1

This chapter provides an introduction to multilevel modeling, including the impact of clustering and the intraclass correlation coefficient. Prototypical research questions in institutional research are examined, and an example is provided to illustrate the application and interpretation of multilevel models.

Hierarchical Data Structures, Institutional Research, and Multilevel Modeling

Ann A. O'Connell, Sandra J. Reed

Introduction

Multilevel modeling (MLM), also referred to as hierarchical linear modeling (HLM) or mixed models, provides a powerful analytical framework through which to study colleges and universities and their impact on students. Due to the natural hierarchical structure of data obtained from students or faculty in colleges and universities, MLM offers many advantages to analysts and policy makers involved in institutional research (IR). This chapter introduces fundamental concepts of hierarchy and its statistical treatment specifically for data structures occurring in IR settings. Our goal is to provide an overview of HLM and set the stage for the chapters that follow as well as highlight the particular advantages of HLM for those involved in IR.

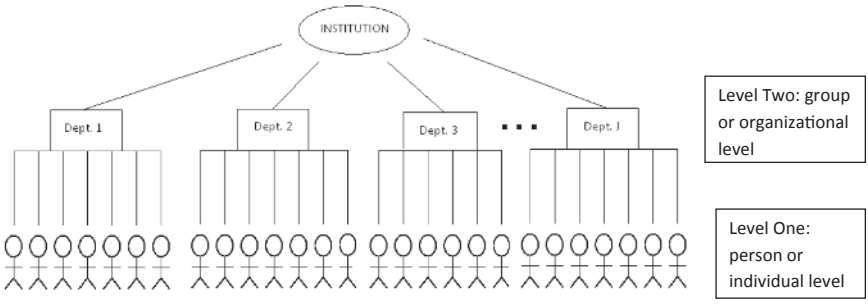
IR professionals routinely encounter the kinds of clustered or nested data structures for which HLM is uniquely suited. Cross-sectional studies of students nested within classes or courses, classes nested within departments or schools, faculty within departments, athletes within sport designations within departments or schools—each of these settings describes lower-level individuals (that is, students or faculty) nested or clustered within one or more higher-level contexts or groups (that is, within classes or within departments). In such cases, the variability in lower-level

New Directions for Institutional Research, no. 154, Summer 2012 © Wiley Periodicals, Inc.
Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ir.20011

outcomes (student retention, faculty satisfaction) might be due in part to differences among higher-level groups or contexts (class size, department size, and so on). Analyses of these data using ordinary linear regression methods are problematic, as the underlying structure of the data often leads to violations of the assumptions of independence intrinsic to these models. Through HLM, we are able to model these dependencies and to examine how differential characteristics in the higher-level contexts help to explain variation in individual or lower-level outcomes. An HLM approach can also be used in place of repeated-measures analysis of variance in longitudinal studies. By viewing a series of repeated observations as lower-level outcomes nested within the individual, researchers are able to explore the effects of higher-level individual characteristics (gender, age) on the patterns of change in the lower-level outcomes over time.

Figure 1.1 represents a prototype situation for nested or clustered cross-sectional data from a single institution. In this figure, potential data of interest such as student persistence, gender, or first-year grade point average (GPA) reside at level one, the lowest level of the hierarchy. These level-one characteristics vary across individuals within the same department as well as between departments. Students are nested within different departments, and these departments may vary in terms of supports in place for mentoring new students or size of faculty in that department. These level-two characteristics vary between departments, but they do not vary between students within the same department. Finally, data representing the institution, such as total endowment or selectivity of undergraduate admissions, are common to all departments and all students within departments at that institution; there is no variability at the institutional level for the prototype model shown. Thus, while there seems to be three levels to the hierarchy, the analysis of outcomes at the student level would be examined through a two-level HLM: students nested within departments. More complex structures are easily accommodated in the HLM environment in both cross-sectional and longitudinal studies. If this cross-sectional data collection scheme were

Figure 1.1. Prototype of Nested or Clustered Data: Students Nested Within Departments



implemented at multiple institutions (see Hox, 1998, and Maas and Hox, 2005, for discussion of factors related to sample sizes at different levels), variability in institutional-level characteristics can be measured and examined, and the influence of institutional context as well as departmental context on student outcomes can be examined through a three-level HLM. If repeated observations of the outcome are collected on all students over a four-year period, such as end-of-year GPA, these repeated measurements reside at the lowest level of data collection; they are nested within students, who are nested within departments, and as a result would add a third level to the analytical design.

Whether the data of interest are longitudinal or cross-sectional, multilevel analyses are concerned with the study of variation. Just as in standard (single-level) regression, the goal of multilevel analysis is to attempt to explain variability, which implies that the outcome of interest can be reliably modeled through a well-chosen or predefined set of predictors, covariates, or explanatory variables. As the multilevel example illustrates, variability exists at each level of a multilevel analysis, and predictors or explanatory variables can exist at different levels as well. Overall, the primary motivation for employing multilevel analysis is to examine and understand the nature of the many different kinds of variability present in the data (Gelman and Hill, 2007). In doing so, we attempt to model the outcomes of interest by examining how group-level or individual-level characteristics are related to lowest-level outcomes.

An assumption in standard regression is that the observations or data subjected to analysis are statistically independent. With nested data, this assumption is clearly violated. Research has consistently shown that for clustered data, observations obtained from persons within the same cluster tend to exhibit more similarity to each other than to observations from different clusters. This similarity leads to underestimation of the standard errors for regression parameter estimates and inflates Type I error even when the similarity is mild (Donner, Birkett, and Buck, 1981; Sudman, 1985; Kenny and Judd, 1986; Murray and Hannon, 1990; Kish, 1995; Fowler, 2001). Cluster homogeneity is commonly measured through the intraclass correlation coefficient (ICC), which can be interpreted as the familiar Pearson correlation between two observations from the same cluster (Donner and Klar, 2000). In a two-level design, the ICC represents the proportion of total variance in the outcome that is captured by differences between the clusters or groups. When no variability is present between the clusters or groups, the value of the ICC is zero, and the assumption of independence among all individuals in the sample is justified. However, in the presence of between-cluster variability, the value of the ICC is positive, indicating a lack of independence, which invalidates standard regression models where clustering is ignored. The presence of ICC supports the adoption of a multilevel approach to analyzing the data, incorporating critical features of the hierarchical structure of the data into the analysis.

Clustered data may arise due to the existence of intact groups within an institution or by design if, for example, first-year students are randomly assigned to small-group mentoring activities to boost student engagement. Another situation in which IR researchers may be presented with clustered data is through sampling convenience. For example, requesting that all students within randomly selected intact courses complete a survey on first-year experiences would yield a more practical and feasible design relative to a sample based on a random selection of students across the entire college or university. However, such clustered samples have limitations as well as strengths that can affect how data may be interpreted. Whether naturally occurring or by intent, the structure of clustered data involves collecting information from clusters or groups of individuals experiencing a common phenomenon or event. In an IR setting, these common phenomena could arise from attending the same class or being in the same degree program. Regardless of the nature of the cluster, the ICC is found through decomposition of total variance in an outcome of interest into its within-group and between-group components; the ICC represents the proportion of variance that is between groups (Raudenbush and Bryk, 2002).

Historical Approaches to Analyzing Data with a Multilevel Structure

Prior to the advent of specialized software devoted to multilevel data, researchers often used two approaches when confronted with clustered or nested data: aggregation and disaggregation. Although multilevel models may eliminate the particular kinds of bias prevalent in these earlier approaches, we review them here to underscore the need for researchers to avoid the kinds of fallacies that earlier methods may have encouraged and to focus instead on methodologically appropriate and ethical practices for multilevel data (American Statistical Association, 1999; Goldstein, 2011).

In an aggregation approach, researchers sometimes averaged the lower-level data within a cluster or group and then used these averages as outcomes or predictor variables in a single-level analysis model. W. S. Robinson's seminal 1950 article on ecological correlations (reprinted in the *International Journal of Epidemiology*, 2009) describes an ecological correlation as the statistical correlation among groups of individuals. It was fairly common at that time to use ecological correlations as if they represented the correlations among the underlying individual-level data; the ecological fallacy refers to the inferential problems inherent in using group-level data to generalize to individual-level relationships. Robinson used 1930 census data to describe correlations among county- or region-level illiteracy rates and the percentage of African Americans in that region. His data showed the ecological correlation to be .946, while the individual-level correlation between illiteracy and race was .203. Since the

publication of Robinson's work, researchers have continued to examine and caution against the ecological fallacy, with implications for the importance of context in multilevel studies (for example, Schwartz, 1994; Susser, 1994; Diez-Roux, 1998; Oakes, 2009; Goldstein, 2011).

In a disaggregation approach, researchers disregarded the tendency for data from persons within distinct groups or geographical regions to be correlated and ignored the multilevel structure of the data completely. Thus, all data was analyzed as if it arose at the individual level, and group-relevant variables would retain the same value for all persons within the same group. Such an approach clearly violates the traditional assumption of independence necessary for valid statistical tests. Consequently, standard errors are underestimated and probability values for statistical tests are too small, leading to the potential for overstating statistical significance of the resulting research findings. These issues have been well documented by sampling methodologists and multilevel researchers (for example, Kish, 1995; Murray, 1998; Raudenbush and Bryk, 2002).

Much of the literature on levels-of-analysis problems has focused on the ecological fallacy, but researchers are also cautioned against the atomistic fallacy, which occurs while drawing inferences based on individual-level data and generalizing these inferences to group-level associations (Diez-Roux, 1998). Both kinds of fallacies can be avoided by careful consideration of the level at which data are collected (individual versus group) and by consistent representation of these levels in the statistical model. Hierarchically structured data, such as those that occur with most institutional research data, are uniquely represented through multilevel models.

Importance of Advancements in Statistical Methods in Institutional Research

Pascarella and Terenzini (1991, 2005) are renowned for their emphasis on methodological rigor in understanding how colleges affect students. Their work documents the importance of remaining current in statistical and research methods for those conducting institutional research. In addition to advancing theories and models for student change, theory development and research must be matched by advances in statistical methodologies. For example, two decades ago, Pascarella and Terenzini's 1991 volume discussed, in part, the strengths and limitations of the use of meta-analysis as an approach to aggregating and comparing results across research studies. In their 2005 volume, while still characterizing limitations to meta-analysis in their updated literature for the new edition, they specifically recognize the profound advances in statistical methods that have occurred over the past 20 years or so, including MLM. The capacity for multilevel models to strengthen our understanding of how colleges and universities affect students cannot be overemphasized. In the next section, we

highlight some of the ways in which multilevel models may be used in institutional research, before turning to our introduction of model notation and interpretation.

Research Questions in Institutional Research

Colleges and universities are complex organizations involving countless interactions among students, faculty, staff, and administration. These interactions occur among organizational entities made up of departments, schools, and colleges, each with unique policies, practices, and values. Observations of student achievement, faculty productivity, and other important performance indicators may be affected by group-level similarities based on these organizational structures. In addition, increased reliance on institutional data for strategic planning, accreditation, accountability, and performance improvement presents a significant challenge for IR professionals (Brittingham, O'Brien, and Alig, 2008; Voorhees, 2008). Similarly, the increasing demand for comparative analysis across institutions for the purpose of performance benchmarking requires that analytical models be developed that accommodate potential heterogeneity across institutions, states, and regions (Yorke, 2010). To effectively assess institutional performance, IR professionals require analytical tools that facilitate comparative analysis across these heterogeneous groups and permit the evaluation of group effects on individual-level performance. By learning and employing multilevel techniques to provide actionable information based in this broad institutional perspective, IR offices can position themselves as key partners in organizational dialogue and decision making (Parmley, 2009).

As the higher education landscape is transformed by demands for increased accountability, a growing emphasis on global demographics, aging faculty and facilities, an increased dependency on technology, and ongoing shifts in the economic climate, IR offices have become an invaluable part of institutional strategic positioning and planning efforts (Voorhees, 2008). To properly inform and support these activities, IR professionals must provide analyses of a wide variety of performance indicators involving data collected from all areas of the institution. Enrollment management measures including college choice, transition to college, student flow, attrition and retention, and student graduation rates are used to shape policy, inform practice, and guide strategic investment across the institution. Findings from studies involving student-centered indicators such as engagement, satisfaction, safety, and health and wellness are utilized to enhance student development. Analyses of campus-centered variables such as campus climate, diversity, sustainability, and service may be used to transform organizational culture. Institutional researchers may also be involved with efforts to effectively deploy campus human resources through the analysis of workplace satisfaction, faculty and staff work, and

organizational training effectiveness. Evaluation of performance in academic affairs incorporates measures of student achievement, assessment, program completion, student placements, faculty activity and productivity, and program effectiveness. As a result of this broad involvement, IR is well positioned to facilitate collaborative decision making, which transcends the multilevel organizational structures typical of college and universities (Leimer, 2009).

Introducing Multilevel Models

Our goal in this chapter is to introduce the IR researcher to the concepts and language of multilevel models. We present a discussion of the ways in which institutional researchers can use and interpret multilevel models, and we identify their strengths as well as limitations through this discussion. There are several different notational frameworks with which to represent multilevel models, and here we utilize the approach of Raudenbush and Bryk (2002), although familiarity with other representations, such as that presented in Snijders and Bosker (1999), is recommended. We focus in this introductory chapter on models where outcomes are continuous and measured at least at the interval level, but the literature on multilevel models for dichotomies, ordinal outcomes, counts, or times-to-event data—as well as longitudinal and latent-growth models—is fairly extensive and builds naturally on applications for continuous outcomes (see, for example, Snijders and Bosker, 1999; Raudenbush and Bryk, 2002; Singer and Willett, 2003; Gelman and Hill, 2007; O'Connell, Goldstein, Rogers, and Peng, 2008).

We begin with a simple example and introduce the development and interpretation of multilevel models by posing a series of questions that can be represented by different models. Suppose a university is interested in increasing its enrollment of underrepresented students (African American, Asian American, Hispanic, and American Indian) into STEM disciplines (science, technology, engineering, and mathematics). As a first step, the university wants to examine how graduating seniors' perceptions of their college experiences vary across academic units based on student race/ethnicity and selected characteristics of their home department or unit, including its type (STEM, non-STEM) and the proportion of underrepresented students in the department. A random sample of graduating students from all departments has been obtained.

In this example, variables are included at two levels. At the student level, we have the scores on senior student experiences, which may be captured through a self-reported rating scale of 1 through 10, with 10 representing more favorable experiences. Thus, the outcome Y_{ij} will represent the experiences rating for the ith senior in the jth department. We also have a student-level predictor, X_{ij}, which is a dummy-coded variable indicating whether the ith senior in the jth department is from an

underrepresented group (1 = yes, 0 = no). Two department-level predictors are included, and we use *W* rather than *X* to distinguish between variables at level one (*X*s) versus level two (*W*s). For our example, W_{1j} is dummy-coded to represent whether or not the *j*th department falls under the identification of a STEM discipline (1 = yes, 0 = no), and our second level-two predictor W_{2j} is the proportion of students from underrepresented groups in the *j*th department.

On Average, Do Seniors' Experience Ratings Differ Across Departments? In a multilevel analysis, the first model that is typically fit is referred to as the empty model, and it is the same as a one-way analysis of variance (ANOVA) model with random effects (Raudenbush and Bryk, 2002).

$$\text{Level 1:} \quad Y_{ij} = \beta_{0j} + r_{ij}$$

$$\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + u_{0j}$$

As with all statistical analyses, simplifying assumptions are made regarding the data. Although beyond the scope of this discussion, it is strongly recommended that readers investigate the validity of these assumptions through model comparisons and residual diagnostics. Here we assume that the department rating scores follow a normal distribution, with department-specific means, β_{0j}, and a common variance within all departments, σ^2. The existence of this common variance constitutes the homogeneity of variance assumption, which can readily be tested, and relaxed, in existing software for multilevel models. We also assume that the department means themselves vary based on a normal distribution with an overall mean γ_{00} and variance τ_{00}. Finally, we assume there is no correlation between the residuals at level one and those at level two. This set of assumptions can be collectively written as:

$$r_{ij} \sim iid\, N\left(0, \sigma^2\right)$$

$$u_{0j} \sim iid\, N\left(0, \tau_{00}\right)$$

$$Cov\left(r_{ij}, u_{0j}\right) = 0$$

The level-one model tells us about the variability in experience scores that exists within each academic unit; and the level-two model tells us about the variability between the academic units. All else being equal, a larger σ^2 would suggest that, within departments, there is a great deal of individual variability in seniors' experience rating scores. Similarly, all else being equal, a larger τ_{00} would suggest that there is a large amount of variability between departments in average experience rating scores.

By substituting the level-two model into the level-one equation (where the same set of assumptions hold), the two models representing variability at each level of data can be combined into a mixed model:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

The complex pattern of variation that is the defining characteristic of hierarchically structured data can be seen directly in this mixed-model form. In particular, two sets of residuals are used to represent the variation between and within the academic units, and their relative contribution to total variability is captured by the ICC. As described earlier, the ICC is a measure of the proportion of total variance (between + within) that can be attributed to the group or cluster; it provides an assessment of how strongly the clusters contribute to dependency in the data:

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$$

The value of the ICC will be positive, even when the contextual group effect is very small. Ignoring the existence of a positive ICC, even if close to zero, can have serious consequences for validity of hypothesis tests as well as for understanding and examining patterns of variability in the data.

Due to the complexity of the model and its partitioning of variance into level-specific components, estimation of model parameters requires iterative strategies generally solved through maximum likelihood procedures. In this simplest of models, the single fixed effect is γ_{00}, which represents the single point estimate for the grand mean of all departments' experience rating scores. Simultaneously, estimates are generated for the variance components, which summarize the contribution of random effects to the model at their respective levels (that is, level one: $\sigma^2 = \text{var}(r_{ij})$, and level two: $\tau_{00} = \text{var}(u_{0j})$).

Multilevel research attempts to explain variability in the dependent variable based on the predictors' contributions to their level-specific variation. Thus we use *X*s to attempt to reduce individual variation, σ^2, and we use Ws to try to account for between-group variability, τ_{00}. Inferential accuracy of the statistical tests of the fixed effects and the variance components rests on the validity of assumptions placed on the data. Generally, a t-test is used for testing whether a specific fixed effect (that is, γ_{00}) is equal to zero; and a Wald test or a chi-square test is used to assess whether a variance component (for example, τ_{00}) is statistically different from zero (see Chapter 2, this volume, for alternative approaches to significance testing based on deviance comparisons). As variance is explained or accounted for, improved understanding of the phenomena of interest is achieved.

Does the Senior Student's Status as a Member of an Underrepresented Group Affect Ratings of Perceived College Experiences? Once we have established that differences in outcomes between academic units exist, we next consider the effect of a level-one covariate, senior's group membership status. Status is a dichotomous predictor, with 1 indicating that the student is a member of an underrepresented group at the

university and 0 indicating otherwise. As a student-level predictor, status is entered at level one:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

This particular model is referred to as a random coefficients model. Note that an additional residual term is now included at level two. This residual term implies that the effect of status on ratings of college experience is expected to vary across academic units—that is, the slope for the status variable is not constant across departments, but varies between them. Thus, the intercepts and slopes from the level-one model vary at random. Within any given department, the expected average experience rating for non-underrepresented students (that is, when $X_{ij} = 0$) is β_{0j}, and the expected average experience rating for underrepresented students (that is, when $X_{ij} = 1$) is $\beta_{0j} + \beta_{1j}$.

Using back-substitution similar to the empty model, we can derive the mixed-model expression for this random coefficients model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + u_{1j}X_{ij} + r_{ij}$$

Averaging across all departments in this example, the mean experience rating is γ_{00} for non-underrepresented students, and the mean experience rating for underrepresented students is $\gamma_{00} + \gamma_{10}$. A significant t-test result for γ_{10} implies that the mean difference in experience ratings between non-underrepresented students and underrepresented students is statistically different from zero. The residual for the slope captures variability in the effect of senior's underrepresentation status across schools. If this effect does not vary between departments, then all of the u_{1j}s would equal zero, and the estimated slope for status would be constant for all departments. Thus, the slope parameter can be fixed, or held constant, rather than be free to vary across groups.

The addition of a single predictor at level one has important implications for the covariance structure of the model:

$$r_{ij} \sim iid\ N\left(0, \sigma^2\right)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim iid\ N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right]$$

The covariance structure of the level-two residuals is often written as $iid\ N(\underline{\mathbf{0}}, \mathbf{T})$ where the size of the symmetric covariance matrix $\mathbf{T}$ depends

on the number of randomly varying coefficients at level one. This structure is more complex than in the previous empty model, due to the accommodation of variability in slopes for the level-one covariate. The covariance term, τ_{01}, captures the association between the level-one intercepts and slopes. If we set the slope variance, τ_{11}, to zero, we end up with a simpler model but one in which the effect of a student's underrepresentation status is constant across all departments.

Adding a relevant predictor to the level-one model should account for some of the individual-level variance contributing to σ^2. The reduction in variance achieved over the empty model can be assessed by comparing the two variances:

$$\frac{\sigma^2_{empty\_model} - \sigma^2_{RC\_model}}{\sigma^2_{empty\_model}}$$

This simple proportion can be used to calculate variance accounted for in more complex models as well. If the expression is zero or negative, it suggests that no variance was reduced over the empty model. Negative results may sometimes arise, due to the maximum likelihood estimation procedures used to fit these kinds of models.

Note that we were able to directly interpret the intercepts, β_{0j}, in this random coefficients model. In all regression models, the intercept is interpreted as the prediction when the covariate is zero. For a dummy-coded variable such as underrepresentation status, this process is straightforward, but for many continuous variables—for example, GPA—there may be no interpretable zero. In single-level regression we are rarely bothered by this fact because our attention is focused on the slopes, which represent effects of different predictors on the outcome. In the multilevel framework, it is often optimal to have intercepts that are directly interpretable. A process called centering is usually employed with continuous level-one predictors to provide a more meaningful interpretation to the intercepts.

Centering has several forms in the multilevel framework, and although it yields a meaningful intercept, it can change the degree of variability in the model and thus should be used carefully. Group-mean centering, or centering within contexts (CWC), of a covariate subtracts the mean of each group's covariate score from each participant's original covariate score: $(X_{ij} - \bar{X}_{.j})$. Centering at the grand mean (CGM) of the covariate subtracts the overall mean from each participant's score: $(X_{ij} - \bar{X}_{..})$. Substituting either of these into the level-one random coefficients model has no effect on the estimation for the fixed-effect slopes, but it does change the quantity being estimated, and thus our interpretation, for the intercepts. For example, with CWC the level-one equation becomes $Y_{ij} = \beta_{0j} + \beta_{1j}\,(X_{ij} - \bar{X}_{.j}) + r_{ij}$. The intercept is now the predicted value of Y_{ij} when X_{ij} is at the group mean, that is, for a participant who is at the

average value of the predictor for their group. Similarly for CGM, the intercept becomes the prediction when X_{ij} is at the grand mean for the sample, or for a participant who is at the average value of the covariate for all persons in the sample: $Y_{ij} = \beta_{0j} + \beta_{1j}\ (X_{ij} - \bar{X}_{..}) + r_{ij}$. Centering is used primarily for continuous covariates, but Raudenbush and Bryk (2002) provide a discussion of its use with effect-coded dichotomous variables. Other recommended sources for the use of centering and its implications in multilevel analyses include Hofmann and Gavin (1998), Paccagnella (2006), and Enders and Tofighi (2007).

Can Contextual Variables (Characteristics of Academic Units) Help to Explain Variability in Intercepts and Slopes across Units? In addition to reducing within-unit variance by the inclusion of level-one predictors, the random coefficients model provides us with baseline information about how much variability in the level-one intercepts and slopes exists between the academic departments. In our example, we have two level-two covariates: W_{1j} is a STEM indicator variable for whether the *j*th department is a STEM discipline (1 = yes, 0 = no), and W_{2j} is the proportion of students from underrepresented groups within the *j*th department. We include these variables at level two to examine their contribution to predictions of both the slopes and intercepts from level one.

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_{1j} + \gamma_{12}W_{2j} + u_{1j}$$

Because this model retains the same level-one model as the previous random coefficients model, the structure of the variance and covariance components remains the same, although estimates are likely to differ—and be smaller, if variance has been explained by the inclusion of level-two predictors:

$$r_{ij} \sim iid\ N\left(0, \sigma^2\right)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim iid\ N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right]$$

Models of this type, which include both level-one and level-two predictors, are often called intercepts and slopes as outcomes models, or conditional models. We can assess reduction in between-group variance in the intercepts and slopes using an approach similar to that taken to find the reduction in within-group variance when we discussed the random coefficients model. In the next expression, *q* refers to the *q*th random coefficient

(including the intercept) from the level-one model. Thus, our assessment of reduction in variance would allow us to verify if the collection of *W*s added into each equation helped to improve predictions of the slopes and intercepts, respectively.

$$\frac{\tau_{qq(base)} - \tau_{qq(full)}}{\tau_{qq(base)}}$$

The conditional model can be used to answer questions regarding how characteristics of academic departments, such as whether they represent a STEM discipline or the proportion of students from underrepresented groups currently enrolled in the department, are related to mean perceived experience ratings for non-underrepresented students in that department, intercept β_{0j}, or to differences in perceived experience ratings between non- and underrepresented students, slope β_{1j}.

Averaging across departments in the conditional model, γ_{00} tells us the expected mean perceived experience score for non-underrepresented students not enrolled in a STEM department ($W_{1j} = 0$) and for departments in which the proportion of students from underrepresented groups is zero ($W_{2j} = 0$). Also in the level-two intercept equation, γ_{01} is the fixed effect for W_{1j} and represents the expected difference in the mean perceived experience ratings of non-underrepresented students for those in STEM disciplines, holding W_{2j} constant. The increase or decrease in mean ratings attributed to proportion of students from underrepresented groups in the department, holding STEM discipline constant, is captured by the effect of W_{2j} (γ_{02}). Statistical tests for these coefficients would indicate whether they are statistically different from zero or not.

The fixed effects in the level-two slope equation can be interpreted in a similar fashion. Averaging across departments, γ_{10} represents the mean effect of being a student from an underrepresented group on the perceived experience ratings, and this effect is conditional on whether the student's academic unit is a STEM department (γ_{11}) and on the unit's proportion of underrepresented students (γ_{12}). Again, significance tests for these coefficients provide information on their statistical contribution to the level-two model being examined.

When level-two predictors have a significant effect on a level-one slope estimate, such as may be expected through γ_{11} and γ_{12}, a cross-level interaction has occurred. A cross-level interaction is an interaction between a level-one predictor *X* and a level-two predictor *W*. The mixed-model expression shows this quite clearly:

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \gamma_{10}X_{ij} + \gamma_{11}X_{ij}W_{1j} + \gamma_{12}X_{ij}W_{2j} + u_{0j} + u_{1j}X_{ij} + r_{ij}$$

This model may be simplified by fixing the level-one coefficients that show little or no variability or by removing nonsignificant predictors or

cross-level interactions. Consistent with good statistical modeling practice, decisions to include or exclude variables and interaction terms should also be based on theory and the purpose of the research, not just on the results of a statistical test.

We now have much of the basic notation necessary to support the design and analysis of multilevel models within an IR setting. Other chapters in this book will strengthen these concepts and solidify their application to authentic research practice. Before summarizing, however, we turn to a few remaining issues to complete this chapter.

Designing Research within Institutional Research Settings

Maximum likelihood is a large-sample estimation method and requires sufficiently large samples for valid inferences. Unfortunately, no single magic number indicates when a large enough sample size has been obtained for a specific research question, and many factors can influence the optimal sample size necessary to reliably detect effects of interest. Maas and Hox (2005) discuss the impact that design and sample/population features have on the quality of estimation and the resulting inferences based on a multilevel analysis, including the sizes of variances and covariances and the ICC.

The size of the ICC for a particular outcome variable affects the size of standard errors of the regression coefficients for predictor variables, which in turn form the denominator of many statistical tests and contribute to the endpoints of confidence intervals for point estimates of those regression coefficients. In the sampling literature on cluster and other complex samples, the impact of the ICC is generally characterized through the design effect, or "deff," which represents how much the standard errors from a clustered design are underestimated relative to a simple random sample (SRS): deff = 1 + (m – 1) * ICC, where *m* = average cluster or group size (Kish, 1995; Murray, 1998). A sampling design that mimics an SRS has a deff of 1.0; clustering increases the design effect, which indicates that the assumption of independence of observations is violated, making traditional tests of significance biased. The design effect is also referred to as the inflation factor because it tells us how much the sampling variance is inflated over the sampling variance expected in the gold standard of an SRS, due to the clustering effect (the ICC). In design planning, estimates of expected ICC are singularly important, because the balance between number of level-one and number of level-two units is typically weighted toward a larger number of level-two units as ICC increases, and the power to detect important group differences will diminish as the level-two sample size decreases, all else being equal (Moerbeek, van Breukelen, and Berger, 2000; Spybrook, 2008). The relationships among ICC, sample size at multiple levels, and power are the driving force

behind calls for researchers to publish the ICCs obtained from their research studies (Murray, Varnell, and Blitstein, 2004).

A dilemma in most multilevel research design scenarios arises because recruiting another unit or group is often more difficult—and more costly—than recruiting additional participants from an already sampled unit or group. Institutional researchers often have access to very large samples given the size of their institutions, but a larger level-one sample size will not compensate for the decrease in power for detecting group-specific differences or effects given a small level-two sample size. Thus, studies in IR need to be carefully planned to attain their research goals with sufficient power.

Resources are available for estimating sample sizes based on desired or expected design characteristics. The Optimal Design software (Liu and others, 2006; Spybrook, 2008) is freely accessible on the Internet from the W.T. Grant Foundation (http://www.wtgrantfoundation.org/resources/overview/research_tools) and allows researchers to manipulate features of a design to optimize power under different research or statistical constraints. In addition, the literature and resources for optimizing power given design and cost limitations continue to expand (for example, see Moerbeek, van Breukelen, and Berger, 2000, 2001; Raudenbush and Liu, 2000, 2001; Bloom, 2005; Hedges and Rhoads, 2010). For optimal design of studies in IR, researchers should ensure that their designs are strong enough to match the importance of the research questions being asked in this field (for example, see Moerbeek, van Breukelen, and Berger, 2000, 2001; Raudenbush and Liu, 2000, 2001; Bloom, 2005; Hedges and Rhoads, 2010).

Summary

We conclude with a brief summary of why MLM is important to researchers in IR.

1. If our data are hierarchical, and most IR data are indeed hierarchical, MLM yields valid estimates of variable effects or group differences by directly taking into account the nature of clustering inherent in the data.
2. Ignoring the hierarchical structure of IR data limits the opportunity to examine whether and how differences in patterns of relationships between predictors and outcomes vary across groups, such as academic units, schools, or other definable levels in the data.
3. Study design features can be used to advantage in designing future studies to replicate or qualify observed or expected effects, thus furthering the opportunity to strengthen and build upon research in IR.
4. In multilevel analyses, the ICC can be directly measured and the impact of clustering accounted for in statistical models, thus decreasing tendency for Type I error and flawed conclusions.

5. Contexts do matter, and student's individual and collective college experiences are relevant to many stakeholders (state policy makers, university administrators, professors, parents, and students themselves). IR researchers have an obligation to examine outcomes and enhance this experience for students through ethical and valid methodologies.

References

American Statistical Association. "Ethical Guidelines for Statistical Practice." 1999. Retrieved August 18, 2011, from http://www.amstat.org/about/ethicalguidelines.cfm.

Bloom, H. S. "Randomizing Groups to Evaluate Place-Based Programs." In H. S. Bloom (ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage, 2005.

Brittingham, B., O'Brien, P. M., and Alig, J. L. "Accreditation and Institutional Research: The Traditional Role and New Dimensions." In D. G. Terkla (ed.), *More Than Just Data*. New Directions for Higher Education, no. 141. San Francisco: Jossey-Bass, 2008.

Diez-Roux, A. V. "Bringing Context Back into Epidemiology: Variables and Fallacies in Multilevel Analysis." *American Journal of Public Health,* 1998, *88*(2), 216–222.

Donner, A., Birkett, N., and Buck, C. "Randomization by Cluster: Sample Size Requirements and Analysis." *American Journal of Epidemiology,* 1981, *114*(6), 906–914.

Donner, A., and Klar, N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold, 2000.

Enders, C. K., and Tofighi, D. "Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue." *Psychological Methods,* 2007, *12*(2), 121–138.

Fowler, F. J. *Survey Research Methods*. (3rd ed.). Newbury Park, Calif.: Sage Publications, 2001.

Gelman, A., and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press, 2007.

Goldstein, H. "Ethical Aspects of Multilevel Modeling." In A.T. Panter and S. K. Sterba (eds.), *Handbook of Ethics in Quantitative Methodology*. New York: Routledge, 2011.

Hedges, L. V., and Rhoads, C. "Statistical Power Analysis in Educational Research (NCSER 2010–3006)." Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education, 2010. Retrieved August 18, 2011, from http://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf.

Hofmann, D. A., and Gavin, M. B. "Centering Decisions in Hierarchical Linear Models: Implications for Research in Organizations." *Journal of Management,* 1998, *23*, 723–744.

Hox, J. "Multilevel Modeling: When and Why?" In I. Balderjahn, R. Mather, and M. Schader (eds.), *Classification, Data Analysis, and Data Highways*. *Proceedings of the 21st Annual Conference of the Gesellschaft fur Klassifikation*. New York: Springer-Verlag, 1998.

Kenny, D. A., and Judd, C. M. "Consequences of Violating the Independence Assumption in Analysis of Variance." *Psychological Bulletin,* 1986, *99*, 422–431.

Kish, L. *Survey Sampling*. New York: Wiley Classics, 1995. (Originally published 1965.)

Leimer, C. "Taking a Broader View: Using Institutional Research's Natural Qualities for Transformation." In C. Leimer (ed.), *Imagining the Future of Institutional Research*. New Directions for Institutional Research, no. 143. San Francisco: Jossey-Bass, 2009.

Liu, X, and others. "Optimal Design for Multilevel and Longitudinal Research," Version 1.77. HLM Software, 2006.

Maas, C.J.M, and Hox, J. J. "Sufficient Sample Sizes for Multilevel Modeling." *Methodology,* 2005, *1*(13), 86–92.

Moerbeek, M., van Breukelen, G.J.P., and Berger, M.P.F. "Design Issues for Experiments in Multilevel Populations." *Journal of Educational and Behavioral Statistics,* 2000, *25*, 271–284.

Moerbeek, M., van Breukelen, G.J.P., and Berger, M.P.F. "Optimal Experimental Designs for Multilevel Models with Covariates." *Communications in Statistics, Theory and Methods,* 2001, *30*, 2683–2697.

Murray, D. M. *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press, 1998.

Murray, D. M., and Hannon, P. J. "Planning for the Appropriate Analysis in School-Based Drug-Use Prevention Studies." *Journal of Consulting and Clinical Psychology,* 1990, *58*, 458–468.

Murray, D. M., Varnell, S. P., and Blitstein, J. L. "Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments." *Public Health Matters,* 2004, *94*(3), 423–432.

O'Connell, A. A., Goldstein, J., Rogers, J., and Peng, C. J. "Multilevel Logistic Models for Dichotomous and Ordinal Data." In A. A. O'Connell and D. B. McCoach (eds.), *Multilevel Modeling of Educational Data*. Charlotte, N.C.: Information Age, 2008.

Oakes, J. M. "Commentary: Individual, Ecological and Multilevel Fallacies." *International Journal of Epidemiology,* 2009, *38*(2), 361–368.

Paccagnella, G. "Centering or Not Centering in Multilevel Models: The Role of the Group Mean and the Assessment of Group Effects." *Evaluation Review,* 2006, *30,* 66.

Parmley, K. A. "Raising the Institutional Research Profile: Assessing the Context and Expanding the Use of Organizational Frames." In C. Leimer (ed.), *Imagining the Future of Institutional Research*. New Directions for Institutional Research, no. 143. San Francisco: Jossey-Bass, 2009.

Pascarella, E., and Terenzini, P. *How College Affects Students*. San Francisco: Jossey-Bass, 1991.

Pascarella, E., and Terenzini, P. *How College Affects Students, Vol. 2: A Third Decade of Research*. San Francisco: Jossey-Bass, 2005.

Raudenbush, S. W., and Bryk, A. S. *Hierarchical Linear Models: Applications and Data Analysis Methods*. (2nd ed.). Thousand Oaks, Calif.: Sage Publications, 2002.

Raudenbush, S. W., and Liu, X. F. "Statistical Power and Optimal Design for Multisite Randomized Trials." *Psychological Methods,* 2000, *5*(2), 199–213.

Raudenbush, S. W., and Liu, X. F. "Effects of Study Duration, Frequency of Observation, and Sample Size on Power in Studies of Group Differences in Polynomial Change." *Psychological Methods,* 2001, *6*(4), 387–401.

Robinson, W. S. "Ecological Correlations and the Behavior of Individuals." *Journal of Epidemiology,* 2009, *38*(2), 337–341. (Originally published 1950.)

Schwartz, S. "The Fallacy of the Ecological Fallacy: The Potential Misuse of a Concept and the Consequences." *American Journal of Public Health,* 1994, *84*, 819–824.

Singer, J. D., and Willett, J. B. *Applied Longitudinal Data Analysis: Methods for Studying Change and Event Occurrence*. New York: Oxford University Press, 2003.

Snijders, T.A.B., and Bosker, R. J. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications, 1999.

Spybrook, J. "Power and Sample Size for Classroom and School-Level Interventions." In A. O'Connell and B. McCoach (eds.), *Multilevel Analysis of Educational Data.* Greenwich, Conn.: Information Age, 2008.

Sudman, S. "Mail Surveys of Reluctant Professionals." *Evaluation Review,* 1985, *9*(3), 349–359.

Susser, M. "The Logic in Ecological: I. The Logic of Analysis." *American Journal of Public Health,* 1994, *84*(5), 825–829.

Voorhees, R. A. "Institutional Research's Role in Strategic Planning." In D. G. Terkla (ed.), *More Than Just Data.* New Directions for Higher Education, no. 141. San Francisco: Jossey-Bass, 2008.

Yorke, M. "'Supra-Institutional Research': A Cost-Effective Contribution towards Enhancement." *Journal of Higher Education Policy and Management,* 2010, *32*(3), 261–273.

ANN A. O'CONNELL is a professor of quantitative research, evaluation, and measurement at The Ohio State University.

SANDRA J. REED is a doctoral candidate in quantitative research, evaluation, and measurement at The Ohio State University.