 for an example).

In summary, several suggestions have been made for communicating preliminary and primary results. Preliminary results should be presented that include univariate summaries of the variables under investigation, the correlations among these variables, summaries of the distributions of the random level-one coefficients, and the ICC. Primary results should include

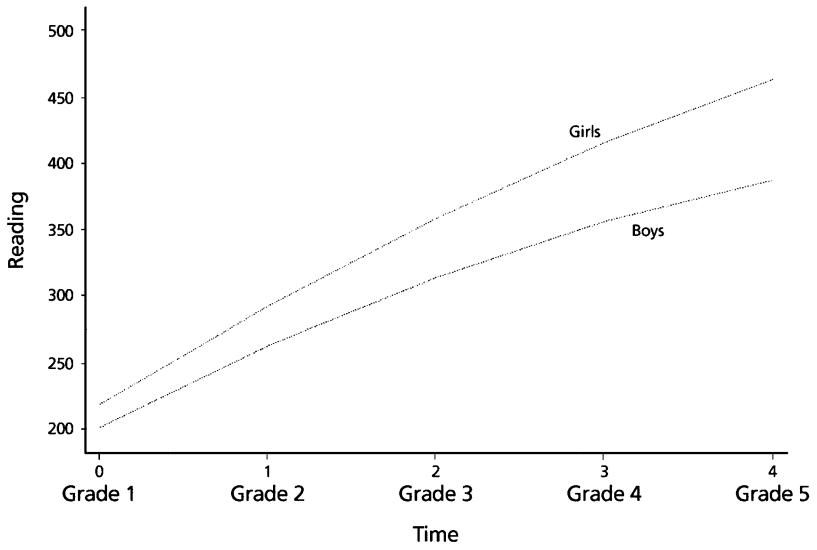

Figure 11.5 Graphical illustration of the predicted reading achievement trajectories of girls and boys based on a quadratic model with time measured in years and centered so zero corresponds to the beginning of first grade.

all parameter estimates of interpreted models (typically in a table) along with indications of the precision of these estimates (i.e., standard errors or confidence intervals). These tabular results should be supplemented with narrative interpretation, effect size calculations, and/or graphical displays to bring attention to the specific estimates or results that address the research questions.

DISCUSSION

The discussion section provides researchers an opportunity to evaluate, interpret, and qualify the results of the study. Researchers should provide a concise statement about the relationship between the results and the original research questions, emphasizing any practical as well as theoretical implications. Description of the limitations of the study resulting from the type of design, sampling, measurement procedures, and analysis should be provided. Researchers can link their findings to past research and articulate the degree to which the results are generalizable based on the study design and analyses.

General guidelines for discussing the results of empirical studies, such as those provided in the *Publication Manual of the American Psychological Asso-*

ciation (American Psychological Association, 2001) or the *American Medical Association Manual of Style* (Iverson et al., 1998), should be augmented by consideration of specific issues relative to multilevel modeling. Such issues might include: (a) what information is provided by the multilevel approach that was not provided in previous investigations that relied on the use of more traditional analyses (e.g., multiple regression) and (b) how the results may have been impacted by the decisions made during the multilevel model development and estimation.

SUMMARY

This chapter has provided a series of suggestions organized around the sections traditionally reported in published research. Some suggestions echo general recommendations made for reporting research, discussing issues such as sampling, variable selection, research design, and the connection between research questions and analyses (e.g., AERA Task Force on Reporting Research Methods, 2006). Other suggestions are more specific to multilevel modeling, focusing on issues such as estimation and inference, the nature of multilevel data, and reporting multilevel results.

To summarize these suggestions, a list of questions was developed that generally should be answerable by the *reader* of a well-written report of a multilevel modeling application. Mirroring the structure of a research report, these questions are organized into five categories: (1) research study (e.g., What sampling strategy was used?), (2) model specification (e.g., How many models were estimated?), (3) estimation and inference (e.g., What method of estimation was used?), (4) data (To what degree were data consistent with distributional assumptions?), and (5) results (Which specific results addressed each research question?). These questions are presented in the form of a checklist in Table 11.10.

Those preparing reports of multilevel modeling applications could use this checklist as a tool, asking themselves whether the consumer of the research report could answer these questions. Alternatively, one might ask colleagues to review a manuscript and attempt to answer the questions. The ability of colleagues to answer the questions may suggest areas that warrant additional attention and clarification.

The manner in which multilevel results are organized and presented has the potential to critically impact the utility, understanding, and credibility of the research. This belief motivated the writing of this chapter and informed the development of suggestions as to *what* to present and *how* to present multilevel results. These suggestions will need to be evaluated critically in the context of novel applications and may need further refinement as the techniques used in multilevel modeling evolve. Nonetheless,

TABLE 11.10 Checklist for Report of Multilevel Study

Could the reader answer the following general questions about the study?	Yes	N/A
What were the purposes/research questions for the study?		
Was the literature reviewed consistent with the study purposes and methods?		
What sampling strategy was used at each level (e.g., probability)?		
What sampling weights, if any, were used?		
How many units were at each level of the analysis?		
How were lower-level units distributed across upper-level units?		
Was a power analysis used to determine the number of units at each level?		
What study design was used (e.g., experimental, quasi-experimental)?		
What variables were used in the analyses?		
What is the validity evidence for each variable?		
What is the reliability evidence for each variable?		
Could the reader answer the following questions about model specification?	**Yes**	**N/A**
How many models were estimated?		
What were the fixed effects in each estimated model?		
What was the covariance structure of each estimated model?		
What process was used to define the fixed effects for each model?		
What process was used to define the covariance structure for each model?		
What method was used to evaluate model fit?		
How was each variable centered, coded, or scaled?		
Could the reader answer the following questions about estimation and inference?	**Yes**	**N/A**
What software and version were used?		
What method of estimation was used (e.g., REML, ML)?		
Were estimation problems (e.g., improper variance estimates) encountered?		
If estimation problems were encountered, how were they addressed?		
What methods were used to make inferential statements?		
Could the reader answer the following questions about the data?	**Yes**	**N/A**
What was the structure of the data (e.g., students nested in schools)?		
How were the variables and level one coefficients distributed?		
To what degree were variables correlated?		
Were data missing? And if so, how were they treated?		
To what degree did missing data impact the results?		
Were there outliers? And if so, how were they identified and handled?		
To what degree did outliers influence the results?		
To what degree were the data consistent with the distributional assumptions?		
To what degree were the results sensitive to questionable assumptions?		
Could the reader answer the following questions about the results?	**Yes**	**N/A**
Which specific results addressed each research question?		
What was the ICC?		
What was the estimated value of each parameter in each interpreted model?		
How precise was each estimate (e.g., SE, CI)?		
How do the limitations impact interpretation?		

Note: N/A = Not applicable

it is hoped that these suggestions will be useful to researchers reporting multilevel modeling applications and will serve to improve the consistency and clarity of the reported results from multilevel analyses.

REFERENCES

AERA Task Force on Reporting Research Methods. (2006). *Standards for reporting on empirical social science research in AERA publications.* Retrieved February 1, 2007, from http//www.aera.net/?id=1480

Akaike, H. (1974). A new look at the statistical model of identification. *IEEE Transaction on Automatic Control, 19,* 716–723.

American Psychological Association. (2001). *Publication manual of the American Psychological Association.* Washington, DC: author.

Boomsma, A. (2000). Reporting analyses of covariance structures: Teacher's corner. *Structural Equation Modeling, 7,* 461–483.

Boote, D. N., & Beile, P. (2005). Scholars before researchers: On the centrality of the dissertation literature review in research preparation. *Educational Researcher, 34,* 3–15.

Carpenter, J., Goldstein, H., & Rasbash, J. (1999). A non-parametric bootstrap for multilevel models. *Multilevel Modeling Newsletter, 11,* 2–5.

Collins, L. M., Schafer, J. L., & Kam, C. M (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6,* 330–351.

Ethington, C. A. (1997). A hierarchical linear modeling approach to studying college effects. *Higher Education: Handbook of Theory and Research, 12,* 165–194.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37,* 379–403.

Ferron, J., Hess, M. R., Hogarty, K. Y., Dedrick, R. F., Kromrey, J. D., Lang, T. R., et al. (2006, April). *Multilevel modeling: A review of methodological issues and applications.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Giesbrecht, F., & Burns, J. (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *Biometrics, 41,* 477–486.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics, 31,* 423–447.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage Publications.

Iverson, C., Flanagin, A., Fontanarosa, P. B., Glass, R. M., Glitman, P., Lantz, J. C., et al. (1998). *American Medical Association Manual of Style* (9th ed.). Philadelphia: Lippincott Williams & Wilkins.

Jiang, J. M. (2001). Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics, 29,* 1137–1164.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345–390). Charlotte, NC: Information Age Publishing.

Kenward, M., & Roger, J. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983–997.

Kochenderfer-Ladd, B., & Wardrop, J. L. (2001). Chronicity and instability of children's peer victimization experiences as predictors of loneliness and social satisfaction trajectories. *Child Development, 72*, 134–151.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel models.* London: Sage.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*, 1–21.

Lee, V. E., & Loeb, S. (2000). School size in Chicago elementary schools: Effects on teachers' attitudes and students' achievement. *American Educational Research Journal, 37*, 3–31.

Littell, R., Milliken, G., Stroup, W., & Wolfinger, R. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute Inc.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

Longford, N. (1993). Regression analysis of multilevel data with measurement error. *British Journal of Mathematical and Statistical Psychology, 46*, 301–311.

Longford, N. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society Series A-Statistics in Society, 164*, 259–273.

Ma, X., Ma, L., & Bradley, K. D. (2008). Using multilevel modeling to investigate school effects. In A. A. O'Connell and D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 59–110). Charlotte, NC: Information Age Publishing.

Maciejewski, M. L., Diehr, P., Smith, M. A., & Hebert, P. (2002). Common methodological terms in health services research and their symptoms. *Medical Care, 40*, 477–484.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*, 153–184.

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self concept. *American Educational Research Journal, 38*, 321–350.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell and D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age Publishing.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analysis. *Psychological Methods, 7*, 68–82.

Meijer, E., Van der Leeden, R., & Busing, F. (1995). Implementing the bootstrap multilevel model. *Multilevel Modeling Newsletter, 7*, 7–11.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modeling Newsletter, 7*, 11–15.

Morrell, C. H., Pearson, J. D., & Brant, L. J. (1997). Linear transformations of linear mixed-effect models. *The American Statistician, 51*, 338–343.

Murray, D. M. (1998). *Design and analysis of group randomized trials.* New York: Oxford University Press.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of methodological developments. *American Journal of Public Health, 94*, 423–432.

Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 199–244). Charlotte, NC: Information Age Publishing.

Purcell-Gates, V., Degener, S. C., Jacobson, E., & Soler, M. (2002). Impact of authentic adult literacy instruction on adult literacy practices. *Reading Research Quarterly, 37*, 70–92.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MLwiN version 2.0.* London: Institute of Education.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173–185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199–213.

Raykov, T., Tomer, A., & Nesselroade, J. R. (1991). Reporting structural equation modeling results in *Psychology and Aging*: Some proposed guidelines. *Psychology and Aging, 6*, 499–503.

Reboussin, D. M., & DeMets, D. L. (1996). Exact permutation inference for two sample repeated measures data. *Communications in Statistical Theory and Methods, 25*, 2223–2238.

Richardson, A., & Welsh, A. (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics, 51*, 1429–1439.

Roberts, J. K., & McLeod, P. (2008). Software options for multilevel models. In A. A. O'Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 427–468). Charlotte, NC: Information Age Publishing.

Roy, J., & Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: Changes in methadone treatment practices. *Journal of the American Statistical Association, 97*, 40–52.

SAS Institute Inc. (2000). *SAS/Proc MIXED* (Version 8) [Computer program]. Cary, NC: SAS Institute Inc.

Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6*, 461–464.

Seltzer, M., Novak, J., Choi, K., & Lim, N. (2002). Sensitivity analysis for hierarchical models employing t level one assumptions. *Journal of Educational and Behavioral Statistics, 27,* 181–222.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods and Research, 22,* 342–363.

Spybrook, J. (2008). Power, sample size, and design. In A. A. O'Connell and D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273–314). Charlotte, NC: Information Age Publishing.

Stapleton, L. M., & Thomas, S. L. (2008). The use of national datasets for teaching and research. In A. A. O'Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 11–58). Charlotte, NC: Information Age Publishing.

Stern, S., & Welsh, A. (2000). Likelihood inference for small variance components. *Canadian Journal of Statistics, 28,* 517–532.

Tate, R. (2004). Interpreting hierarchical linear and generalized linear models with slopes as outcomes. *Journal of Experimental Education, 73,* 71–95.

Teuscher, F., Herrendorfer, G., & Guiard, G. (1994). The estimation of skewness and kurtosis of random effects in the linear model. *Biometrical Journal, 36,* 661–672.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174–195.

Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18,* 325–346.

Verbeke, G. (1997). Linear mixed models for longitudinal data. In G. Verbeke & G. Molenberghs (Eds.), *Linear mixed models in practice: A SAS-oriented approach* (pp. 63–153). New York: Springer.

Wainer, H. (1997). Some multivariate displays of NAEP results. *Psychological Methods, 2,* 34–63.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10,* 395–426.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics—Simulation, 22,* 1079–1106.

Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society-A, 159,* 201–212.

Yu, Q., & Burdick, R. (1995). Confidence-intervals on variance components in regression-models with balanced (Q–1)-Fold nested error structure. *Communications in Statistics -Theory and Methods, 24,* 1151–1167.

Zucker, D., Lieberman, O., & Manor, O. (2000). Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. *Journal of the Royal Statistical Society-B, 62,* 827–838.