22
Multilevel Modeling

D. Betsy McCoach

In the social sciences, a large proportion of our data are hierarchical in nature. Examples of naturally occurring hierarchies include students nested within schools, patients nested within hospitals, workers nested within companies, husbands and wives nested within couple dyads, and observations within people. Most traditional statistical analyses assume that observations are independent of each other. The assumption of independence means that subjects' responses are not correlated with each other. This assumption might be reasonable when data are randomly sampled from a large population. However, when people are clustered within naturally occurring organizational units (e.g., schools, classrooms, hospitals, companies), the responses of people from the same cluster are likely to exhibit some degree of relatedness with each other, given that they were sampled from the same organizational unit. Multilevel modeling techniques allow researchers to adjust for and model this non-independence.

Multilevel models are also often referred to as *hierarchical linear models, mixed models, mixed effects models*, or *random effects models*. These terms are generally used interchangeably, although there are slight differences in the meanings of the terms. For instance, hierarchical linear model is a more circumscribed term than the others, as it assumes a normally distributed response variable. In contrast, mixed effects or random effects models are the most general terms, as they denote non-independence within a data set, but that non-independence does not necessarily need to be hierarchically nested. For instance, cross-classified random effects account for non-independence that is crossed, rather than nested. In longitudinal educational studies, students often change teachers or schools. Therefore, each student is crossed within a particular combination of teachers or schools. In such scenarios, students are *cross-classified* by two teachers, and in some cases, by two schools. This chapter focuses specifically on multilevel models, or models that exhibit a purely hierarchical data structure.

With clustered data, traditional statistical analyses that assume independence produce incorrect standard errors. In such a scenario, the estimates of the standard errors are smaller than they should be. Therefore, the Type I error rate is inflated for all inferential statistical tests that make the assumption of independence. In multilevel analyses, we explicitly estimate and model the degree of relatedness of observations within the same cluster, thereby correctly estimating the standard errors and eliminating the problem of inflated Type I error rates.

The advantages of multilevel modeling, however, are not merely statistical in nature. Multilevel analyses allow us to exploit the information contained in cluster samples to explain both the between- and within-cluster variability of an outcome variable of interest. These models allow us to use predictors at both the individual (or lowest) level (level 1), and the organizational (or higher) level (level 2) to explain the variance in the dependent variable. We can also allow the relation between an independent variable and the dependent variable to randomly vary across clusters. If we find that the impact of the independent variable on the dependent variable varies across clusters, we can try to explain the variability in this relation using cluster-level variables. For example, we can allow the relation between students' SES and achievement to vary by school. If we find that this relation does vary by school, we can try to explain that variability using school-level predictors, such as type of school, school SES, or average per-pupil expenditures. If a level 2 variable, such as average per-pupil expenditure, moderates the relation between a level 1 variable (SES) and the dependent variable (achievement), this is called a *cross-level interaction*. Thus, multilevel modeling allows us to simultaneously model the impact of both individual (or lower-level) and institutional (or higher-level) variables on the dependent variable of interest, as well as to model the cross-level interactions between higher-level and lower-level variables on the outcome of interest. Such analyses allow us to ask and answer far more nuanced questions than are possible within traditional regression analyses.

Finally, growth curve and other longitudinal analyses can be reframed as multilevel models, in which observations across time are nested within individuals. Using this framework, we can partition residual or error variances into those that are within-person and those that are between people. In such a scenario, between-person residual variance represents between-person variability in any randomly varying level 1 parameter of interest, such as the intercept (which is commonly centered to represent initial status in growth models) and the growth slope.

Contemporary expositions of multilevel modeling include textbooks by Raudenbush and Bryk (2002), Hox (2010), and Snijders and Bosker (2012), and an edited volume by O'Connell and McCoach (2008). Table 22.1 presents specific desiderata for applied studies that utilize multilevel modeling, and the remainder of this chapter is devoted to the explication of these desiderata.

Table 22.1 Desiderata for Hierarchical Linear Modeling.

Desideratum	*Manuscript Section(s)**
1. Model theory and variables included in the model are consistent with the purposes of the study and the research questions or study hypotheses.	I
2. The decision to include/exclude random effects should be justified theoretically. The number of random effects to include should be as realistic and yet as parsimonious as possible. If random effects are eliminated during the model-building process, this decision should be justified both empirically and theoretically.	M, R
3. Statistical model is presented, preferably using equations. Otherwise, minimally, the statistical model is described in enough verbal detail to be replicable by other researchers, and for the reader to determine the fixed effects and the random effects at each level for each model.	M
4. Sample size is specified at each level, and is sufficient for conducting the proposed analysis. Sampling strategy and mode(s) of data collection are identified and justified. If appropriate, weighting methods are described and justified.	M
5. Measurement of the outcome/response variable is described and justified. Measurement of all explanatory variables is described and justified; evidence of reliability and validity is provided.	M

(continued)

Table 22.1 *(continued)*

Desideratum	*Manuscript Section(s)**
6. Scaling and centering of predictor variables are described and justified. Coding of all categorical predictors is fully described. Special attention must be paid to the centering/coding of all lower-level independent variables and to the implications of these centering decisions for interpretation of the model results.	M
7. Extent of missing data is clearly reported for all variables at all levels, and methods for accommodating missing data are described. The final analytical sample is described.	M, R
8. For longitudinal models, the shape of the growth trajectory is described, and the modeling of this trajectory is described and justified.	M, R
9. The software or program, including version number, used to run the models should be identified. Parameter estimation strategy (e.g., REML, ML) is identified and justified.	M, R
10. Assumptions of the model are described and checked. This may include discussions of normality, outliers, multicollinearity, homogeneity or heterogeneity of variances, and residual diagnostics.	M, R
11. The assumed error covariance structure should be described, and any plausible alternative error covariance structures should be described and tested. This is especially important for longitudinal models.	M, R
12. Descriptive statistics for variables at each level of the analysis should be reported. These should include means, standard deviations, and correlations.	R
13. The intraclass correlation coefficient for the unconditional model should be reported and interpreted.	R
14. Generally, multilevel models are built sequentially, using a series of models: an unconditional model, a random coefficients model (containing lower-level predictors), and a full model (containing predictors at all level of the analysis). This series of models is described.	M, R
15. The write-up includes a table that presents the results of the analysis. These results should include both fixed effect parameter estimates and variance component estimates.	R
16. Model fit issues are addressed. Deviance is reported for any estimated models. Additionally, other measures of model fit (e.g., AIC, BIC) are reported for all estimated models. Competing nested models are compared using the likelihood ratio/chi-square difference test.	R
17. Some description/summary of the final model's predictive ability should be provided. This could include a proportion reduction in variance at each level / proportion of variance accounted for at each level. In models with random slopes, proportion reduction in variance measures can be misleading, as the predictive ability of the model varies by cluster.	R
18. Some measure of effect size or practical importance should be reported for the targeted coefficients of interest.	R, D
19. Language used in the presentation and discussion of results appropriately reflects the study design. Causal language is not used except when justified through study design.	R, D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Model/Theory Alignment

When using modeling techniques such as multilevel modeling, a coherent conceptual base should inform and guide the statistical analyses. The model theory and the variables included in the model need to be consistent with the purposes of the study and the research questions or study hypotheses. Of course, any study should be guided by a coherent theory. However, given that excluding an important potential confounder creates the potential for bias in the estimates, the temptation with

regression-type models is to try to add any variable that might be related to the outcome variable. While it is true that failing to include important potential confounders can create bias in the estimates of the effects of other variables, given the complexity of the error structure and the number of potential cross-level interactions, models that include large numbers of fixed and random effects can become unwieldy, difficult to interpret, and perhaps even impossible to estimate. Therefore, researchers should spend a great deal of time determining the variables for inclusion based on theory and relevant literature prior to undertaking the data analysis.

2. Random Effects

As in multiple regression (see Chapter 23, this volume), we estimate a within-cluster residual (r) that represents the deviation of a person's score from his or her predicted value. In a multilevel model, the intercept and the slopes for each of the level 1 variables can randomly vary across the level 2 units. In general, we allow the intercept to randomly vary across level 2 units. Therefore, we estimate a residual for each cluster (u_0). This is the deviation of a cluster's value from the overall intercept. It is this ability to partition variance into within-cluster variance and between-cluster variance that is the essence of the multilevel model. For simplicity, imagine a model in which there are no predictors. Each person's score on the dependent variable is composed of three elements: the overall mean (γ_{00}), the deviation of the cluster mean from the overall mean (u_{0j}), and the deviation of the person's score from his/her cluster mean (r_{ij}). The u_0 term allows us to model the dependence of observations from the same cluster because u_{0j} is the same for every student within school j (Raudenbush & Bryk, 2002). The u_0 term is referred to as a *random effect* for the intercept because we assume that the value of u_0 randomly varies across the level 2 units (clusters). We also assume that u_0 is normally distributed with a mean of 0 and a variance of τ_{00}.

Now, imagine a model in which there is one predictor at the lowest level. For this example, assume that we are predicting reading achievement (Y_{ij}) using socio-economic status (SES). We continue to allow the intercept to randomly vary across schools. However, now we can allow the SES slope to randomly vary across schools as well by including u_1. By allowing the SES slope to randomly vary across schools, we are specifying a model in which the relation between SES and reading achievement is different for different schools. Therefore, in some schools, there could be no relation between students' SES and their reading achievement, whereas in other schools the relation between students' SES and their reading achievement could be quite strong. The set of equations for this model is

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(SES)_{ij} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{1}$$

Generally speaking, in multilevel modeling, a fixed effect represents the average effect across an entire population and is expressed by the regression coefficient (Snijders, 2005). In contrast, a random effect varies randomly across the population of level 2 units and is estimated as a residual for each of level 2 units (Snijders, 2005). Multilevel techniques allow us to model, estimate, and test the variances (and covariances) for these random effects. The variances and covariances of the random effects are referred to as *variance components*. In the set of equations above (1), the γ terms are the *fixed effects* and the u terms are the *random effects*.

In a two-level model, the number of possible random effects is equal to the number of variables at level 1 plus 1 (the random effect for the intercept). Therefore, in a model that contains 10 different level 1 variables, there could be up to 11 random effects. The number of random effects included should be as

realistic and yet as parsimonious as possible. At first glance, it might seem desirable to try to allow the slopes for all level 1 variables to vary randomly across the level 2 clusters and then to eliminate empirically any random effects that are not statistically significant. However, Raudenbush and Bryk (2002, p. 256) cautioned against this practice: "If one overfits the model by specifying too many random level 1 coefficients, the variation is partitioned into many little pieces, none of which is of much significance." Instead, researchers should make the decision to include/exclude random effects based on theoretical grounds, rather than blindly allowing all level 1 slopes to vary randomly across level 2 clusters.

Even when using theory as a guide, sometimes analysts make changes to the random portion of the model during the model-building process. If random effects are eliminated during the model building process, this decision should be justified both empirically and theoretically. One common reason for eliminating random effects is that the level 2 variables in the model are able to explain the between-cluster variability in the slopes. For example, imagine a model in which students' SES is a positive predictor of reading achievement and the SES slope randomly varies across schools. However, when the level 2 model includes the percentage of students within the school who are eligible for free lunch, the between-school variability in the SES slope is greatly reduced, and it is no longer statistically significant. In such a scenario, the between-school variability in the relation between SES and reading achievement is explained by a school-level variable: the percentage of students within the school who are eligible for free lunch. If the fixed effect for this cross-level interaction is negative, schools with greater percentages of students who are eligible for free lunch have less positive SES/reading slopes. If the fixed effect for this cross-level interaction is positive, then schools with larger percentages of free lunch students have larger, more positive SES/reading achievement slopes. In such a scenario, the slope of vocabulary on reading achievement is neither fixed nor randomly varying. Instead, it systematically varies as a function of the two level 2 variables.

3. Presentation of the Statistical Model

It is important for readers to be able to understand and potentially replicate the reported multilevel analyses. Therefore, the full hierarchical linear model must be specified clearly within the Methods section. There are many decisions that a researcher must make when building a multilevel model. Are the slopes of the level 1 coefficients allowed to vary randomly across the level 2 units? Which cross-level interactions between level 1 variables and level 2 variables are specified? Given the complexity of most multilevel models, the clearest and easiest way to communicate the exact specification of the model is to present the statistical model using equations. The equations for the multilevel model can be presented in one of two ways: using separate equations for the level 1 and level 2 variables or using a combined model.

To illustrate the multilevel and combined specifications, imagine a model in which the researcher wants to predict the reading achievement scores for students nested within schools. The level 1 independent variable is socio-economic status (SES), and the effect of SES is assumed to vary randomly across schools. The level 1 intercept is also allowed to vary randomly across schools. The level 2 independent variable is percentage of students receiving free-lunch (FREELNCH), which serves as an indicator of School SES. The multilevel, multiple equation notation is:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(FREELNCH)_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(FREELNCH)_j + u_{1j} \end{aligned}, \tag{2}$$

The γ_{00} term is the intercept of the school intercepts, indicating the predicted value of reading achievement when all other variables in the model are held constant at 0. The γ_{01} term represents

the unit change in the predicted value of the intercept per unit change in the free lunch variable. The γ_{10} term is the intercept of the SES slope, indicating the relation between SES and achievement when FREELNCH = 0. Finally, γ_{11} is the cross-level interaction between FREELNCH and SES, indicating the degree to which the percentage of students within a school who are eligible for free lunch moderates the relation of SES with reading achievement. The u_{0j} term indicates that the intercept (β_{0j}) is allowed to vary randomly across schools. The u_{1j} term indicates that the slope of the SES variable (β_{1j}) is allowed to vary randomly across schools. The combined model is the same model and contains the same information as the multilevel, multiple equation notation. However, in the combined model, we substitute the expressions to the right of the equals sign for β_0 and β_1. Thus, the combined notation for the same model would be:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(FREELNCH) + u_{0j} + \gamma_{10}(SES)_{ij} + \gamma_{11}(FREELNCH)_j(SES)_{ij} + u_{1j}(SES)_{ij} + r_{ij}. \quad (3)$$

Generally, these terms are regrouped so that the fixed effects are in the beginning of the equation and the random effects are at the end of the equation; so, the standard combined form would be as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(FREELNCH)_j + \gamma_{10}(SES)_{ij} + \gamma_{11}(FREELNCH)_j(SES)_{ij} + u_{0j} + u_{1j}(SES)_{ij} + r_{ij}. \quad (4)$$

In reality, the model that is estimated is the combined model. Users of SAS, Stata, R, and SPSS must specify the combined model, whereas users of the software package HLM (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004), for example, may use the multiple equation notation to estimate multilevel models. Thus, certain multilevel modelers prefer to use the combined notation while others prefer the multiple equation notation. Either convention is acceptable, as both sets of equations are equivalent and contain the same information.

There might be audiences who would be confused by the multilevel equations. In such a situation, a reasonable solution is to present the equations and then to explain them verbally within text. Occasionally, a researcher might present research findings to an audience who would be completely overwhelmed by the presentation of equations, and the editor may request that the equations be removed from the manuscript. In such a circumstance, the statistical model should be described in enough verbal detail to be replicable by other researchers based on the description. The reader should be able to determine the fixed effects and the random effects at each level and the cross-level interactions. Some models are so complex that describing them verbally might actually be more difficult than using equations. Even so, the author should make every effort to present his or her multilevel models both verbally and through the use of combined or multilevel equations.

4. Sample Size Issues

Issues related to sample size are critically important in hierarchical linear models. Further, sample size issues are complicated by the multilevel nature of the data. In a sense, there are two salient sample sizes in a two-level model. Consider an organizational model in which people (level 1) are nested within organizations (level 2). The number of individuals represents the level 1 sample size, and the number of organizations represents the level 2 sample size. The number of level 1 units divided by the number of level 2 units provides an estimate of the average cluster size (the average number of level 1 unit within each of the level 2 units.) In a longitudinal model, on the other hand, observations across time are nested within people. Therefore, the number of observations across time (and people) is the level 1 sample size, and the number of people is the level 2 sample size. For example, in a longitudinal model where 100 people are each measured across four time points, the level 1 sample size would be 400 (4 × 100), and the level 2 sample size would be 100.

Generally speaking, the overall sample size is less important than the number of level 2 units and average number of level 1 units within each of the level 2 units. Thus, it is important to report the sample size at each level. At a minimum, the researcher should report the number of level 1 units, the number of level 2 units, and the mean and standard deviation of the average cluster size. If there are a small number of clusters ($N < 50$), including a frequency table that shows the number of observations within each cluster can be useful (Ferron, Hogarty, Dedrick, Hess, Niles, & Kromrey, 2008). In addition, the researcher should identify and justify the sampling strategy and the mode of data collection. When using sampling weights, it is important to describe the method for weighting the data and justify the decision to use sample weights. For more information on the use of sampling weights, see Stapleton and Thomas (2008).

There are two important considerations related to sample size. First, the number of units at each level of analysis must be large enough to estimate the multilevel model. Second, the analysis should be adequately powered to detect the effect of interest.

The total sample size is far less important than the average sample size at each level. Of course, the number of level 1 units can vary greatly from cluster to cluster. The number of level 1 units within clusters places an upper limit on the number of slopes that can randomly vary across clusters. The maximum number of random effects that can be estimated is one fewer than the number of level 1 units within the level 2 clusters; however, it is better to far exceed that minimal criterion. Although small average numbers of level 1 units within level 2 clusters limit the number of random effects that a researcher can estimate, the number of level 2 units is the more important sample size to consider when conducting or evaluating multilevel analyses. The sample size must be large enough to produce estimates, and these estimates must be reasonably free from bias. Maximum likelihood estimation (see Desideratum 9) is a large-sample technique that provides asymptotically unbiased estimates. However, in multilevel modeling, having a large overall sample size is not sufficient. The number of clusters (or the sample size at the highest level) must be large enough to support the estimation technique and to produce relatively unbiased estimates of the parameters and standard errors. What is the minimum number of clusters for a multilevel analysis? Certainly, it seems clear that multilevel analyses require a bare minimum of 10 clusters (Snijders & Bosker, 1999). However, such small sample sizes at level 2 might still produce biased estimates. The number of clusters impacts the estimates of the variance components and the standard errors, as well as the parameter estimates themselves and their standard errors.

Maas and Hox (2005) conducted a series of simulation studies to determine the smallest level 2 sample size that would produce unbiased parameter estimates and standard errors. With only 10 level 2 units, the regression parameters and the level 1 variance components exhibited little bias. However, the level 2 variance components were overestimated by approximately 25%, and the standard errors for all parameter estimates were underestimated, leading to inflated Type I error rates for all statistical hypothesis tests. With at least 30 clusters, the parameter estimates for the regression slopes and both the level 1 and level 2 variance components tended to exhibit very little bias in samples. However, there were issues with the estimation of the standard errors, especially for the variance components. Although the standard errors for the fixed effects and the level 1 variance components seemed to exhibit reasonable coverage with as few as 30 clusters, the standard errors for the level 2 variance components tended to be underestimated when there were fewer than 100 clusters (Maas & Hox, 2005). This means that studies with small to moderate numbers of clusters might have a higher Type I error rate for the level 2 variance components, which could lead to concluding mistakenly that the between-group variance is more pronounced than it actually is. Therefore, while it is possible to produce unbiased estimates of the fixed effects with as few as 10 higher-level units, at least 30 clusters are required to produce unbiased estimates

of the variance components and at least 100 clusters are necessary to have reasonable estimates of the standard errors of the level 2 variance components. In conclusion, while it may be possible to estimate a model with as few as 10 clusters, models with at least 30 clusters should provide reasonable estimates of variance components and standard errors for the fixed effects. However, standard errors for higher-level variance components are likely to be underestimated in studies with small to moderate numbers of clusters and a model comparison approach should be used instead (see Desideratum 16).

Maas and Hox (2005) considered only normal dependent variables. Non-normal (discrete) dependent variables appear somewhat analogous. Using Monte Carlo simulation techniques, Paccagnella (2011) considered the effects of sample size, the number of quadrature points, and the magnitude of the intra-class correlation coefficient accuracy of parameter estimates and standard errors of estimates in logistic multilevel models using Gaussian quadrature estimation. His results generally parallel those of Maas and Hox (2005): as in the continuous case: estimates of the fixed effects are unbiased, even with small sample sizes. Generally, the standard errors of the fixed effect estimates were well estimated with as few as 50 clusters. The estimation of the variance components and their standard errors was more problematic. The estimates of the variance components are underestimated, although the magnitude of this bias decreased substantially as the number of clusters increased. The standard errors of the variance components also exhibited downward bias, even with very large sample sizes. Increasing the number of quadrature points helped to mitigate the bias variance components.

Of course, an additional sample size consideration involves statistical power and precision: the number of level 1 and level 2 units must be large enough to detect the effect of interest. In the simplest scenario, power in multilevel modeling is a function of the number of clusters, the number of units per cluster, the intraclass correlation coefficient (see Desideratum 13), and the effect size. Although increasing sample size at either level increases power, in general, increasing the number of clusters boosts statistical power much more than increasing the average number of units per cluster does. This effect is even more pronounced as the intraclass correlation increases. Several free software programs are available to conduct a priori power analyses for multilevel models. These include the Optimal Design software program (Spybrook, Raudenbush, Liu, Congdon, & Martinez, 2011; http://hlmsoft.net/od) and the Power Up software and program (Dong & Maynard, 2013; www.causalevaluation.org).

5. Measurement Issues

As with any analysis, it is important to describe the scale of measurement of the outcome variable. Hierarchical linear models are appropriate for analyzing continuous, normally distributed outcome variables whereas hierarchical *generalized* linear models allow for the estimation of non-normal response variables (O'Connell, Goldstein, Rogers, & Peng, 2008; Raudenbush & Bryk, 2002).

In addition, the Methods section should include a description of the scale of measurement for all of the explanatory variables in the model. As with any statistical analysis, the researcher should provide evidence of reliability and validity of each of the variables in the model. Because multilevel modeling is a regression-based technique, the assumptions of linear regression models (aside from the assumption of independence, which applies at each level) continue to apply (see Chapter 23, this volume). One commonly overlooked and rarely satisfied assumption of linear regression is that the independent variables are measured with perfect reliability. When one or more predictor variables are measured with error, the regression coefficients are likely to be biased: such biases can result in misleading inferences. Therefore, it is especially important to provide evidence of reliability of scores for all of the continuous independent variables in the model.

6. Centering

In multilevel modeling, it is especially important to describe and justify the scaling and centering of all the predictor variables. Decisions about centering impact the interpretation of the parameter estimates. Centering decisions are especially important for the lower-level independent variables because the choice of centering at the lower level(s) impacts the interpretation of both the lower- and higher-level parameter estimates. For organizational models, the two main centering techniques for lower-level independent variables are *grand mean centering* and *group mean centering*. In grand mean centering, the overall mean of the variable is subtracted from all scores. Therefore, the new score captures a person's standing relative to the full sample. In group mean centering, the cluster mean is subtracted from the score for each person in that cluster. As such, the transformed score captures a person's standing relative to his or her cluster. Whereas grand mean centering is a simple transformation of the raw score, group mean centering is not. There is some debate within the multilevel literature about whether grand mean centering or group mean centering is preferable from a statistical point of view. However, most experts in multilevel modeling agree on three issues related to centering. First, the decision to use grand mean or group mean centering should be based on substantive reasons, not just statistical ones. For instance, if the primary research question involves understanding the impact of a level 2 variable on the dependent variable and the level 1 variables serve as control variables, grand mean centering may be the most appropriate choice. On the other hand, when level 1 variables are of primary research interest, group mean centering may be more appropriate. This is because group mean centering removes between cluster variation from the level 1 covariate and provides an estimate of the pooled within cluster variance (Enders & Tofighi, 2007). Second, it is important to explain the centering decision and procedures and to interpret the parameter estimates accordingly. Third, when using group mean centering, it is important to introduce an aggregate of the group mean centered variable (or a higher-level variable that measures the same construct) into the analysis. Without an aggregate or contextual variable at level 2, all of the information about the between-cluster variability is lost. See Enders and Tofighi (2007) for an excellent discussion of centering in organizational multilevel models.

In growth models, the time or age variable also needs to be centered so that the intercept represents an interpretable value. For linear growth models, the most common technique is to center time at initial status or age at the beginning of the study. When time is centered at initial status, then the intercept represents an individual's starting value. However, the time variable can be centered at any point in the data collection period. For certain research problems, analysts may prefer to center time at the final time point or at the middle of the data collection cycle. As Biesanz and colleagues stated:

> The choice of where to place the origin of time has to be substantively driven. Because this choice determines that point in time at which individual differences will be examined for the lower order coefficients, the answer to which coding(s) of time to examine in detail lies with the researcher's specific substantive questions of interest.
>
> (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004, p. 37)

Again, various options for centering time are appropriate and interpretable; however, it is incumbent upon the researcher to describe the centering procedure, the rationale for selecting the procedure, and the correct interpretation of the parameters, given the chosen centering procedure.

In addition to describing the centering and scaling of continuous variables, it is also important to describe the coding of all categorical predictors. Researchers should use the same conventions that they would use when conducting multiple regression to code the categorical variables in their models. Thus, researchers should use dummy coding, weighted or unweighted effects coding, or

contrast coding for all categorical variables (see Cohen, Cohen, West, & Aiken, 2003, for an excellent discussion of coding for multiple regression analyses). The decision about the type of coding scheme should be conceptually driven and result in easily interpretable parameter estimates. Again, researchers should describe the chosen coding scheme used and explain the appropriate interpretations of the parameter estimates that result from such a coding scheme. Finally, researchers need to consider the necessity of and model all same-level interactions among categorical and/or continuous variables in the same manner as they would if they were conducting a multiple regression analysis. The interpretation of the same level interaction parameter estimates depends on the coding schemes used for the lower-order variables. Often it is easiest to describe such interactions visually, using figures. Alternatively, creating tables of prototypical predicted values for different types of participants may help to illustrate such interaction effects. See Aiken and West (1991) for an excellent discussion of creating and interpreting same-level interactions within a multiple regression framework.

7. Missing Data

The percentage of missing data should to be reported for all variables at all levels, and the author should describe the methods used to address the issue of missing data. Missing data are a problem for any analysis. However, in multilevel modeling, dealing with missing data can be especially complex. First, in most commercial multilevel software programs, units with missing data on any of the covariates are eliminated from the analysis by default. This becomes especially problematic when higher level units have missing data on any covariates, as the deletion of one higher level unit could result in the loss of tens, hundreds, or even thousands of lower level units, depending on the within cluster sample size of that higher level unit. For example, any school with missing data on any of the school-level covariates (e.g., percentage of free lunch eligible students, average per pupil expenditures) is eliminated from the multilevel model. Thus, it is easy to see how even small amounts of missing data at the higher levels of analysis could drastically reduce the size of the sample as well as the generalizability of the results.

Several modern data techniques exist for dealing with the problem of missing data. *Multiple imputation* (MI; Rubin, 1987, 1996) and full maximum likelihood estimation (FIML) (Enders, 2010) are generally considered two of the best methods of dealing with missing data. The use of multiple imputation has become increasingly common. When using MI with clustered data, there is one important caveat: multiple imputation of either the dependent variable or lower-level covariates should take the clustered nature of the data into account (Black, Harel, & McCoach, 2011). Standard multiple imputation procedures assume that the observations are independent. Using normal theory/standard approach does not provide valid inferences about variance components with any amount of missing data, and it does not provide reasonable estimates for fixed effects with high rates of missing data (Black et al., 2011). Although listwise deletion is generally considered a less desirable method of dealing with missing data than multiple imputation, there is some evidence to suggest that listwise deletion outperforms standard multiple imputation in terms of recovering parameter estimates when the data are multilevel in nature in some conditions (Black et al., 2011), however more advanced approaches would be prefer (Hox, van Buuren, & Jolani, 2016). When describing the sample, the author should explicitly describe the amount of missing data and justify his or her method of handling missing data.

8. Fitting Growth Trajectories

Fitting longitudinal growth models using hierarchical linear modeling techniques is becoming increasingly popular. In such a model, observations across time (at level 1) are nested within people (at level 2). Then standard unconditional linear growth model is

$$
\begin{aligned}
Y_{ti} &= \pi_{0i} + \pi_{1i}(TIME)_{ti} + e_{ti} \\
\pi_{0i} &= \beta_{00} + r_{0i} \\
\pi_{1i} &= \beta_{10} + r_{1i}
\end{aligned} \tag{5}
$$

The dependent variable (y_{ti}) is the score for student *i* at time *t*, which is a function of the randomly varying intercept, π_{0i} (which is the predicted value of y_{it} when time=0), the randomly varying growth slope, π_{1i} (time_{ti}), and time-specific individual error.

Both the slope and the intercept contain a subscript *i*, indicating that a separate slope and intercept are estimated for each person in the sample. The deviation of a particular observation from the model-predicted trajectory is captured in the error term, (e_{ti}) which represents the within-person error associated with that individual's data at that time point. The pooled error variability within individuals' trajectories is estimated by the variance of e_{ti}[var(e_{ti}) = σ^2] (Raudenbush & Bryk, 2002), and this error variance is generally assumed to be constant across time.

Because the time slope, π_{1i}, enters the equation as a predictor of the outcome value at a given occasion, each participant can have his/her own unique data collection schedule. Therefore, multilevel models seamlessly handle time unstructured data. Centering the time variable around some meaningful value within the data collection period helps to ensure the interpretability of the intercept and the variance/covariance components. In longitudinal studies, time is often centered at the beginning of the study period so that the intercept represents the expected value at the beginning of the study. If age is used as the time variable, it is quite common to center at a particular age (for example, at age 6). This strategy has the added advantage of controlling for age in addition to centering time. For example, if we center time at age 6, then the intercept represents the model predicted score at age 6.

The level 2 equations model the average growth trajectory across people and can capture between-person differences in the model-implied growth trajectories based on level 2 (time invariant) covariates. The second level of the multilevel model specifies that the randomly varying intercept (π_{0i}) for each individual (*i*) is predicted by an overall intercept (β_{10}), the effects of any level 2 predictors on the intercept, and , the level 2 residual, which represents the difference between person *i*'s model predicted intercept (based on the overall intercept, β_{10}, and level 2 predictors) and his or her actual intercept. Likewise, the randomly varying linear growth slope (π_{0i}) for each individual (*i*) is predicted by an overall intercept (β_{10}), the effects of level 2 variables on the linear growth slope, and r_{1i}, the level 2 residual, which represents the difference between person *i*'s model predicted linear growth slope and his or her actual growth slope. The inclusion of the r_{0i} and r_{1i} in the level 2 equations allows for between-person variability in the intercepts and slopes. If the intercept is centered around initial status, then the variance in r_{0i} (τ_{00}) represents the between-person variability in initial status, or where people start. Likewise, the variance in r_{1i} (τ_{11}) represents the between-person variability in peoples' growth rates. The standardized covariance of the two level 2 residuals, τ_{01} from the unconditional linear growth model provides the correlation between initial status (or, more generally, the intercept) and growth.

As the name implies, a linear growth model assumes a straight-line growth trajectory. However, many growth processes do not follow a linear trajectory. Assuming a linear growth trajectory is very limiting, and it may result in a serious misspecification of the model. Other shapes are accommodated easily using a variety of strategies. These include estimating piecewise models, polynomial models, or other non-linear models, as well as introducing time-varying covariates (McCoach & Kaniskan, 2010; McCoach, Madura, Rambo, O'Connell, & Welsh, 2013; McCoach & Yu, 2016; Singer & Willett, 2003). Therefore, the researcher should empirically examine the shape of the individual and average growth trajectories descriptively prior to fitting any statistical models.

This information, in combination with the theory, can help guide decisions about the shape of the growth trajectory. When using multilevel modeling to fit longitudinal models, it is imperative that the researcher describe the shape of the growth trajectory, describe the level 1 model, and justify how the modeling procedure used at level 1 was able to capture the shapes of the growth trajectories for the sample.

9. Software and Parameter Estimation

The methods section should include the program or software package and version used to conduct the analysis. Many general purpose statistical software packages such as R, SPSS, SAS, and Stata have multilevel capabilities. In addition, specialized multilevel software programs such as HLM, MLwin and latent variable modeling programs such as Mplus and LISREL are popular choices for estimating multilevel models. All of these programs handle straightforward two-level models with normal response variables with ease. Where the programs differ is in their ability to handle more complicated models such as cross-classified models, three-level models, multilevel meditational models, or models with non-normal outcome variables. For an overview and comparison of these different software programs, see McCoach *et al.* (2018) and Roberts and McLeod (2008), as well as the reviews provided at the Multilevel Centre (www.bristol.ac.uk/cmm/learning/mmsoftware).

The two most common estimation techniques for hierarchical linear models with normal response variables are *maximum likelihood* (ML) and *restricted maximum likelihood* (REML). The two methods should produce similar results in terms of the fixed effects (regression parameters); however, they do produce different estimates of the variance components (Snijders & Bosker, 1999). In ML estimation the estimates of the variance and covariance components are conditional upon the point estimates of the fixed effects, whereas in REML they are not (Raudenbush & Bryk, 2002). Thus whereas REML estimates of variance-covariance components adjust for the uncertainty about the fixed effects, ML estimates do not. When estimating the variance components, REML takes "into account the loss of degrees of freedom resulting from the estimation of the regression parameters, whereas the ML method does not" (Snijders & Bosker, 1999, p. 56). When the number of clusters is very large, REML and ML results should produce similar estimates of the variance components. However, when the number of level 2 units is relatively small, the ML estimates of the variance components (τ_{qq}) are underestimated by a factor of $(J - F)/J$, where J is the number of level 2 units and F is the number of fixed effects. Therefore, REML is the preferred estimation strategy for models with relatively few level 2 units.

While REML may be preferable to ML for estimating the variance components, ML is often preferable to REML for testing model fit. The deviances of any two nested models that differ in terms of their fixed and/or random effects can be compared when using ML. In contrast, REML only allows for comparison of nested models that differ in their random effects (Snijders & Bosker, 1999, p. 89). In addition, information criteria, such as the AIC and BIC, should be based on the ML estimates of the deviance (see Desideratum 16 for information about deviance and model fit.)

For binary or ordinal response variables, the most common estimation techniques use quasi-likelihood estimators such as penalized quasi-likelihood (PQL) or Maximum likelihood approaches using Gauss–Hermite quadrature, adaptive quadrature, or Laplace algorithms (Bauer & Sterba, 2011). Although penalized quasi-likelihood tends to be faster, especially in models with large numbers of random effects, it does not produce a deviance statistic that can be used to compare competing models. However, both PQL and ML approaches using adaptive quadrature appear to perform well for binary and ordinal models (Bauer & Sterba, 2011).

10. Assumptions and Residual Analyses

As with any statistical analysis, it is important to check the assumptions of the model and to describe any violations of the assumptions. Many regression diagnostics for single-level models are applicable within the multilevel framework as well. These may include discussions of normality, linearity, outliers, multicollinearity, homogeneity or heterogeneity of variances, and residual diagnostics. However, because the regression model is operating on multiple levels, tests of the assumptions become a bit more complex and time consuming.

O'Connell, Yeomans-Maldonado, and McCoach (2016) recommend a three-stage approach to conducting residual analyses within a multilevel framework. Stage 1 focuses on the level 1 model and residuals, and includes checking assumptions such as homogeneity of level 1 variance, linearity, and normality using statistics and graphical displays including histograms and box plots, graphs of predicted means versus residuals, and normal probability plots. Stage 2 focuses on exploring the level 2 residuals using the level 1 model/residuals at level 2. In addition, stage 2 includes examining for influential, outlying or unusual level 2 units. Stage 3 examines the residuals at level 1 and level 2 using the level 2 model (O'Connell et al., 2016).

Most residual analyses can and should be conducted at each level of the analysis. For example, in a two-level model where, say, students are nested within schools, it is possible to have an outlier at the student level or at the school level. Researchers should carefully check the assumptions of their models, and they should include a short description of the procedures that they used to check their assumptions. In addition, they should describe any violations of the assumptions and the procedures that they used to rectify those violations (e.g., Were any outliers deleted? Were any variables transformed?).

11. Error Covariance Structure

The researcher should briefly describe the assumed error covariance structure. Any plausible alternative error covariance structures should be described and tested. Generally, the assumed error covariance structure is quite reasonable for organizational models. The simplest error structure for a two-level model with a random intercept is depicted in equation (6). In this matrix, there are as many rows and columns as there are level 1 units. In this example, the first six level 1 units are shown. The first three level 1 units belong to cluster 1 and the second three units belong to cluster 2. The total residual variance for each person in the model is the sum of the within cluster residual (σ^2) and the between-cluster residual (τ_{00}). The covariance between any two people who are members of the same cluster is accounted for by τ_{00} , the between cluster residual. Finally, the residual covariance between 2 members of two different clusters is assumed to be 0.

$$\begin{bmatrix} \sigma^2+\tau_{00} & \tau_{00} & \tau_{00} & 0 & 0 & 0 & \\ \tau_{00} & \sigma^2+\tau_{00} & \tau_{00} & 0 & 0 & 0 & \\ \tau_{00} & \tau_{00} & \sigma^2+\tau_{00} & 0 & 0 & 0 & \\ 0 & 0 & 0 & \sigma^2+\tau_{00} & \tau_{00} & \tau_{00} & \cdots \\ 0 & 0 & 0 & \tau_{00} & \sigma^2+\tau_{00} & \tau_{00} & \\ 0 & 0 & 0 & \tau_{00} & \tau_{00} & \sigma^2+\tau_{00} & \\ \cdots & & & & & & \end{bmatrix} \tag{6}$$

Describing the error covariance structure is especially important for longitudinal models. Models that fail to adequately account for the covariances among repeated measurements may result in

misleading inferences (Fitzmaurice, Laird, & Ware, 2004). On the other hand, when modeling these longitudinal covariances, the analyst's goal should be to "select the most parsimonious covariance structure that reasonably fits the data" (Wolfinger, 1996, p. 208).

The standard multilevel linear growth model imposes a very particular structure on the composite within-person/across time covariances (the composite of the covariances across waves). The structure is dependent on the number of random effects in the model. The maximum number of random effects that can be estimated in a repeated measures model is the number of waves of data minus 1. The standard multilevel linear growth model estimates a random effect for the intercept, a random effect for the linear growth slope, and a covariance between the intercept and the slope. Using the standard multilevel model, the model-implied variance-covariance matrix for a model with four waves of data is

$$\begin{bmatrix} \tau_{00}+\sigma^2 & & & \\ \tau_{00}+\tau_{01} & \tau_{00}+2\tau_{01}+\tau_{11}+\sigma^2 & & \\ \tau_{00}+2\tau_{01} & \tau_{00}+3\tau_{01}+2\tau_{11} & \tau_{00}+4\tau_{01}+4\tau_{11}+\sigma^2 & \\ \tau_{00}+3\tau_{01} & \tau_{00}+4\tau_{01}+3\tau_{11} & \tau_{00}+5\tau_{01}+6\tau_{11} & \tau_{00}+6\tau_{01}+9\tau_{11}+\sigma^2 \end{bmatrix}. \tag{7}$$

Thus, all 10 unique elements of the variance covariance matrix for the four repeated measurements are estimated using four parameters: τ_{00}, the between person variance in the intercept, τ_{11}, the between person variance in the linear growth slope, τ_{01}, the covariance between the slope and the intercept, and σ^2, the within person residual variance. Other options for estimating the covariance structure of the repeated measurements include fitting models with heterogeneous σ^2 across the time points, first order autoregressive models, first order moving average models, and unrestricted covariance matrices, to name a few. A complete treatment of this topic is beyond the scope of this chapter. However, researchers who are interested in learning more about covariance structures for repeated measures multilevel models should consult Singer and Willett (2003), and Wolfinger (1996).

12. Descriptive Statistics

As in any research study, it is important to provide the reader with tables of descriptive statistics. Minimally, the author should provide a table of means and standard deviations and sample sizes for all of the continuous level 1 variables used in the analysis as well as a table of means and standard deviations, and sample sizes for all of the continuous level 2 variables in the model. Dichotomous variables should be reported as proportions or percentages. In addition, the document should include a table of correlations corresponding to each level in the analysis. So, for a two-level model, one correlation matrix should detail the correlations among the level 1 variables, whereas another correlation table should provide the correlations among the level 2 variables, computed at the cluster level.

13. Intraclass Correlation Coefficient

The *intraclass correlation coefficient* (ICC) is the proportion of variance in the outcome variable that is between clusters, that is, the proportion of variance that can be explained by the clustering or grouping structure (Hox, 2002). Alternatively, one may interpret the ICC as the "expected correlation between any two randomly chosen units that are in the same group" (Hox, 2002, p. 15). The formula for the ICC is

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \tag{8}$$

Where τ_{00} represents between cluster variance and σ^2 represents within cluster variance. The ICC is important to report because it indicates the degree of non-independence in the data. The higher the ICC, the more homogeneity there is within clusters (or the more heterogeneity there is between clusters). An ICC of 0 indicates independence of observations, and any ICC above 0 indicates some degree of dependence in the data. The smaller and more homogeneous the cluster is, the higher the expected ICC is (McCoach & Adelson, 2010). For example, in school effects research, ICCs typically range from .10 to .20. In dyadic research, on the other hand, ICCs above .50 are not uncommon.

The computation of the *design effect*, which indicates the degree to which the parameter estimates' standard errors are underestimated when assuming independence, utilizes the ICC (ρ) and the average number of units per cluster ($\overline{n}_j$):

$$\text{design effect} = \sqrt{1 + \rho\left(\overline{n}_j - 1\right)} \tag{9}$$

Generally, design effects below 2.0 are considered fairly small. However, keep in mind that even with a design effect as low as 1.5, the standard errors in a model that assumes independence of observations are underestimated by a factor of 1.5. Therefore, the Type I error rate is already noticeably inflated, even with such a small design effect (McCoach & Adelson, 2010).

14. Model Building

Generally, multilevel models are built sequentially, using a series of models. First, researchers estimate an *unconditional* (or *null*) *model*, which contains no predictors. The purpose of this model is to obtain estimates of the level 1 and level 2 variance components for comparison to later, more parameterized models and to estimate the ICC. The second model estimated is a *random coefficients model*, which contains the level 1 predictors (Raudenbush & Bryk, 2002). Depending on the researcher's theoretical framework as well as the sample size at level 1, the slopes for some of the level 1 predictors may be estimated as randomly varying across level 2 units, or they can be estimated as fixed across all level 2 units. Raudenbush and Bryk (2002, p. 256) cautioned against the "natural temptation," which "is to estimate a 'saturated' level 1 model . . . where all potential predictors are included with random slopes." Any level 1 slopes that do not have statistically significant variability across level 2 units should be fixed prior to conducting the full contextual analysis. The next model to be estimated is the *full contextual model*, which contains both level 1 and level 2 predictors. Level 2 predictors can be used to predict the intercept or the mean value of the dependent variable (when all of the level 1 variables are held constant at 0). Level 2 predictors also can help explain the variability of level 1 slopes across clusters. In such a scenario, the level 2 variable is used to predict the level 1 slopes, or the relationship of the level 1 predictor and the dependent variable across level 2 units. For example, imagine that SES is a level 1 predictor of math achievement. Sector, a level 2 variable that indicates whether a school is public or private, can be added as a predictor of the relation between SES and math achievement. This cross-level interaction indicates whether sector moderates the relationship between SES relates and math achievement. Finally, if any fixed or random effects are eliminated from the full model, a final contextual model should be estimated and compared to the prior model.

It is important to describe the process of building these sequential models. Analysts differ somewhat in their approaches to building multilevel models. Thus, authors must be sure to describe the model building process in enough detail that another analyst could replicate the entire model and decision sequence.

15. Tables

The Results section should include a table that presents the results of the analyses. If space allows, presenting the results of the entire series of models can be quite informative; however, minimally, the table should include the complete results from the final, full contextual model. These results should include the fixed effect parameter estimates, the random effect parameter estimates (the variances of the random effects), the standard errors for all parameter estimates, and tests of statistical significance for both the fixed and random effects. Also, the table may include covariances among the random effects. Reporting the covariances is expected in longitudinal models; it is less customary to report all covariances among level 2 residuals in organizational models.

16. Deviance and Model Fit

It is important to address model fit issues as part of the model building and testing process. The deviance compares the log-likelihood of the specified model to the log-likelihood of a saturated model that fits the sample data perfectly (Singer & Willett, 2003, p. 117). Specifically, deviance = −2LL, where LL is the log-likelihood of the current model minus the log-likelihood of the saturated model. Therefore, deviance is a measure of the badness of fit of a given model; it describes how much worse the specified model is than the best possible model. Deviance statistics cannot be interpreted directly since deviance is a function of sample size as well as the fit of the model.

When one model is a subset or special case of the other, the two models are said to be "nested" (e.g., Kline, 1998). In nested models, "the more complex model includes all of the parameters of the simpler model plus one or more additional parameters" (Raudenbush, Bryk, Cheong, & Congdon, 2000, pp. 80–81). When two models are nested, their deviance can be compared directly using the chi-square difference test. The deviance of the simpler model (D_1), which has p_1 degrees of freedom, minus the deviance of the more complex model (D_2), which has p_2 degrees of freedom ($p_2 < p_1$), provides the change in deviance ($\Delta D = D_1 - D_2$). As the number of parameters in a model increases, the deviance value decreases. In sufficiently large samples, the difference between the deviances of two hierarchically nested models is distributed as an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters being estimated between the two models (e.g., de Leeuw, 2004).

In evaluating model fit using the chi-square difference test, the more parsimonious model is preferred, as long as it does not result in statistically significantly worse fit. In other words, if the model with the larger number of parameters fails to reduce the deviance by a substantial amount, the more parsimonious model is retained. However, when the change in deviance (ΔD) exceeds the critical value of chi-square with $p_2 - p_1$ degrees of freedom, then the additional parameters have resulted in statistically significantly improved model fit. In this scenario, the more complex model (i.e., with p_1 degrees of freedom) is favored.

Under ML estimation, the number of reported parameters includes the fixed effects (the γ terms) as well as the variance/covariance components. When using REML, the number of reported parameters includes only the variance and covariance components. To compare two nested models that differ in their fixed effects, it is necessary to use ML estimation, not REML estimation. REML only allows for comparison of models that differ in terms of their random effects but have the same fixed effects. Because most programs use REML as the default method of estimation, it is important to remember to select ML estimation to use the deviance estimates to compare two nested models with different fixed effects (McCoach & Black, 2008).

17. Predictive Ability of the Model

In single-level regression models, an important determinant of the utility of the model is the proportion of variance explained by the model, or R^2. Unfortunately, there is no exact multilevel analog

to the proportion of variance explained. Variance components exist at each level of the multilevel model; therefore, variance can be accounted for at each level of the multilevel model. In addition, in random coefficients models, the relation between an independent variable at level 1 and the dependent variable can vary as a function of the level 2 unit or cluster. Consequently, there is no constant proportion of variance in the dependent variable that is explained by the independent variable. Instead, the variance in the dependent variable that is explained by the independent variable varies by cluster. Finally, because the variance components are estimated using ML estimation, the estimation of the variance can differ slightly from model to model. Therefore, it is impossible to compute an R^2 value for the entire model. However, both Raudenbush and Bryk (2002) and Snijders and Bosker (2012) have proposed multilevel analogs to R^2. In both cases, the authors provided two separate formulas: one to explain variance at level 1 and another to explain variance at level 2.

Perhaps the most common statistic used to estimate the variance explained is the *proportional reduction in variance* statistic (Raudenbush & Bryk, 2002). The proportional reduction in variance can be estimated for any variance component in the model. This statistic compares the variance in the more parameterized model to the variance in a simpler baseline model. To compute the proportional reduction in variance, subtract the remaining variance within the more parameterized model from the variance within a baseline model. Then divide this difference by the variance within the baseline model. That statistic is computed

$$\frac{\hat{\sigma}^2_{\mathrm{b}} - \hat{\sigma}^2_{\mathrm{f}}}{\hat{\sigma}^2_{\mathrm{b}}} \tag{10}$$

where $\hat{\sigma}^2_{\mathrm{b}}$ is the estimated level 1 variance for the baseline model and $\hat{\sigma}^2_{\mathrm{f}}$ is the estimated level 1 variance for the fitted model (Raudenbush & Bryk, 2002). At level 2, population variance components estimates are represented by $\hat{\tau}_{qq}$ and are given for the intercepts (β_{0j}) and each slope estimate (β_{1j}, β_{2j}, . . ., β_{qj}) that is allowed to randomly vary across clusters. The proportional reduction in the variance of a given slope, β_{qj}, is

$$\frac{\hat{\tau}_{qq_{\mathrm{b}}} - \hat{\tau}_{qq_{\mathrm{f}}}}{\hat{\tau}_{qq_{\mathrm{b}}}} \tag{11}$$

where $\hat{\tau}_{qq_{\mathrm{b}}}$ is the estimated variance of slope q in the base model and $\hat{\tau}_{qq_{\mathrm{f}}}$ is the estimated variance of slope q in the fitted model.

It should be noted, however, that the proportion reduction in variance statistic does not behave like the familiar R^2. First, the proportional reduction in variance statistic proposed by Raudenbush and Bryk (2002) represents a comparison of one model to another model, and as such it cannot be interpreted as an explanation of the absolute amount of variance in the dependent variable. In addition, the proportion reduction in variance statistic can be negative. This actually happens with some regularity when comparing the level 2 intercept variance of a completely null model (a random effects ANOVA model which includes no predictors at level 1 or level 2) to the level 2 intercept variance of a model that includes a group mean centered predictor at level 1. Finally, it is inappropriate to use this technique to compute the proportion reduction in variance for two models that differ in terms of the number of random slopes being estimated.

The second method of deriving a multilevel R^2 type statistic (Snijders & Bosker, 1994, 1999) produces measures of *proportional reduction in prediction error* for level 1 (the prediction of Y_{ij}) and level 2 (the prediction of $\overline{Y}_{.j}$). These statistics are only available for models that include random intercepts but not for random coefficients models, which include randomly varying slopes. Like the proportional reduction in variance static presented above, the proportional reduction in prediction error for

level 1 (the prediction of Y_{ij}) compares the amount of residual variance in the more parameterized model to a simpler baseline model. However, this formula uses the total estimated variance, $\hat{\sigma}^2 + \hat{\tau}_{00}$, to compare the two models. The rationale is that $\hat{\sigma}^2 + \hat{\tau}_{00}$ provides a reasonable estimate of the total sample variance of the outcome variable *Y* (Snijders & Bosker, 1994). Because $\hat{\sigma}^2 + \hat{\tau}_{00}$ is being used as a proxy for the total variance in the dependent variable, this formula is only appropriate for models without randomly varying slopes. Given a random intercepts only model, the prediction error for individual outcomes (Y_{ij}) is equal to the sum of the level 1 and level 2 variance components, $\hat{\sigma}^2 + \hat{\tau}_{00}$.

The proportional reduction of prediction error at level 1 compares the total residual variance of a fitted (or more parameterized) model, f, to that of a baseline (or less parameterized) model, b. The formula for R_1^2 is

$$R_1^2 = 1 - \frac{(\hat{\sigma}^2 + \hat{\tau}_{00})_f}{(\hat{\sigma}^2 + \hat{\tau}_{00})_b} \tag{12}$$

Where the fraction's numerator is the prediction error for the fitted model and the fraction's denominator is the prediction error for the baseline model.

With respect to level 2, Snijders and Bosker's (1999, p. 103) explained proportion of variance at level 2 is the proportional reduction in the mean squared prediction error for the cluster mean "$\overline{Y}_{.j}$ for a randomly drawn level-two unit *j*." The prediction error for the group mean is

$$\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00} \tag{13}$$

Thus, the level 2 proportional reduction in the prediction error, R_2^2, is

$$R_2^2 = 1 - \frac{\left(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00}\right)_f}{\left(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00}\right)_b} \tag{14}$$

Where the fraction's numerator is the prediction error variance for the fitted model and the fraction's denominator is the prediction error variance for the baseline model. In this case $\overline{n}_j$, is a representative value for average group size.

The various multilevel R^2-type statistics described above provide heuristics to compare models in terms of their ability to "explain variance." However, it is important to remember their shortcomings. First, these estimates do not provide unequivocal estimates of the variance explained by a model. Instead, they compare two models in terms of their ability to reduce some type of variance at one of the levels of the hierarchy. Second, when a model contains random slopes, R^2 does not have a unique definition (Hox, 1998; Kreft, deLeeuw, & Aiken, 1995). The relation between the level 1 predictor and the dependent variable varies across level 2 units, and the level 2 variance estimate is not constant in these models (Snijders & Bosker, 1999). Therefore, the notion of a unitary proportion of variance explained ceases to exist. Finally, these statistics can produce negative estimates, which provides a clear indication that they are not actually proportions of variance explained. However, even given these shortcomings, multilevel R^2 analogs do help researchers to compare predictive ability of various multilevel models. Therefore, they should be reported within the Results section of a multilevel paper. When reporting their R^2 results, researchers should be sure to specify whether

they used Raudenbush and Bryk's (2002) or Snijders and Bosker's (1999) method to compute these proportional reduction in variance estimates, and they also should clearly specify which model they used as the baseline model and which model they used as the fitted (or more parameterized model) for each of their computations.

18. Effect Size

As with any statistical analyses, it is important to report effect size measures for multilevel models. The R^2 analogs described above can help researchers and readers to determine the impact that a variable or a set of variables has on a model. In addition, researchers can compute Cohen's *d*-type effect sizes to describe the mean differences among groups (see Chapter 6, this volume). To calculate the equivalent of Cohen's *d* for a group-randomized study (where the treatment variable occurs at level 2), use the following formula:

$$\delta = \frac{\hat{\gamma}_{01}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}_{00}}} \tag{15}$$

(Spybrook et al., 2011). Assuming the two groups have been coded as 0/1 or −.5/+.5 (or any centering that retains a one point difference between the two groups), the numerator of the formula represents the difference between the treatment and control groups. The denominator utilizes the σ^2 and τ_{00} from the unconditional model. In the unconditional model, the total variance in the dependent variable is divided into two components: the between-cluster variance, τ_{00}, and the within-cluster variance, γ_{01}.

To facilitate understanding among readers, researchers should consider including figures that illustrate cross-level interactions among variables. Just as plotting same level interactions facilitates an understanding of interaction effects (Aiken & West, 1991), similar visual graphics of interactions between two variables at different levels of the data hierarchy can effectively display cross-level moderation. In addition, researchers should include predicted values for prototypical participants. These predicted values also can help the reader to make sense of the magnitude of the effects that are being reported. Thus, they serve as a form of "unstandardized" effect size.

19. Causal Claims

Multilevel modeling solves certain statistical issues that arise from non-independent or clustered data, and it allows for more nuanced analyses of variables that occur at different levels of the hierarchy. However, any causal claims that can be made from a multilevel analysis are determined by the strength of the research design. As Kelloway (1995, p. 216) stated, "No amount of sophisticated analyses can strengthen the inference obtainable from a weak design." It is common to refer to "effects" in multilevel modeling. In fact, the entire lexicon of the technique is replete with references to fixed effects, random effects, cross-level interaction effects, and so forth. However, none of these "effects" should ever be interpreted as indicative of causation or a causal mechanism except under certain randomized designs. When writing the Results and Discussion sections of a multilevel article, researchers should choose their language carefully so as not to imply causal claims that cannot be substantiated or defended given the design of the study.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*, 373–390.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, *9*, 30–52.

Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics*, *38*(9), 1845–1865.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

De Leeuw, J. (2004). Multilevel analysis: Techniques and applications (book review). *Journal of Educational Measurement*, *41*, 73–77.

Dong, N., & Maynard, R. A. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138.

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Charlotte, NC: Information Age Publishing.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley-Interscience.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). New York: Springer Verlag.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. J., van Buuren, S., & Jolani, S. (2016). Incomplete multilevel data. In J. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 39–62). Charlotte, NC: Information Age.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior*, *16*, 215–224.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

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.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86–92.

McCoach, D. B., & Adelson, J. (2010). Dealing with dependence (Part I): Understanding the effects of Clustered Data. *Gifted Child Quarterly*, *54*, 152–155.

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

McCoach, D. B., & Kaniskan, B. (2010). Using time-varying covariates in multilevel growth models. *Frontiers in Quantitative Psychology and Measurement*, *1*, 17.

McCoach, D. B., Madura, J., Rambo, K., O'Connell, A. A., & Welsh, M. (2013). Longitudinal data analysis. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 199–230). Rotterdam: Sense Publishers.

McCoach, D. B., Rifenbark, G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A., & Bellara, A. (2018). Does the package matter? A Comparison of Five Common Multilevel Modeling Software Packages. *Journal of Educational and Behavioral Statistics*, *43,* 594–627.

McCoach, D. B. & Yu, H. H. (2016). Using Individual Growth Curves to Model Reading Fluency. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 269–308). New York: Springer.

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. B. McCoach (Eds.) *Multilevel modeling of educational data.* (pp. 199–244). Charlotte, NC: Information Age Publishing.

O'Connell, A. A., & McCoach, D. B. (Eds.) (2008). *Multilevel modeling of educational data*. Charlotte, NC: Information Age Publishing.

O'Connell, A. A., Yeomans-Maldonado, G., & McCoach, D. B. (2016). Residual diagnostics and model assessment in a multilevel framework: Recommendations toward best practice. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 97–135). Charlotte, NC: Information Age.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *7*(3), 111–120.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2000). *HLM 5: Hierarchical linear and nonlinear modeling*. Statistical software manual. Skokie, IL: Scientific Software International.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Roberts, J. K., & McLeod, P. (2008). Software options for multilevel models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 427–467). Charlotte, NC: Information Age Publishing.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473–489.

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

Snijders, T. A. B. (2005). Fixed and random effects. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (vol. *2*, pp. 664–665). New York: Wiley.

Snijders, T., & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*, 342–363.
Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.
Snijders, T., & Bosker, R. (2012). *Multilevel analysis* (2nd ed.). Thousand Oaks, CA: Sage.
Spybrook, J., Raudenbush, S. W., Liu, X., Congdon, R., & Martinez, A. (2011). Optimal design for longitudinal and multilevel research. V1.*77* [computer software]. Retrieved February 20, 2016 from http://hlmsoft.net/od/
Stapleton, L. M., & Thomas, S. L. (2008). Sources and issues in the use of national datasets for pedagogy and research. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel analysis of educational data* (pp. 11–57). Charlotte, NC: Information Age Publishing.
Wolfinger, R. D. (1996). Heterogeneous variance covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, *1*, 205–230.