Rapid #: -20075695

CROSS REF ID: **1220863**

LENDER: **IPL :: Main Library**

BORROWER: **PUL :: Interlibrary Services, Firestone**

TYPE:	Book Chapter
BOOK TITLE:	Introduction to Modern Modelling Methods /
USER BOOK TITLE:	Introduction to Modern Modelling Methods /
CHAPTER TITLE:	Introduction to Modern Modelling Methods
BOOK AUTHOR:	D. Betsy McCoach; Dakota Cintron
EDITION:	
VOLUME:	
PUBLISHER:	
YEAR:	2022
PAGES:	1-45
ISBN:	9781529711103
LCCN:	
OCLC #:	

Processed by RapidX: 1/9/2023 7:50:41 AM

1

CLUSTERING AND DEPENDENCE: OUR ENTRY INTO MULTILEVEL MODELLING

Chapter Overview

Frequently in the social sciences, our data are nested, clustered or hierarchical in nature: individual observations are nested within a hierarchical structure. 'The existence of such data hierarchies is neither accidental nor ignorable' (Goldstein, 2011, p. 1). Examples of naturally occurring hierarchies include students nested within classrooms, teachers nested within schools, schools nested within districts, children nested within families, patients nested within hospitals, workers nested within companies, husbands and wives nested within couples (dyads) or even observations across time nested within individuals. 'Once you know that hierarchies exist, you see them everywhere' (Kreft & de Leeuw, 1998, p. 1).

Multilevel modelling (MLM) provides a technique for analysing such data. It accounts for the hierarchical structure of the data and the complexity that such structure introduces in terms of correctly modelling variability (Snijders & Bosker, 2012). Multilevel models are often referred to as **hierarchical linear models**, *mixed models*, *mixed-effects models* or *random-effects models*. Researchers often use these terms interchangeably, although there are slight differences in their meanings. For instance, *hierarchical linear model* is a more circumscribed term than the others: it assumes a normally distributed outcome variable. In contrast, mixed-effects or random-effects models are more general than multilevel models: they denote **non-independence** within a data set, but that non-independence does not necessarily need to be hierarchically nested.

In this book, we focus on one particular type of random-effects model: the multilevel model, in which units are hierarchically nested within higher level structures. Other common random-effects models include cross-classified random-effects models, which account for non-independence that is crossed, rather than nested. For example, in longitudinal educational studies, students often change teachers or transfer from one school to another; hence, students experience distinct combinations of teachers or schools. In such scenarios, students are *cross-classified* by two teachers or two schools. Multiple-membership models allow for membership in multiple clusters simultaneously. Although cross-classified and multiple-membership models are random-effects models, they are not purely multilevel models because they do not exhibit clean, hierarchical data structures. This book focuses on **hierarchical linear modelling (HLM)/multilevel modelling (MLM)**. Interested readers should refer to the following resources for more information about cross-classified and multiple-membership models: Airoldi et al. (2015), Beretvas (2008) or Fielding and Goldstein (2006). In addition, this book assumes normally-distributed continuous outcomes. To learn more about using MLM techniques with non-normal (binary, ordinal or count) outcomes, see O'Connell and McCoach (2008) or Raudenbush and Bryk (2002).

In MLM, **organisational models** are models in which people are clustered within hierarchical structures such as companies, schools, hospitals or towns. Multilevel models

also prove useful in the analysis of longitudinal data, where observations across time are nested within individuals.

In this chapter, we introduce the MLM framework and discuss two-level multilevel models in which people are clustered within organisations or groups. We begin our introduction to MLM by introducing basic terms and ideas of MLM as well as introducing one of the most fundamental concepts in the analysis of **clustered data**: the **intra-class correlation coefficient** (ICC). Subsequently, Chapter 2 provides an overview of standard two-level multilevel organisational models, and Chapter 3 illustrates fundamental MLM techniques with an applied example. In Chapters 4–6, we turn our attention to structural equation modelling. In Chapters 7 and 8, we return to MLM, demonstrating its application to individual growth models.

Nested data and non-independence

Most traditional statistical analyses assume that observations are independent of each other. In other words, the assumption of independence means that subjects' responses are not correlated with each other. For example, imagine that a survey company administers a survey to a sample of participants. Under the assumption of independence, one participant's responses do not correlate with the responses of any of the other participants. The assumption of independence might be reasonable when data are randomly sampled from a large population. However, the responses of people clustered within naturally occurring organisational units (e.g. schools, classrooms, hospitals, companies) are likely to exhibit some degree of relatedness, given that they were sampled from the same organisational unit. For instance, students who receive instruction together in the same classroom, delivered by the same teacher, tend to be more similar in their achievement (and other educational outcomes) than students instructed by different teachers.

Observations within a given cluster often exhibit some degree of dependence (or interdependency). In such a scenario, violating the assumption of independence produces incorrect **standard errors** that are smaller than they should be. Therefore, inferential statistical tests that violate the assumption of independence have inflated Type I error rates: they produce statistically significant effects more often than they should. The Type I error rate is the probability of rejecting the null hypothesis when the null hypothesis is correct. Alpha, the desired/assumed Type I error rate, is commonly set at .05. However, alpha may not equal the actual Type I error rate if we fail to meet the assumptions of our statistical test (i.e. normality, independence, homoscedasticity etc.). MLM techniques allow researchers to model the relatedness of observations within clusters explicitly. As a result, the standard errors

from multilevel analyses account for the clustered nature of the data, resulting in more accurate Type I error rates.

The advantages of MLM are not purely statistical. Substantively, it may be of great interest to understand the degree to which people from the same cluster are similar to each other and to identify variables that help predict variability both within and across clusters. Multilevel analyses allow us to exploit the information contained in clustered samples and to partition the variance in the outcome variable into **between-cluster variance** and **within-cluster variance**. We can also use predictors at both the **individual level** (level 1) and the group level (level 2) to try to explain this between- and within-cluster variability in the outcome variable.

In general, MLM techniques allow researchers to model multiple levels of a hierarchy simultaneously, partition variance across the levels of analysis and examine relationships and interactions among variables that occur at multiple levels of a hierarchy. In MLM, a level is 'a focal plane in social, psychological, or physical space that exists within a hierarchical structure' (Gully & Phillips, 2019, p. 11). Generally, the levels of interest within an analysis depend on the phenomena and research questions (Gully & Phillips, 2019). For example, in a study of instructional techniques, where students are nested within teachers, students are level-1 units and teachers are level-2 units. In contrast, in a study of teachers' perceptions of their principals' leadership, teachers are nested within principals. In this case, teachers are level-1 units and principals are level-2 units. Often, researchers use the term *organisational model* to refer to cross-sectional MLM where individuals (level-1 units) are clustered within some sort of organisational, administrative, social or political hierarchy (level-2 units).

Traditional correlations and regression-based approaches estimate the relationship between two variables. However, standard single-level analyses (which ignore the clustered/hierarchical nature of the data) assume that the relationship between the variables is constant across the entire sample. MLM allows the relationships among key substantive variables to randomly vary across clusters. For example, the relationship between socio-economic status (SES) and achievement may vary by school. In some schools, student SES may be a strong (positive) predictor of students' subsequent academic achievement; in other schools, SES may be completely unrelated to academic achievement (Raudenbush & Bryk, 2002).

Additionally, in MLM, researchers can study relationships among variables that occur at multiple levels of the data hierarchy as well as potential interactions among variables at multiple levels while allowing relationships among lower-level variables to randomly vary by cluster. How much of the between-cluster variability in these relationships (or in the cluster means) can be explained by cluster-level variables? For instance, imagine we want to study the relationships between student ability, teaching style and academic achievement. The data are clustered: students are nested within teachers (classrooms). For simplicity, assume that each teacher teaches only

one class. Therefore, the teacher and classroom levels are synonymous, and student ability varies across different students taught by the same teacher. Consequently, student ability is an individual-level (or level-1) variable. Although teaching style varies across teachers, every student within a given teacher's class is exposed to a single teacher with one individual teaching style. Therefore, teaching style varies across classrooms, but not within classrooms, so teaching style is a classroom/teacher (cluster) level, or level-2 variable.

Of course, the effect of a teacher's teaching style does not necessarily have the same effect on all students. In our current example, we might hypothesise that teaching style moderates the effect of student ability on student achievement. In other words, the relationship between student ability and student achievement varies as a function of teachers' teaching style. For example, some teachers may strive to ensure that all students in the class meet the same set of grade-level standards and are exposed to the same content at the same level, ensuring that all students in the class have the same set of skills and knowledge. In contrast, other teachers may differentiate instruction to meet the needs of individual students. We hypothesise that the relationship between ability and achievement would likely be stronger in the classrooms where teachers differentiate instruction than in the standards-based classrooms. In a standard linear regression model, we can include an interaction between teaching style and student ability. However, the multilevel framework allows the slope for the effect of students' ability on achievement to randomly vary across classrooms, even after controlling for all teacher- and student-level variables in the model. If the ability/achievement slope randomly varies, even after including teaching style in the model, the relationship between ability and achievement does indeed vary across classrooms but the teachers' teaching style does not fully explain the between-class variability in the ability/achievement relationship. Perhaps, other omitted classroom-level variables may explain the variability in the ability/achievement relationship across classes. MLM allows us to ask and answer more nuanced questions than are possible within traditional regression analyses.

As the preceding paragraphs highlight, multilevel models are incredibly useful for studying organisational contexts like schools, companies or families. However, many other types of data exhibit dependence. For instance, multiple observations collected on the same person represent another form of nested data. Growth curve and other longitudinal analyses can be reframed as multilevel models, in which observations across time are nested within individuals. Using the MLM framework, we partition residual or error variance into within-person **residual variance** and between-person residual variance. In such a scenario, between-person residual variance represents between-person variability in any randomly varying level-1 parameters of interest, such as the intercept (which we commonly centre to represent initial status in growth models) and the growth slope. Within-person residual variance represents

the variance of time-specific residuals, which is generally referred to as measurement error. We explore multilevel growth models in Chapters 7 and 8 of this book and demonstrate how to reframe the basic MLM framework to analyse longitudinal data. However, for the remainder of Chapters 1 to 3, we focus exclusively on cross-sectional organisational models.

Intra-class correlation coefficient

This section introduces one of the most fundamental concepts in MLM: the intraclass correlation coefficient (ICC). The ICC measures the proportion of the total variability in the outcome variable that can be explained by cluster membership. The ICC also provides an estimate of the expected correlation between two randomly drawn individuals from the same cluster (Hox et al., 2017).

Most traditional statistical tests (multiple regression, analysis of variance [ANOVA] etc.) assume that observations are independent. The assumption of independence means that cases 'are not paired, dependent, correlated, or associated in any way' (Glass & Hopkins, 1996, p. 295). Nested or clustered data violate this assumption because clustered observations tend to exhibit some degree of interdependence. In other words, observations nested within the same cluster tend to be more similar to each other on a given outcome variable than observations drawn from two different clusters. This interdependence, resulting from the sampling design, affects the variance of the outcome, which in turn affects estimates of the standard errors for model parameters.

Of course, the degree of dependence also varies by outcome variable, and some outcome variables may not exhibit any discernible dependence, even though the observations are clustered. For example, students are clustered within classrooms, so we would expect to see that academic outcomes such as mathematics and reading achievements exhibit some degree of within-class/within-teacher dependence. However, other variables may exhibit little to no dependence, even though they are clustered. Therefore, we compute the ICC separately for each outcome variable of interest.

To better understand this phenomenon, let's imagine that a research assistant named Igor receives the task of surveying 1000 people about how many hours they sleep per night (on average). Instead of randomly sampling 1000 people, he decides that he can accomplish the task much more quickly if he samples 250 households (each of which has four members) and asks all members of each household to respond to the sleep survey. For simplicity, let's assume that each of the 250 households are drawn from different neighbourhoods so that outside noise such as car alarms or sirens that affect one household do not affect any other households. Of course, in

households where one member is not sleeping well, other members of the household also tend to sleep less well. (Classic sleep disturbances that affect the sleep patterns of entire households to some degree include crying babies, barking dogs, visitors, late or early household events etc.). However, in this scenario, there is no reason to believe that sleep patterns exhibit any dependencies across households.

So, what happens to variability in the outcome variable, sleep, within households? Because people's sleep hours are more similar within a given household, the expected variability in sleep hours for members of the same household is smaller than the expected variability of members who reside in different households. Therefore, knowing the household in which a person resides can help explain some of the variability in sleep hours. This is *between-cluster variability*. Another way of thinking about between-cluster variability is to imagine computing the mean number of sleep hours for each household in the study. The degree to which those household-aggregated means vary across clusters (households) represents *between-cluster variance*. At first, it may seem counter-intuitive that similarities within clusters actually relate to between-cluster variance. However, a quick thought experiment may help. Imagine that Igor samples households in Stepford, and in these households, every member must sleep exactly as much as the patriarch. In such a scenario, all members of each household have the exact same sleep hours. Therefore, all variance in the outcome variable (sleep hours) must be between-cluster variance; there is no *within-cluster variance*.

Of course, Stepford doesn't actually exist. Even though people within the same household may be more similar to each other than to people from different households, they are not exactly the same. With clustered data, knowing the cluster helps to explain some (but not all) of the variability in the outcome variable of interest.

Instead, we can partition the total variability in sleep time into the portion that is within clusters (i.e. how much do members of the same household differ from their household average in terms of sleep time?) and the portion that is between clusters (i.e. how do households differ from each other in terms of sleep time?). The degree to which people within the same household (or cluster) differ from the household average is *within-cluster variability*. In other words, it is the (pooled) variability across people within the same cluster. Conceptually, *between-cluster variability* represents the variability in the cluster means. Between-cluster variance is analogous to aggregating data to the cluster level and computing cluster means for each cluster and then estimating how much the cluster means vary.

The ICC describes how similar, or homogeneous, individuals are within clusters and how much they vary across clusters: it quantifies the degree of dependence, or the degree of relationship among units from the same cluster (Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). The ICC is the proportion of between-cluster variance, or the proportion of the total variability in the outcome variable that can be explained by cluster membership. The calculation of the ICC (often symbolised

as ρ, 'rho') involves partitioning the total variability in the outcome variable into within-cluster variance (σ^2) and between-cluster variance (τ_{00}). To compute the ICC, we simply divide the between-cluster variability (τ_{00}) by the total variability ($\tau_{00} + \sigma^2$), as the following equation shows:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{1.1}$$

A large ICC indicates that there is a large degree of similarity within clusters (σ^2 is small) and/or a large degree of variability across clusters (τ_{00} is large). An ICC of 1 indicates that all observations within a cluster are perfect replicates of each other and all variability lies between clusters: the within-cluster variance is 0. An ICC of 0 indicates that observations within a cluster are no more similar to each other than observations from different clusters: the between-cluster variance is 0. We expect a **simple random sample** from a population to have an ICC of 0; the assumption of independence is the assumption that the ICC = 0.

Returning to our sleep example, imagine the ICC (the proportion of between-cluster variability) is .40. This means 40% of the variance in sleep time lies between households and 60% of the variance in sleep time lies within households. It also means that the expected correlation in sleep time for two members of the same household is .40.

To recap, cluster means vary (*between-cluster variance*). People in the same cluster also differ from each other (though two people from a single cluster differ less than two randomly selected people) (*within-cluster variance*). The sum of the within- and between-cluster variances represents total variance in the outcome variable. The ICC indicates the proportion of total variability explained by group membership; an ICC of 1.00 suggests members of each cluster are perfect replicates of one another (so, all variation occurs across clusters), whereas an ICC of .00 implies that cluster members are completely independent of one another (i.e. uncorrelated), akin to a theoretical random sample.

Effective sample size

A concept related to the ICC is **effective sample size** (n_{eff}). Is Igor's sample of 1000 people within 250 households really the same as sampling 1000 people from 1000 different households? No – given the built-in dependence of people within the same household, Igor hasn't really obtained as much information about people's sleep habits as he would have if he had actually sampled 1000 people from 1000 different households.

So, why does ignoring this non-independence (as traditional tests of significance do) increase the possibility of making a Type I error (rejecting the null hypothesis when we should have failed to reject it)? Because people within clusters are more homogeneous than people from different clusters, the variance of the clustered

data is smaller than the variance from a truly independent sample of observations. Therefore, treating clustered data as independent underestimates the sampling variance, which in turn produces underestimated standard errors. In other words, given that people in the sample are non-independent (or somewhat redundant with each other), the n_{eff} is smaller than the actual sample size for the study. The n_{eff} for a given sample is a function of the degree of non-independence in the sample and the number of people per cluster. The degree to which the effective sample size and the actual sample size differ determines the degree to which the standard errors from traditional statistical tests are underestimated. To estimate how much the clustered nature of the data impacts the standard errors, we must account for both the homogeneity within clusters (ICC) and the average cluster size ($\bar{n}_j$).

So, what is the effective sample size for Igor's sample? It is certainly less than 1000. One might be tempted to aggregate the data up to the household level and then use the mean sleep score of each household as the outcome variable, resulting in a sample size of 250. However, such an approach is overly conservative. There is still considerable variability in sleep hours within each household. By aggregating to the household level, we would lose all of the information about within-household variability in sleep time. So, aggregating to the cluster level both undersells the amount of information in the sample data and discards substantively interesting information about how different people within the same cluster differ from each other.

So, Igor faces a difficult question: is his sample more like a sample of 250 people, a sample of 1000 people or something in between? The n_{eff} provides the answer. Let's first consider two extremes. When the ICC is .00, then the n_{eff} is equal to the total number of observations. When the ICC is 1.00, observations within a cluster are complete replicates of each other, so sampling more than one unit per cluster is completely unnecessary. Thinking back to our sleep example, an ICC of 1.00 indicates that people within the same household receive identical amounts of sleep. Therefore, sampling more than one person per household would provide no additional information. In such an unlikely situation (at least for those of us who study humans!), the effective sample size would equal the number of clusters. When the ICC is larger than .00 but smaller than 1.00, the effective sample size is somewhere between the total number of people in the sample (as it is when the ICC = .00) and the number of clusters (as it is when the ICC = 1.00).

Computing the effective sample size n^eff

Computing the effective sample size requires both the ICC (ρ) and the average number of observations per cluster ($\bar{n}_j$). Using the effective sample size, we can adjust our standard errors to account for the non-independence.

Using the ICC and the average number of observations per cluster, first we compute the **design effect** (**DEFF**; Kish, 1965). The design effect is a ratio of the **sampling variability** for the study design compared to the sampling variability expected under a simple random sample (SRS). We calculate the DEFF using the following equation:

$$\text{DEFF} = \frac{\text{var}(design)}{\text{var}(SRS)} = 1 + \rho(\bar{n}_j - 1) \tag{1.2}$$

where $\bar{n}_j$ is the average sample size within each cluster and ρ is the ICC. If this ratio equals 1.00 (which only happens when the ICC = .00), then the clustering has no effect. However, DEFF greater than 1.00 indicates some degree of dependence of observations within clusters; this increases the actual Type I error rate above the nominal Type I error rate (α). Using the design effect, we can calculate the n_{eff}, or the sample size that we should use to more appropriately compute the standard errors for our study. The formula for n_{eff} is simply (Snijders & Bosker, 2012):

$$n_{eff} = \frac{N}{\text{DEFF}} = \frac{N}{1 + \rho(\bar{n}_j - 1)} \tag{1.3}$$

where $\bar{n}_j$ is the average cluster size and ρ is the ICC and N is the total sample size.

Now, let's calculate the DEFF and the n_{eff} for Igor's study, assuming that the ICC = .40. The DEFF for Igor's study is 1 + .4(4 – 1) = 2.2. This means that the n_{eff} for Igor's study is 1000/2.2, or 454.55, which is about half as large as the actual sample size. So, how do we fix Igor's error? We have two options: (1) we could use MLM techniques or (2) we could adjust the standard errors from a traditional statistical analysis to account for the non-independence in our data. To compute the corrected standard errors, we simply substitute the n_{eff} for n. To illustrate, let's correct the standard error of the mean for the degree of clustering in our sample. The standard error of the mean is the square root of the variance (σ^2) divided by the sample size, $\sqrt{\sigma^2/n}$, which equals the standard deviation divided by the square root of the sample size, $\sigma/\sqrt{n}$. In Igor's sample, the standard deviation in the number of sleep hours per night is 2.00. Therefore, the standard error using simple random sampling is = $2/\sqrt{1000}$ = 0.063. However, the effective sample size of Igor's sample is 454.22, which is much smaller than it would have been if he had sampled 1000 people from 1000 different households. We replace n with n_{eff} in the denominator of the standard error formula to correct the standard error of the mean. Replacing n with n_{eff}, $2/\sqrt{454.22}$ results in a standard error of 0.094. Thus, the corrected standard error is almost 50% larger than it would have been if we incorrectly assumed our sample of 1000 people were completely independent.

Alternatively, we can also adjust previously computed standard errors using the square root of the DEFF, called the *root design effect* (DEFT; Thomas & Heck, 2001). The DEFT indicates the degree to which the standard errors need to increase to account for the clustering (non-independence). Recall that the DEFF for our study was 2.2. The square root of 2.2 is 1.48 (DEFT). Multiplying 1.48 by the original standard error, 0.063, provides the corrected standard error, 0.094, which matches the standard error computed using n_{eff}.

To summarise, two factors influence the design effect: (1) the average cluster size (i.e. the average number of individuals per cluster) and (2) the ICC. Holding average cluster size constant, as the ICC increases, the design effect increases. Similarly, holding ICC constant, as the average cluster size increases, the design effect increases. The effective sample size is the actual sample size divided by the DEFF. In our sleep example, if the average cluster size were 10 (instead of 4) with an ICC of .40, the DEFF would be 4.6, resulting in a DEFT of 2.14 and an effective sample size 217.4. In this scenario, our corrected standard error estimate would be 0.063 * 2.14= 0.135, meaning the corrected standard error is over twice as large as the original standard error.

Figure 1.1 illustrates the effect of increasing average cluster size on the effective sample size. The actual sample size is 1000. This graph presents curves for two common ICC values for school-based research: .15 and .30. The *y*-axis depicts the drop in effective sample size as the average cluster size increases. Holding cluster size constant, n_{eff} is lower when the ICC is higher: n_{eff} is consistently lower when ICC = .30.

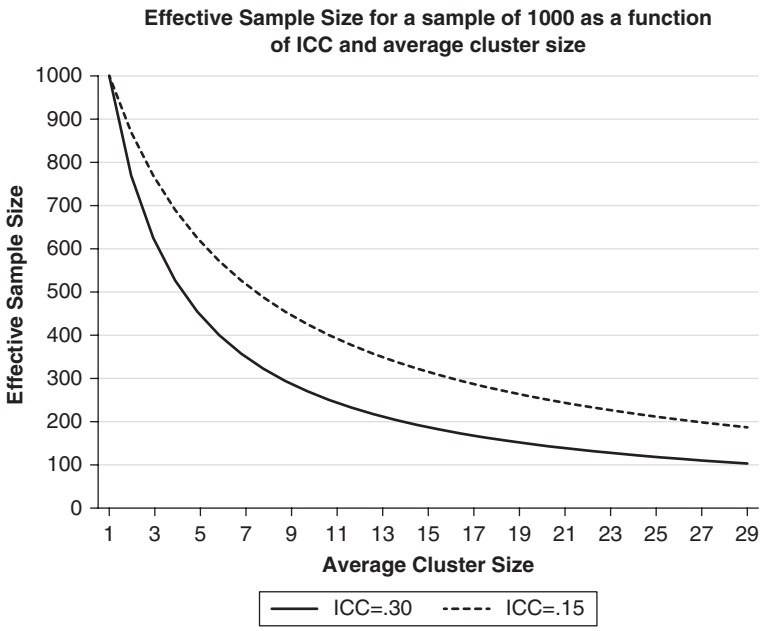

Figure 1.1 Effective sample size for a sample of 1000 as a function of intra-class correlation coefficient and average cluster size

Some researchers mistakenly believe that they can safely ignore small ICCs. However, small ICCs, coupled with large average cluster sizes, can still result in severely underestimated standard errors. Figure 1.2 illustrates the dangers of ignoring small ICCs when the average cluster size is large. The n_{eff} is a function of ICC and average cluster size for three very small ICC values: .01, .02 and .05, again assuming an original sample size of 1000. For example, with an ICC of .05 and an average of 50 units per cluster, the DEFF is 3.45, so the DEFT is 1.86 ($\sqrt{3.45} = 1.86$) and n_{eff} is 289.86. With an ICC of .01 and 50 people per cluster, n_{eff} is 671.14. With an ICC of .01 and a cluster size of 250 people, n_{eff} is 286.53. In general, if either the ICC or the average cluster size is large, then design effect is non-ignorable. These corrections, which enlarge the standard error and increase the *p*-value, have a major impact on tests of statistical significance. Luckily, when we use MLM, it is not necessary to correct standard errors. MLM produces standard errors that account for the dependency induced by clustering.

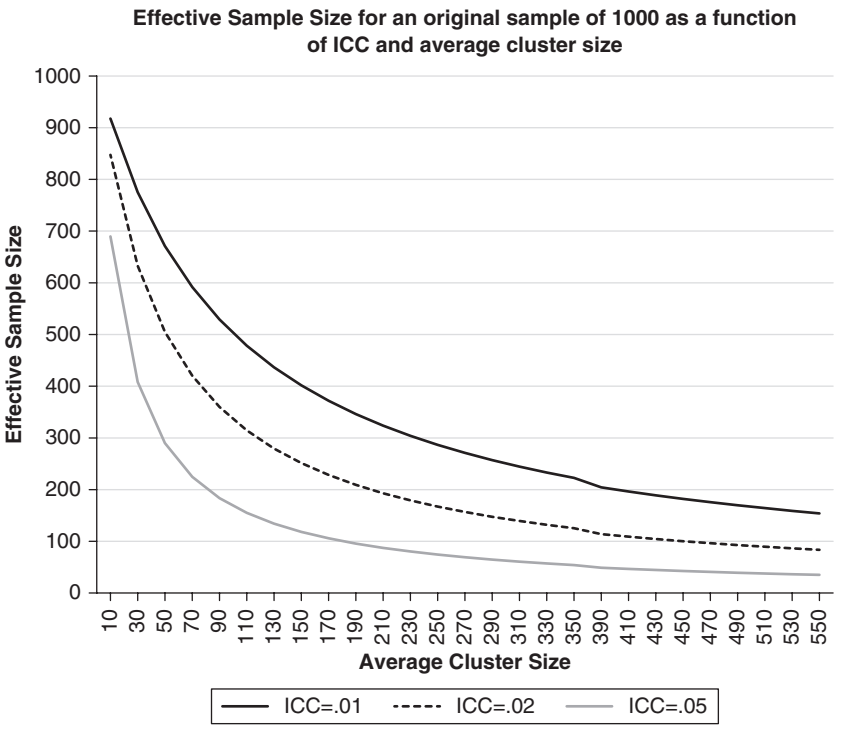

Figure 1.2 Effective sample size for an original sample of 1000 as a function of intra-class correlation coefficient and average cluster size. This figure demonstrates that even a seemingly-small intra-class correlation coefficient can have a large effect on standard errors and tests of statistical significance if the average cluster size is very large

Chapter Summary

- Observations within a given cluster often exhibit some degree of dependence (or interdependency). In such a scenario, violating the assumption of independence produces incorrect standard errors that are smaller than they should be. Therefore, inferential statistical tests that violate the assumption of independence have inflated Type I error rates: they produce statistically significant effects more often than they should.
- Multilevel modelling (MLM) allows researchers to model the relatedness of observations within clusters explicitly. The standard errors from multilevel analyses account for the clustered nature of the data, resulting in more accurate Type I error rates.
- Multilevel analyses can partition the variance in the outcome variable into between-cluster variance and within-cluster variance.
- Predictors at both the individual level (level 1) and the group level (level 2) may explain this between-cluster variability in the outcome variable. Level-1 predictors may explain within-cluster variability in the outcome variable.
- The intra-class correlation coefficient (ICC) measures the proportion of the total variability in the outcome variable that is explained by cluster membership. The ICC is also the expected correlation between two randomly drawn individuals from the same cluster (Hox et al., 2017).
- The effective sample size (n_{eff}) for a given sample is a function of the degree of non-independence in the sample and the number of people per cluster. The degree to which the effective sample size and the actual sample size differ indicates the degree to which the standard errors from traditional statistical tests are underestimated.

Further Reading

McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence: Part I. Understanding the effects of clustered data. *Gifted Child Quarterly, 54*(2), 152–155.
This article is a good supplement to the material covered in this chapter. It provides a conceptual introduction to the issue of clustering and dependence as well as an illustration of the effect of non-independence on the standard error.

Kreft, I. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.
This book provides a non-technical, accessible and practical introduction to multilevel modelling. The book also provides a broad overview of multilevel modelling, applications of multilevel modelling and its historical development.

2

MULTILEVEL MODELLING

A CONCEPTUAL INTRODUCTION

Chapter Overview

Having discussed the importance of taking the nested nature of data into account when conducting statistical analyses, we now introduce the multilevel model. This chapter provides a conceptual introduction to multilevel modelling (MLM). Chapter 3 provides additional guidance on building models in MLM and presents an applied example using real data.

Review of single-level regression

A multilevel regression model is still a regression model, so let's briefly review the standard regression equation before extending to the multilevel case. Typically, we represent a regression model with one predictor as $Y_i = \beta_0 + \beta_1(X_i) + r_i$, where Y_i = person *i*'s score on the outcome variable, *Y*. The intercept, β_0, represents the expected value of *Y* when *X* (the independent variable) is equal to 0. Generally, the intercept receives relatively little attention in multiple regression. However, in multilevel models, the intercept is often the star of the show! (We will have much more to say about this later.) The error term or residual, r_i, represents individual *i*'s actual score on the outcome variable (*Y*) minus their model-predicted score on the outcome variable ($\hat{Y}$), which is $\beta_0 + \beta_1 (X_i)$. A positive residual indicates that the person's actual score is higher than their predicted score, whereas a negative residual indicates that a person's actual score is lower than their predicted score. In multiple regression, we assume that these errors are normally distributed with a mean of 0 and a constant variance σ^2.

Regression model with no predictors

Let's begin with the simplest possible regression model: a standard regression model with no predictors: $Y_i = \beta_0 + r_i$. In this case, the intercept β_0 represents the expected value on the outcome *Y*; absent any other information, β_0 is the mean of *Y*. The error term r_i denotes the difference between person *i*'s actual score (Y_i) and his/her predicted score on *Y*; in this model, the person's predicted score is simply β_0, the mean of *Y*. Furthermore, with standard regression approaches, we make the assumption of independence, which means that we assume the r_i's are uncorrelated with each other.

Multilevel model with no predictors

Recall from Chapter 1, in a clustered sample (like Igor's), people within a given cluster are more similar to each other than to individuals from other clusters. Therefore, in clustered samples, we expect the r_i's to be correlated within clusters, but independent

across clusters. So, how does a multilevel model with no predictors differ from a multiple regression model with no predictors? Given that the residuals for observations within clusters co-vary, some of the variance in the dependent variable can be explained by cluster membership. Therefore, we introduce an additional error term, u_{0j}, to capture the portion of the residual that is explained by membership in cluster *j*. The residual for the intercept for each cluster (u_{0j}) represents the deviation of a cluster's intercept from the overall intercept. The term u_{0j} allows us to model the dependence of observations from the same cluster because u_{0j} is the same for every person within cluster *j* (Raudenbush & Bryk, 2002).

Returning to our sleep example, some of the variability in sleep time is between households, meaning that households differ in terms of their expected/mean sleep time. Households do vary in terms of their average sleep time, so we allow the intercept β_0, which is the mean of the outcome variable (sleep time) to vary across clusters. Conceptually, allowing the intercept to randomly vary across clusters is analogous to allowing separate intercepts for each cluster. Therefore, our level-1 equation is now $Y_{ij} = \beta_{0j} + r_{ij}$, where *j* indexes the cluster and *i* indexes the person. Instead of having only one intercept (as we did in the multiple regression equation), we now have *j* intercepts, one for each cluster. This means that person *i* in cluster *j*'s score on the outcome variable *Y* equals the expected **cluster mean** for cluster *j*, β_{0j}, plus person *i*'s deviation from this expected cluster mean, r_{ij}. So in Igor's sample, each household has its own intercept, β_{0j}, which is the predicted household (cluster) mean. Given that there are 250 clusters (households), there are also 250 β_{0j}'s or intercepts.

For simplicity, let's assume we have no predictors at level 2. The level-2 equation is then $\beta_{0j} = \gamma_{00} + u_{0j}$. In MLM, we refer to these β_{0j}'s as *randomly varying* intercepts. The randomly varying intercept, which was on the right-hand side of the **level-1 equation** (acting as a predictor of *Y*) is now on the left-hand side of the **level-2 equation** (acting as an outcome variable). The 250 intercepts (β_{0j}'s) are predicted by an overall intercept, γ_{00}, and a level-2 residual (error), u_{0j}, which captures the deviation of cluster *j*'s predicted intercept, β_{0j}, from the overall intercept, γ_{00}. Each of the *j* clusters has its own level-2 residual, u_{0j}, which allows each cluster to have its own intercept (β_{0j}). Rearranging the level-2 equation so that $u_{0j} = \beta_{0j} - \gamma_{00}$, it becomes clear that the level-2 residual, u_{0j}, is the difference between β_{0j} (the expected cluster mean for the outcome variable) and γ_{00} (the overall expected value on the outcome variable). Thus, our set of multilevel equations for a completely unconditional model is

$$\begin{aligned} Y_{ij} &= \beta_{0j} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \end{aligned} \tag{2.1}$$

The subscript for γ_{00} contains no *i*'s or *j*'s, meaning that γ_{00} is *fixed*: there is only one value of γ_{00}, the overall intercept. Because we have no predictors, γ_{00} is also the predicted

mean (average) on the outcome variable, Y. Notice that β_{0j} occurs in both equations. Therefore, we can substitute $\gamma_{00} + u_{0j}$ for β_{0j} to obtain one combined (or mixed) equation,

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \tag{2.2}$$

What does this mean? Person i in cluster j's score on Y (Y_{ij}) is equal to the overall expected/mean Y score (γ_{00}) plus the amount by which his/her cluster (cluster j) deviates from that overall mean Y score (u_{0j}) plus the amount by which he/she (person i in cluster j) deviates from his/her cluster mean (r_{ij}). So each person's score (Y_{ij}) equals the expected (predicted) mean (γ_{00}) plus their cluster's deviation from the overall mean (u_{0j}) plus their deviation from their own cluster's mean (r_{ij}).

Box 2.1

A note on multilevel versus combined equations

Conceptually, separating the equations from a multilevel model into multiple levels is often more intuitive than the combined equation. In reality, the multilevel model that is estimated is the **combined model**, and different statistical software packages require users to convey models in different formats. For example, users of SAS, Stata, R and SPSS must specify the combined model, whereas users of the software packages HLM, Mplus and MLWin can use the multiple-equation notation to estimate multilevel models. Thus, some 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. In this book, we tend to favour the multilevel equations, but we sometimes present the combined form as well.

For a more concrete example, let's return to Igor's sample. Imagine that, on average, people report sleeping 8 hours per night ($\gamma_{00} = 8$). Suzie lives in a house where the average number of sleep hours per night is 6 ($\beta_{0j} = 6$), but Suzie herself sleeps 7 hours per night ($Y_{ij} = 7$). Conceptually, $Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$ for Suzie would be $7 = 8 + (-2) + 1$. In a non-multilevel framework (with a non-clustered, simple random sample), the prediction equation for Suzie would simply be $Y_i = \beta_0 + e_i$, or $7 = 8 + (-1)$. The single-level regression equation contains only one error term, Suzie's deviation from the overall average (or predicted) score. The multilevel regression equation contains two residuals: (1) the deviation of Suzie's household from the overall mean (which in this case is –2) and (2) Suzie's deviation from her household mean (which in this case is +1). So, the overall mean (the overall intercept or predicted score) is the same in the multilevel and single-level frameworks above. What differs is our treatment of the residual(s).

Random effects and variance components

Without predictors, each person's score on the dependent variable is composed of three elements: (1) the expected mean (γ_{00}), (2) the deviation of the cluster mean from the overall mean (u_{0j}) and (3) the deviation of the person's score from his/her cluster mean (r_{ij}). In this equation, γ_{00} is a **fixed effect**: γ_{00} is the same for everyone. The u_{0j} term is called a **random effect** for the intercept because u_{0j} randomly varies across the level-2 units (clusters). In MLM, *fixed effects* are parameters that are fixed to the same value across all clusters (or individuals), whereas *random effects* differ (vary) across clusters (or individuals) (West et al., 2015).

Residual variances in MLM

Multilevel models and standard regression models do not differ in terms of their fixed effects. However, they differ in terms of the complexity of their residual variance/covariance structures. This more complex residual variance/covariance structure is at the heart of MLM. Therefore, understanding the meaning and utility of the included random effects is essential.

To account for the dependence/clustering, we break the residual into two pieces, u_{0j} and e_{ij}: u_{0j} captures the deviation of the cluster mean (intercept) from the overall mean (intercept), and r_{ij} captures the deviation of the individual's score from the mean for that individual's cluster. We can then compute variances for each of these residuals. (You may have noticed that we spend a lot more time thinking about intercepts and residuals in MLM than we ever did in ordinary least squares [OLS] regression!) The variance of r_{ij}, σ^2, represents the within-cluster residual variance in the outcome variable, and the variance of u_{0j}, τ_{00}, represents the between-cluster residual variance in the outcome.

We also make several important assumptions related to our model's residual variance terms: (a) the set of *u*'s is normally distributed with a mean of 0 and a variance of τ_{00}, (b) the set of *r*'s is normally distributed with a mean of 0 and a variance of σ^2 and (c) the within-cluster residuals (r_{ij}'s) and between-cluster residuals (u_j's) are uncorrelated. This last assumption allows us to cleanly partition the variance in the outcome variable into within- and between-cluster **variance components**. Therefore, in the simplest unconditional model with no predictors, the total variance in the outcome variable (var(Y_{ij})) equals the sum of the between-cluster variance, τ_{00}, and the within-cluster variance, σ^2. The ability to partition variance into within-cluster variance and between-cluster variance is one of MLM's greatest assets.

Intercepts as outcomes models

Because the intercepts vary across clusters, we can build a regression equation at level 2 to try to explain the variation in these randomly varying intercepts. For instance, in our sleep example, we could include household-level covariates to predict between-cluster variance in households' sleep time. For example, the number of dogs in the house or the average age in the household are potential level-2 covariates. Raudenbush and Bryk (2002) refer to these models as *means as outcomes models* because the level-2 model predicts differences in the intercepts across clusters (level-2 units).

The level-2 covariates may help to explain why some households sleep more than others. However, level-2 variables can never explain within-cluster variance (i.e. household-level variables cannot explain why certain members of the family sleep more or less than other family members). To explain within-cluster (level-1) variance, we need to include within-cluster (level-1) covariates.

Adding level-1 predictors

Now, let's consider a model in which there is one predictor at the lowest level (level 1). Imagine that we want to predict sleep hours using age. (The age–sleep relationship might actually be non-linear: people in middle adulthood might sleep less than children and older adults. However, for simplicity, let's assume a linear relationship between age and sleep time.) We regress sleep hours (Y_{ij}) on age (X_{ij}). Now our level-1 model is as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}\,(X_{ij}) + r_{ij} \tag{2.3}$$

Remember that in standard linear regression, the intercept is the predicted value on Y when all predictors are held constant at 0. Similarly, we interpret the intercept (β_{0j}) as the predicted mean sleep hours in cluster j when X_{ij} (age) is equal to 0. Because age is equal to 0 at birth, the intercept is the expected amount of sleep time for a newborn infant. The slope β_{1j} (the effect of age on sleep time) can vary by cluster, just like β_{0j} does. If we allow β_{1j} to randomly vary by cluster, β_{1j} becomes an outcome variable in a level-2 equation and has its own residual term, u_{1j}. Equation (2.4) contains the multilevel model with a randomly varying intercept and a randomly varying slope.

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}\left(X_{ij}\right) + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{2.4}$$

In Equation (2.4), γ_{00} represents expected (predicted) number of sleep hours when age = 0, and γ_{10} represents the average effect of age on sleep time across the entire sample. So, if age is measured in years, we expect a γ_{10}-hour change in sleep for every additional year. The error term, u_{1j}, represents the difference between the average slope and cluster *j*'s slope. In our example, u_{1j} is the difference between cluster *j*'s age–sleep slope and the overall age–sleep slope. If the 'effect' of age on sleep time does not vary across clusters, then all clusters should have the same (or very similar) age–sleep slopes. In such a scenario, the value of u_{1j} for each cluster would be 0 (or near zero), and the variance of u_{1j} would also be approximately 0. If the slope is the same across all clusters (i.e. the slope does not vary across clusters), it is not necessary to estimate a randomly varying slope. Instead, we could estimate a model in which the intercept for sleep time randomly varies across clusters, but the age–sleep slope (the effect of age on sleep) remains constant across clusters. In that scenario, our multilevel model equations would be

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(X_{ij}) + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} \end{aligned} \tag{2.5}$$

Again, using substitution to combine these level-specific equations into one mixed-format model produces the combined model for each set of multilevel equations above. If the age–sleep slope does not randomly vary across clusters, the combined model is simple. Substituting $\gamma_{00} + u_{0j}$ for β_{0j} and γ_{10} for β_{1j}, the mixed-format equation is

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij}) + u_{0j} + r_{ij} \tag{2.6}$$

such that person *ij*'s score on *Y* is a function of γ_{00}, the overall intercept (the predicted score when $X_{ij} = 0$, which in this case is when age = 0), γ_{10} (the slope of age on sleep hours) multiplied by person *ij*'s age (X_{ij}), the deviation of his/her household's intercept (the predicted number of sleep hours at age = 0) from the overall intercept (u_{0j}), and r_{ij}, the deviation of person *ij*'s score from his/her model-predicted score.

If the age–sleep slope does randomly vary by cluster, then substituting $\gamma_{10} + u_{1j}$ for β_{1j} results in the following combined equation:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij}) + u_{0j} + u_{1j}(X_{ij}) + r_{ij} \tag{2.7}$$

Now person *ij*'s score is a function of γ_{00}, the overall intercept (the predicted score when $X_{ij} = 0$), γ_{10}, the slope of age on sleep hours, multiplied by person *ij*'s age (X_{ij}), the deviation of his/her household's intercept from the overall intercept (u_{0j}), the deviation of his/her household's slope from the overall slope (u_{1j}) multiplied by person *ij*'s age (X_{ij}) and the deviation of person *ij*'s score from their model-predicted score, r_{ij}.

Allowing the age–sleep slope to randomly vary across households by including a *random effect* for the slope (u_{1j}) specifies a model in which the age–sleep slope is different for different households. Therefore, in some households, there could be no relationship between age and sleep time, resulting in an age–sleep slope of 0; in other households, the sleep slope could be negative, indicating that older members of the household tend to sleep less than younger members of the household. Finally, the age–sleep slope could be positive, indicating that older members of the household tend to sleep more than younger members of the household. The fixed effect, γ_{10}, indicates the expected (average) value of the age–sleep slope across the entire sample. The variance in the age–sleep slope, var(u_{1j}), indicates how much households vary from the overall average. If the variance of u_{1j} is large, there is a lot of between-household variability in the age–sleep slope. In contrast, if the variance of u_{1j} is 0, then there is no variability across households in terms of their age–sleep slopes: in this case, we would want to fix u_{1j} to 0, as that would greatly simplify our model.

Figure 2.1 illustrates the concept of randomly varying intercepts and randomly varying slopes by graphing the relationship between a hypothetical independent variable (X), such as age, and a hypothetical dependent variable (Y), such as hours of

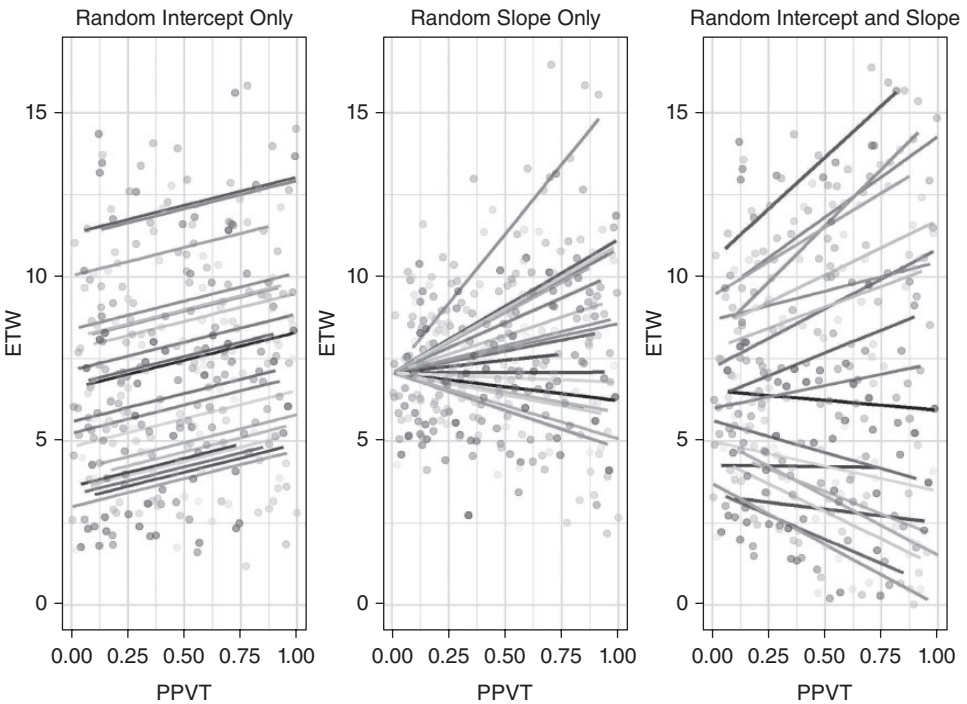

Figure 2.1 (1) Randomly varying intercepts, (2) randomly varying slopes and (3) randomly varying intercepts and slopes

Note. ETW = expressive target word assessment; PPVT = Peabody Picture Vocabulary Test.

sleep, under three conditions: (1) when the intercept randomly varies but the slope is fixed, (2) when the slope randomly varies but the intercept is fixed and (3) when both the slope and the intercept randomly vary. (Note each line in Figure 2.1 represents the regression line for a specific cluster.) When only the intercepts randomly vary, all clusters have equal slopes, but they differ in their intercepts, producing parallel regression lines. Likewise, when only the slopes randomly vary, all clusters have equal intercepts, but they differ in their slopes. Therefore, all regression lines appear to originate in the same location (at $X = 0$) but then diverge. Lastly, when the slopes and intercepts randomly vary, each cluster has its own unique intercept and slope.

Full contextual (slopes-as-outcomes) model

The **full contextual/theoretical model** contains both level-1 and level-2 predictors. Level-2 predictors may help to explain between-cluster differences in the intercept (the expected value of the outcome variable when all level-1 variables are held constant at 0). Level-2 predictors may also help explain between-cluster differences in level-1 slopes. In other words, the level-2 variable helps to predict how the relationship between the level-1 predictor and the outcome variable differs across clusters. Returning to our example, age is a level-1 predictor of sleep hours. We could include a household-level variable, such as the average noise level in the home, to predict the average number of sleep hours within the household (the intercept). A level-2 variable (i.e. the noise level in the house) could also predict a level-1 slope (i.e. the age–sleep hours slope). We refer to a level-2 predictor of a level-1 slope as a **cross-level interaction** because it represents an interaction between a level-2 variable and a level-1 variable. In this example, the cross-level interaction indicates whether the noise level in the house moderates the relationship between age and sleep hours.

Variance–covariance components

The γ terms are the *fixed effects* and the u terms are the *random effects*. All of the γ terms could be estimated using multiple regression models with interaction terms. However, the u terms, the random effects, are unique to mixed/multilevel models. Multilevel techniques allow us to model, estimate and test the variances (and covariances) of these random effects (also known as *variance components* and denoted by the symbol, τ_{qq}). Specifically, τ_{00} represents the variance of the randomly varying intercepts (u_{0j}), τ_{11} signifies the variance of the first randomly varying slope (u_{1j}) and so on. In addition, we generally allow the random effects (within a given level) to co-vary with each other.

Therefore, in our simple example above, τ_{01} represents the covariance between residuals for the randomly-varying intercept and the randomly varying slope.

$$\text{var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} \tag{2.8}$$

If we standardise τ_{01}, it represents the correlation between the residuals of the intercept and slope. If τ_{01} is positive, then clusters with more positive intercepts also tend to have more positive/less negative slopes. If τ_{01} is negative, then clusters with more positive intercepts tend to have less positive/more negative slopes. Although we do allow random effects to co-vary within a given level, we assume that the residuals are uncorrelated across the levels of our analysis. As a result, although τ_{00} and τ_{11} are allowed to co-vary, we assume that both τ_{00} and τ_{11} are uncorrelated with σ^2.

Advice on modelling randomly varying slopes

Depending on the researcher's theoretical framework and 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. A random-coefficients model is a model in which one or more level-1 slopes randomly vary (Raudenbush & Bryk, 2002). Although our simple example contains only one level-1 variable (age), often multilevel models contain several level-1 variables. For example, many multilevel educational studies in which students are nested within schools include a variety of level-1 control variables, such as gender, race/ethnicity (which is often a set of four to six dummy-coded variables), free lunch status, English learner status and special education status. A set of control variables could easily include 10 or more level-1 variables. In such a situation, the researcher must decide which level-1 slopes to allow to randomly vary across schools and which level-1 slopes to fix to a single value across all schools. Why not allow all 10 level-1 covariates to randomly vary across schools? First, remember the structure of the residual covariance matrix. The unstructured tau (τ) matrix contains a variance for the randomly varying intercept and every randomly varying slope as well as all possible covariances among the slopes and intercepts. Therefore, the number of unique variance–covariance components in the tau matrix is equal to $r(r + 1)/2$, where r equals the number of random effects. As we saw earlier, with a **random intercept** and a **random slope**, the Tau matrix contains (2 * 3)/2 = 3 parameters (two variances and a covariance). However, in a model that contains five randomly varying slopes and a randomly varying intercept, the tau matrix contains (6 * 7)/2 = 21 unique parameters, and the tau matrix for a model with 10 randomly varying slopes and a randomly varying intercept contains

(11 * 12)/2 = 66 unique parameters. In other words, in a model with 10 randomly varying slopes, we need to estimate a total of 67 different residual parameters: σ^2 and an 11 * 11 tau matrix, containing 66 unique level-2 variance components. Partitioning the residual variance in a model into 67 separate pieces feels like a Sisyphean task. (Remember, in multiple regression, we estimate just one residual variance parameter.)

Raudenbush and Bryk (2002) cautioned against succumbing to the 'natural temptation to estimate a "saturated" level-1 model' in which all level-1 predictors are specified to have randomly varying slopes (p. 256). We cannot overstate the importance of this advice. It is essential to be parsimonious when specifying randomly varying slopes for several reasons:

1 First, as demonstrated above, adding random slopes radically increases the complexity of the model.
2 There is an upper limit on the number of random slopes based on the sample size at level-1. Minimally, cluster size must be larger than the number of variance components for the model to be identified. Therefore, in dyadic data, it is only possible to estimate one random effect. Our sleep example had a cluster size of four, so the maximum number of potential random effects is three (presumably a randomly varying intercept and two randomly varying slopes). This does not mean that it is a good idea to estimate such nearly saturated models; we generally prefer to fit level-1 models in which the number of random effects is comfortably less than the cluster size.
3 It is common to experience convergence problems when trying to estimate randomly varying slopes that are unnecessary. Multilevel models that contain a random slope that has no between-cluster variance often fail to converge (or require thousands of iterations to converge). Because variances cannot be less than 0, trying to estimate randomly varying slopes that are actually 0 in the population often leads to boundary issues, resulting in models that fail to converge. Unfortunately, such results may not provide guidance about which of the random effects to eliminate (McCoach et al., 2018).

Of course, it is easy to think of at least one logical reason that each slope in a multilevel model might randomly vary, and it is tempting to allow most or all of them to do so 'just to see' what happens. However, we implore you – don't do it! In our experience, some people who learn about randomly varying slopes become 'greedy' and want to be able to allow every slope parameter in large regression models to randomly vary across clusters. Attempting to estimate and interpret dozens of residual variance–covariance components is unrealistic and unreasonable under most circumstances.

We recommend including random slope effect only if the randomly varying slope is central to your research question or if you have compelling evidence from prior research that the slopes are likely to randomly vary. Use randomly varying slopes carefully, sparingly and cautiously: be judicious and parsimonious about which random slopes to include in multilevel models. Also, eliminate any unnecessary

random effects for level-1 coefficients that do not vary across level-2 units (McCoach et al., 2018).

Centring level-1 predictors

In regression models, we often *centre* covariates for both substantive and analytic reasons. As mentioned earlier, the intercept is the predicted value of the outcome variable when all of the predictor variables are held constant at 0. So, in our sleep example, the intercept for sleep was the predicted sleep hours at age 0. In single-level regression, one common strategy is to *centre* continuous predictor variables by subtracting the mean of the variable ($\bar{X}$) from each person's score (X_i). This transforms person *i*'s score on *X* into a deviation score, which indicates how far above or below the mean person *i* scored. Therefore, the mean of a centred variable is 0 and the variance is the same as the variance of the score in its original metric (because all scores change only by a single constant value, the mean). In single-level regression, centring continuous covariates is especially important when including interaction terms. The choice of centring influences the main effects for the predictor variables included in the interaction term: the regression coefficient is the predicted effect of *X* on *Y* when the other predictor variable in the interaction term equals 0 (Aiken & West, 1991).

In MLM, we may centre continuous predictor variables for substantive and/or analytic reasons. First, centring continuous covariates allows for a more substantively useful and interpretable intercept. Second, the magnitude of the between-person (residual) variance in the intercept, τ_{00}, and the correlation between the intercept and any randomly varying slopes is dependent on the location of the intercept.

In organisational applications of MLM, the two main centring techniques for lower-level covariates are **grand mean centring** and **group mean centring**. Grand mean centring subtracts the overall mean of the variable from all scores. Therefore, the grand mean–centred score captures a person's standing relative to the full sample. Group mean centring subtracts the cluster's mean from each score in the cluster. As such, the transformed score captures a person's standing relative to their own cluster.

As an example, let us grand mean and group mean centre age (X_{ij}) for person *i* in cluster *j*. In our example, the grand mean represents the mean age across all individuals *i* and all clusters *j* ($\bar{X}_{\bullet\bullet}$), and the cluster mean represents the mean age of all individuals *i* in a household *j* ($\bar{X}_{\bullet j}$). To grand mean centre age, we subtract the mean age in the entire sample from each person *ij*'s age ($X_{ij} - \bar{X}_{\bullet\bullet}$), so under grand mean centring X_{ij} is person *ij*'s deviation from the average age in the entire sample ($\bar{X}_{\bullet\bullet}$). To group mean centre age, we subtract the average household age from each person's age ($X_{ij} - \bar{X}_{\bullet j}$); so, under group mean centring, X_{ij} is person *ij*'s deviation from his/her household's average age ($\bar{X}_{\bullet j}$).

Obviously, the decision about how to centre independent variables has major implications for the interpretation of the intercept. Grand mean centring age sets the intercept at the overall mean. This holds age constant at the overall mean, thereby controlling for age. When grand mean centring age, the randomly varying intercept, β_{0j}, denotes the predicted number of sleep hours for household *j* assuming that this household's average age is the same as the overall average age. The intercept is the predicted number of sleep hours in household *j*, holding age constant at the overall mean ($\bar{X}_{\bullet\bullet}$). Person *i* in cluster *j*'s score on the grand mean–centred *X* variable represents the deviation of that person's score from the overall average. In this case, grand mean–centred age represents each person's deviation from the average age across the entire sample. Grand mean centring represents a simple linear transformation of the original variable.

One problem with grand mean centring arises when no one in a given cluster has scores near the overall mean. In such cases, the intercept for that cluster is extrapolated outside the range of data for the cluster. For example, if the average age across households is 40, but in household *j*, the four members are 55, 55, 65 and 65 years old, then the grand mean–centred scores are 15, 15, 25 and 25, respectively. No one in the household has a centred score near 0. Thus, the intercept in household *j* is the predicted sleep score for a 40-year-old, even though there are no 40-year-olds in that household. For a detailed discussion of the statistical and interpretational issues that such extrapolation can cause, see Raudenbush and Bryk (2002).

On the other hand, if we group mean centre age, then the randomly varying intercept (β_{0j}) is the mean number of sleep hours in household *j*. Having subtracted each cluster's own mean ($\bar{X}_{\bullet j}$) from each score, the mean of the cluster-mean centred age variable is 0 in every cluster. Therefore, the randomly varying intercept for each cluster is the mean (expected/predicted) number of sleep hours in that household (*j*). Person *i* in cluster *j*'s score on the group mean–centred *X* variable represents the deviation of their score from their cluster's average score. In this case, group mean–centred age represents each person's deviation from their household's average age. So, in a household where the ages are 55, 55, 65 and 65 years, the household's mean age is 60 years. To group mean centre, we subtract 60 from each score, producing group mean–centred scores of –5, –5, 5 and 5, respectively. The mean of the group mean–centred variable is 0 in every cluster, so the overall intercept is the mean of cluster means: it is the overall average household sleep time.

Important guidance on group mean centring

A group mean–centred score provides information about individuals' relative standing as compared to their cluster, but it provides no information about the individual

or the group's relative standing as compared to the overall sample. So, for example, grand mean–centred scores of 15, 15, 25 and 25 indicate that two of the members of the household are 15 years older than the sample average and two of the members of the household are 25 years older than the sample average. In contrast, the group mean–centred scores of –5, –5, 5 and 5 tell us nothing about how the ages in this household compare to ages in the other households in the sample.

Group mean centring removes between-cluster variation from the level-1 covariate, so the variance of a group mean–centred variable provides an estimate of the pooled within-cluster variance (Enders & Tofighi, 2007). Using group mean centring, we can partition the variance in the predictor, the outcome and the relationship between the predictor and the outcome into within-cluster and between-cluster components.

However, using group mean centring does not preserve information about between-cluster differences on the *X* variable. Using a different cluster mean to centre each cluster results in a centred *X* variable that contains information about how much a person deviates from his/her group, but it contains no information about how much the person deviates from the overall mean on *X*. Therefore, when using group mean centring, be sure to introduce the aggregate of the group mean–centred 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 between-cluster variability in the *X* variable would be lost – a considerable but avoidable drawback. In our age example, the grand mean–centred mean age for our cluster, +20, provides information indicating that the average age in this cluster is 20 years older than the average age in the overall sample. In contrast, the group mean–centred age for our cluster (and every other cluster in the sample) is 0. However, adding the cluster mean into the model as a level-2 predictor preserves between-cluster component of the age variable. Finally, in group mean centring, a different cluster mean is subtracted in every cluster. Therefore, group mean centring is not a simple linear transformation, and it does not produce results that are statistically equivalent to the uncentred and grand mean–centred results.

Within the multilevel literature, some debate exists about whether to use grand mean centring or group mean centring. Because centring decisions affect the interpretations of important model parameters involving the intercept, it is important to carefully and thoughtfully decide if and how to centre covariates. The decision to use grand mean or group mean centring may vary depending on the context of the study, the research questions asked and the nature of the variables in question. 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 centring may be an appropriate choice. On the other hand, when level-1 variables are of primary research interest or for research on contextual and

compositional effects, group mean centring may be more appropriate. In addition, group mean centring aids in the computation of variance explained (***R*-squared**) measures, a point we discuss more fully in Chapter 3. To preserve between-cluster information from the covariate, we recommend including the aggregates of group mean–centred variables at level 2.

What about centring level-2 variables? Grand mean centring is the only available option at level 2. As a general rule, it is advisable to grand mean centre all level-2 continuous variables. When using level-2 variables as part of a cross-level interaction, grand mean centring is especially important. However, even for level-2 variables that predict only randomly varying intercepts (not randomly varying slopes), grand mean centring the level-2 variable usually facilitates interpretation of the intercept. When reporting MLM results, it is important to explain centring decisions and procedures and to interpret the parameter estimates accordingly. See Enders and Tofighi (2007) for an excellent discussion of centring in organisational multilevel models.

Estimation

This book does not delve into the computational details required to actually estimate multilevel models. However, it is helpful to conceptually understand the analytic challenges of multilevel data and the estimation strategies that MLM employs.

MLM does not require balanced data: the number of units per cluster can vary across clusters. In fact, there is no minimum or maximum number of units per cluster, and multilevel models can easily accommodate data sets that include some clusters with very few level-1 units and other clusters with very large numbers of level-1 units. MLM employs a variety of estimation strategies to handle unbalanced data.

To keep things simple, we contextualise this discussion in the context of the unconditional random effects model: $Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$. Let's start by identifying a simple but important issue that arises when estimating parameters from unbalanced data: how to determine the expected value of the outcome variable (γ_{00}) from multiple clusters of multiple sizes. In non-clustered data, the sample mean provides our 'best guess' about the population mean (the *expected value*). In clustered data, what is the expected value of the outcome variable, γ_{00}? Let's imagine that we have randomly sampled 100 schools, and the sample sizes within the schools vary widely: the smallest cluster size is two students and the largest cluster size is 1000 students. How should we determine expected achievement?

One option would be to ignore clustering and compute the sample mean. In such a scenario, every person is weighted equally; however, the schools with larger numbers of sampled students have a much larger influence on the expected mean than the

small schools do. On the other hand, we could compute the mean of school means. However, in that case the schools with very few students have an outsized influence on the expected mean. (In addition, the school mean that is computed from a school with 1000 students is likely to be a much better estimate of the school's performance than a school mean that is computed from only two students, a point to which we return when we discuss empirical Bayes estimates.)

In MLM, larger clusters do have a larger influence on the expected mean. However, the ICC tempers that influence. Remember, the ICC provides valuable information about the proportion of between-school variance in the outcome, which indicates the degree of dependence (redundancy) within a cluster. If the ICC is 0, then students within a given school are no more similar to each other than students from different schools. In such a situation, taking the sample mean ignoring clustering seems reasonable. On the other hand, if the ICC is 1.0, all students in a school are complete replicates of each other. In such a scenario, the mean of school means might be of greater interest, given the deterministic nature of within-school performance. When the ICC is 0, the influence of each cluster on γ_{00} is proportional to its cluster size. When the ICC is 1.0, each cluster has an equal influence on γ_{00}, regardless of its size (Snijders & Bosker, 2012). In reality, the ICC lies between 0 and 1.0. So γ_{00} is a compromise between a proportional weighted average (as it would be when ICC = 0) and a mean of cluster means (as it would be when ICC = 1). The higher the ICC, the more γ_{00} approaches a mean of cluster means; the lower the ICC, the more γ_{00} approaches a proportional weighted average.

Conceptual introduction to maximum likelihood

Both MLM and structural equation modelling (SEM) use maximum likelihood (ML) estimation techniques. In ML estimation, we estimate parameters that maximise the probability of observing our data. This section provides a very rudimentary, conceptual introduction to ML. The probability of observing an event implicitly assumes a model. We make statements about the probability of observing some event, based on the model parameters. For example, take the case of a coin toss. Everyone knows that the probability of tossing a head is .5. Let's formalise this notion. Our model contains one parameter, p, the probability of tossing a head, and that parameter p is equal to .5. In probability, we know the value of the parameter, and we try to predict future outcomes based on that known parameter. The likelihood, on the other hand, turns probability on its head. With likelihood, we already have the data, and we try to determine the most likely value for a parameter, given the data. The goal of ML estimation is to find the set of parameter values that makes the actual data most likely to have been observed. So imagine we know nothing about the probability of tossing

heads, but we want to use data empirically to determine the value for that parameter, so we flip a coin 100 times. The coin lands on heads 55 times and on tails 45 times. Then we can ask – what is the most likely parameter value for p, the probability that I will flip a head, given the data that I have collected? The answer in this case would be .55 (not .50): the parameter value of $p = .55$ maximises the probability of observing our results. For a much more detailed and nuanced discussion of likelihood estimation, see Myung (2003).

Maximum likelihood estimation in MLM

The most common estimation techniques for estimating variance components for multilevel models with normal response variables are full information maximum likelihood (FIML) and restricted maximum likelihood (REML).

In FIML, the estimates of the variance and covariance components are conditional upon the point estimates of the fixed effects (Raudenbush & Bryk, 2002). FIML chooses estimates of the fixed effects, the level-2 variance–covariance components (T) and the level-1 residual variance (σ^2) 'that maximise the joint likelihood of these parameters for a fixed value of the sample data, Y' (Raudenbush & Bryk, 2002, p. 52). Thus, the number of parameters in the model includes both the fixed effects and the variance–covariance components. In contrast, REML maximises the joint likelihood of the level-2 variance–covariance components (T) and the level-1 residual variance (σ^2) given the observed sample data, Y. Thus, when estimating the variance components, REML takes the uncertainty due to loss of degrees of freedom from estimating fixed parameters into account, while FIML does not (Goldstein, 2011; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

When the number of clusters (level-2 units) is large, REML and FIML results produce similar estimates of the variance components. However, when there are small numbers of clusters, the FIML estimates of the variance components ($\tau_{qq'}$) are smaller than those produced by REML. With few clusters, FIML tends to underestimate variance components, and the REML results may be more realistic (Raudenbush & Bryk, 2002). A simple formula, $(J - F) / J$, where J is the number of clusters and F is the number of fixed effects in the model, provides a rough approximation of the degree of underestimation of the FIML estimates (Raudenbush & Bryk, 2002). For example, when estimating a model with three fixed effects using a sample containing observations from 20 clusters, we estimate that the level-2 variance components are .85 = ((20 – 3)/20) as large in FIML as they are in REML; this means the FIML variance components are underestimated by 15%. However, there is an advantage to using FIML: it allows us to compare the fit of two different models, as we explain in the section 'Deviance and Model Fit'.

Reliability and estimation of randomly varying level-1 coefficients

The randomly varying level-1 coefficients (β_{0j}, β_{1j},…, β_{qj}) are not parameters in the model; they are a function of the fixed effects and the cluster-level residuals. Generally, standard MLM uses empirical Bayes estimation to generate the 'best estimates' of β_{qj} (Raudenbush & Bryk, 2002). The computation of empirical Bayes residuals follows a different logic and process than the computation of OLS residuals. Below, we present a very simple, conceptual introduction to empirical Bayes estimation of the randomly varying level-1 intercept for an unconditional random effects model. For a more detailed description, we recommend Goldstein (2011) and Raudenbush and Bryk (2002).

For simplicity, consider the estimation of the randomly varying intercept, β_{0j}, the 'true cluster mean' under the simplest model, the random effects ANOVA model, $\beta_{0j} = \gamma_{00} + u_{0j}$. We do not know the true cluster mean for cluster *j*; cluster *j* has sample size n_j. MLM allows for unbalanced data: the within-cluster sample sizes (n_j) can vary greatly across clusters. Some clusters could have many observations (i.e. – n_j is large); other clusters could have small n_j.

How could we estimate the *true* mean for cluster *j*? The sample mean, $\bar{Y}_{.j}$, provides an estimate of the *true* mean. However, the smaller the cluster size (n_j), the less confidence we should have in using $\bar{Y}_{.j}$ (the cluster's observed mean) as an estimate of the cluster's *true* mean. In the most extreme situation, imagine we had no observations from cluster *j* with which to estimate the *true* cluster mean. With a sample size of 0, what is our best guess about the true mean of cluster *j*? It is the overall mean, γ_{00}. Why? We know nothing about this cluster, but we have information about lots of other clusters. Our best guess for the mean of this cluster is the overall mean (expected value). So, there are two potential competing estimates for the true mean for a cluster: the overall mean (expected value) across the entire sample, which allows us to 'borrow' information from other clusters to estimate the true mean of cluster *j*, and the observed sample mean of cluster *j*, which contains some degree of error or imprecision. As the sample size in cluster *j* increases, the precision with which we can estimate the true mean from the sample mean increases; there is less error in our measurement of the true cluster mean based on the sample mean. Of course, another factor influences our ability to estimate the true school mean from the sample mean: the ICC. Again, imagine an extreme example: if the ICC were 1.0, every observation within a cluster is a replicate of every other observation. When there is very little or no within-cluster variance, $\bar{Y}_{.j}$ is an especially good estimate of the true mean of cluster *j*. When there is a great deal of within-cluster variance, $\bar{Y}_{.j}$ is a poor estimate of the true cluster mean (especially with small sample sizes; Raudenbush & Bryk, 2002).

Reliability of Cluster *j*

Our two potential estimates of the *true* cluster mean are γ_{00} and $\bar{Y}_{.j}$. Empirical Bayes estimation combines these estimates of the true cluster mean, based on the reliability of cluster *j*. The reliability of cluster *j* incorporates three pieces of information: the within-cluster variability (σ^2), the between-cluster variability (τ_{00}) and the number of observations per cluster, n_j (Raudenbush & Bryk, 2002):

$$\text{Reliability of } \hat{\beta}_{0j} = \frac{\tau_{00}}{\tau_{00} + \sigma^2 / n_j} \tag{2.9}$$

When the reliability in cluster *j* is higher, more weight is placed on the sample mean as the estimate of the true school mean. When the reliability of cluster *j* is lower, more weight is placed on γ_{00} as an estimate of the true school mean. Holding between- and within-school variance constant, larger cluster sizes (n_j's) result in higher reliability. Each cluster has its own estimate of reliability; however, variance estimates τ_{00} and σ^2 remain constant across clusters. Therefore, larger clusters have larger reliability estimates. Nevertheless, larger between-cluster variance (relative to within-cluster variance) also increases reliability. In other words, reliability is higher when the group means vary substantially across level-2 units (holding constant the sample size per group). So, increasing group size, increasing homogeneity within clusters and increasing heterogeneity between clusters all increase reliability. The formula for the ICC, τ_{00} / (τ_{00} + σ^2), features prominently in the reliability formula above. With a bit of algebra, we can re-express the reliability formula in terms of ICCs (Raudenbush & Bryk, 2002). Larger ICCs, which indicate that within-cluster group variance is small relative to between-cluster variance, result in higher reliability. Although reliability can range from 0 to 1, the lower bound for the reliability in any given sample is the ICC, and that occurs when n_j = 1.

Empirical Bayes estimates of randomly varying parameters

Imagine we need to estimate the true political attitudes for a set of counties, and we have an incomplete set of information. In most counties, pollsters randomly sampled 1000 or more respondents. However, in one county, pollsters randomly sampled only two respondents. In the counties where the pollsters sampled 1000 respondents, the best guess about the true political attitudes would be near the sample mean for the 1000 respondents. In the county where the pollsters sampled only two of the respondents, we can compute the sample mean. But how confident would we be that the mean of the two respondents accurately reflects the true political attitudes in that county?

If one of the two people in the sample is extreme, our sample mean could actually be a very poor estimate of the true political attitudes in the county. Imagine an even more extreme example: what if the pollsters missed one county entirely? What would be our best estimate of the political attitudes in that county? There are two logical possibilities for estimating the true county political attitudes. One is the sample mean in the county, and this seems like a good estimate of the true mean in counties where we have many observations (more information). However, in the counties without much information, what is our best guess about the county's political attitudes? We could use the overall mean across all of the counties as an estimate of the county's political attitudes. If we know nothing else about the county and we have no information from the county, using the overall mean provides our best estimate. What do we do for the county with only two respondents? The small sample of respondents does give us some information about the political attitudes in the county, but we cannot completely trust that the sample mean of those two respondents provides a good estimate of the true county mean. In such a situation, we could use a combination of the overall mean and the sample county mean to derive an estimate of the true mean for the county. To do so, we would want to give more weight to the cluster (county) mean when we have more information, and we would want to place more weight on the overall mean (expected value) when we have less information from the cluster (county). Conceptually, this is the essence of empirical Bayes estimates of the randomly varying parameters (intercepts and slopes).

Again, assuming an unconditional random effects ANOVA model, the empirical Bayes estimate of the true cluster mean (β_{0j}^{*}) weights the two potential estimates for each cluster as a function of the reliability for that cluster.

$$\beta_{0j}^{*} = \lambda_j \bar{Y}_{.j} + \left(1 - \lambda_j\right)\hat{\gamma}_{00} \tag{2.10}$$

where λ_j is the reliability of the sample mean, $\bar{Y}_{.j}$ (Raudenbush & Bryk, 2002) and $\hat{\gamma}_{00}$ is the expected value of the intercept (which is the expected value of the outcome variable in the unconditional random effects ANOVA model).

The sample mean ($\bar{Y}_{.j}$) is weighted by the reliability for that cluster (λ_j); the model-based mean (expected value, $\hat{\gamma}_{00}$) is weighted by 1 minus the reliability ($1 - \lambda_j$). Thus, the empirical Bayes estimate of the true cluster mean is a compromise between the sample mean and the model-based mean, and the degree to which we trust the sample-based mean ($\bar{Y}_{.j}$) determines the weight (λ_j) that we place on the sample-based mean ($\lambda_j \bar{Y}_{.j}$). Our lack of trust in the sample mean ($1 - \lambda_j$) determines the weight that we place on the overall expected value. Thus, the higher the reliability of the estimate for cluster *j*, the more weight is placed on the sample mean ($\bar{Y}_{.j}$) as the estimate of the true cluster mean, β_{0j}. In contrast, the lower the reliability of cluster *j* estimate, the more weight is placed on the model-based estimate $\hat{\gamma}_{00}$ as the estimate of the

true cluster mean. In the extremes, if the reliability were 1, the sample mean would be the estimate of the true cluster mean. If the reliability were 0, $\hat{\gamma}_{00}$ would serve as the estimate of the true cluster mean. Sometimes the empirical Bayes estimators are referred to as shrinkage estimators because the $\bar{Y}_{.j}$ estimate is 'shrunken' towards the model-based estimate; empirical Bayes residuals are like OLS estimates of the residuals which are 'shrunken' towards 0 (Raudenbush & Bryk, 2002).

Deviance and model fit

Deviance

Using ML to estimate the parameters of the model also provides the likelihood, which easily can be transformed into a deviance statistic (Snijders & Bosker, 2012). The **deviance** is –2 multiplied by difference of the log likelihood of the specified model and the log likelihood of a saturated model that fits the sample data perfectly. Therefore, deviance is actually a measure of the *badness* of fit of a given model: higher deviances are indicative of greater model misfit (Singer & Willett, 2003). Although lower deviances indicate better **model fit**, we cannot interpret deviance in isolation, and it is a function of sample size as well as model fit. However, we can interpret differences in deviance for competing models as long as the models (a) are hierarchically nested, (b) use same observations and (c) use FIML to estimate the parameters (if we wish to compare two models that differ in terms of their fixed effects).

Likelihood ratio (deviance difference) test

When one model is a subset of the other, the two models are said to be hierarchically nested (e.g. Kline, 2015), such that 'the more complex model includes all of the parameters of the simpler model plus one or more additional parameters' (Raudenbush et al., 2004, pp. 80–81). In sufficiently large samples, under standard normal theory assumptions and using the same set of observations, the difference between the deviances of two hierarchically nested models follows an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters being estimated between the two models (de Leeuw, 2004; Raudenbush & Bryk, 2002; Singer & Willett, 2003). Using the likelihood ratio test (LRT), we can compare two hierarchically nested models. The simpler model (the model with fewer parameters) is the null model (M_0); the more parameterised model is the alternative model (M_1). The deviance of the simpler model (D_0) has p_0 parameters; the deviance of the more parameterised model (D_1) has p_1 parameters. The simpler model must have fewer parameters ($p_0 < p_1$), and the deviance of the simpler model must be at

least as large as the deviance of the more parameterised model ($D_0 \geq D_1$). We compare the difference in deviance ($\Delta D = D_0 - D_1$) to the critical value of chi-square with degrees of freedom equal to the difference in the number of estimated parameters ($\Delta p = p_1 - p_0$). Using the LRT, we prefer the more **parsimonious model**, as long as it does not result in (statistically significantly) worse fit. Put another way, if the model with the larger number of parameters fails to reduce the deviance by a substantial amount, we retain the simpler model (M_0). However, when the change in deviance (ΔD) exceeds the critical value of chi-square with $p_2 - p_1$ *df* (degrees of freedom), then the additional parameters result in statistically significantly improved model fit. In this scenario, we favour the more-complex model (i.e. Model M_1, with p_1 *df*).

Having described the LRT, we must now attend to a few subtle but important details about using the LRT to compare two nested models within multilevel modelling.

First, comparing two nested models that differ in their fixed effects (γ) requires FIML, not REML. In FIML, the number of reported parameters includes the fixed effects (the γ terms) as well as the variance–covariance components. In REML, the number of reported parameters includes only the variance and covariance components. REML allows for comparison of models that differ in terms of their random effects, but both models must have the same fixed effects structure. Therefore, comparisons of models with differing fixed and random effects should utilise the deviance provided by FIML (Goldstein, 2011; McCoach & Black, 2008; McCoach et al., 2018; Snijders & Bosker, 2012). The major advantage of using FIML over REML is the ability to compare the deviances of models that differ in terms of their fixed and/or random effects. Most statistical programs use REML as the default method of estimation, so remember to select FIML estimation to use the deviance estimates to compare two nested models with differing fixed effects (McCoach & Black, 2008).

Second, when comparing the fit of two models that differ in terms of their variance components, we sometimes encounter a boundary issue that affects the way in which we must conduct such model comparisons. Variances cannot be negative. Therefore, if the variance of the random effect is 0 in the population, then the estimation of this variance component hits a lower boundary (variance = 0). (Similar issues can arise when testing correlation coefficients, which are bounded by ±1.00; however, covariance values are generally not bounded is this way.) Given the lower bound of 0, the sampling distribution for a variance with a population value of 0 is not normally distributed. Instead, it has a median and mode of 0 and is leptokurtic and positively skewed. Therefore, for multilevel models with random effects, the standard LRT is too conservative (Self & Liang, 1987; Stram & Lee, 1994).

To adjust for this issue, if the model has only one random effect, and therefore only one τ, we can use the chi-square value for $p = .10$ to test for statistical significance when we set the Type I error rate (alpha) at .05 (Snijders & Bosker, 2012). The critical

value of chi-square with 1 *df* is 3.841 for $p < .05$ and 2.706 for $p < .10$. To compare two models that differ in terms of one variance component, we should use the critical value of 2.706. In contrast, to compare two models that differ in terms of one fixed effect (and no random effects), the critical value is 3.841.

Comparing a model with one random effect to a model with two random effects is a bit more complex. The simpler model has two fewer parameters: it eliminates both a variance and a covariance. Although the variance has boundary issues (the variance cannot be less than 0), the covariance does not. Therefore, the correct critical value of χ^2 comes from the $\bar{\chi}^2$ distribution, which is actually a mixture of χ^2 distributions (Snijders & Bosker, 2012). Technically, the correct critical value of χ^2 for a model that eliminates one random slope variance (τ_{11}) and one covariance (τ_{01}) is 5.14, rather than 5.99, as would normally be the case for a model that differs by two parameters (Snijders & Bosker, 2012, p. 99). The rejection regions for LRT that include variance components are 2.706 for a single variance parameter, 5.14 for a variance and a covariance, 7.05 for a variance and two covariances and 8.76 for a variance and three covariances (Snijders & Bosker, 2012). Snijders and Bosker (2012) present a more detailed discussion of this issue, as well as a table with the correct critical values to compare nested models that differ in terms of one or more randomly varying slopes.

Akaike information criterion and Bayesian information criterion

Information criteria such as *Akaike information criterion* (AIC) and the *Bayesian information criterion* (BIC) also provide a method to compare the fit of competing model. There is an advantage to using *information criteria* for model comparison: they allow for comparison of non-nested models. Using AIC and BIC, we can compare competing models fit using the same sample, whether or not they are hierarchically nested. Lower information criteria (ICs) are indicative of better fitting models; therefore, the model with the lowest IC is considered the best fitting model (McCoach & Black, 2008). For additional details regarding the conceptual and methodological underpinnings of the AIC and the BIC, see Bozdogan (1987), Burnham and Anderson (2004), Raftery (1995), Schwarz (1978), Wagenmakers and Farrell (2004), Weaklim (2004, 2016) and Zucchini (2000).

The Akaike information criterion (AIC)

To compute the AIC, simply multiply the number of parameters by 2 and add this product to the deviance statistic. The formula for the AIC is

$$AIC = D + 2p \tag{2.11}$$

where D is the deviance and p = the number of estimated parameters in the model. The model with the lowest AIC value is considered the best model.

The deviance (or –2 log likelihood [–2LL]) represents the degree 'of inaccuracy, badness of fit, or bias when the maximum likelihood estimators of the parameters of a model are used' (Bozdogan, 1987, p. 356). The second term, $2p$, imposes a penalty based on the complexity of the model. This penalty indicates that the deviance must decrease by more than two points per additional parameter to favour the more parameterised model.

Compare this to the LRT for model selection. The critical value of χ^2 with 1 *df* at $\alpha = .05$ is 3.84 (or 2.706 for a 1 *df* change involving a variance). Therefore, when comparing two models that differ by 1 *df*, the LRT imposes a more stringent criterion for rejecting the simpler model. In fact, for comparisons of models that differ by seven or fewer parameters,[i] using the LRT results in an equivalent or more parsimonious model than the AIC. Conversely, when comparing models that differ by more than seven parameters, the AIC favours more parsimonious models than the LRT.

The Bayesian information criterion (BIC)

The BIC equals the sum of the deviance and the product of the natural log of the sample size and the number of parameters. The formula for the BIC is

$$BIC = D + \ln(n) * p \tag{2.12}$$

where D is deviance (–2LL), p is the number of parameters estimated in the model and n is the sample size. As with the AIC, the model with the lowest BIC is considered the best fitting model.

Therefore, the penalty the BIC imposes for each additional parameter is a function of the sample size (n). However, in MLM, it is not entirely clear which sample size should be used with the BIC: the total number of observations, the number of clusters at the highest level or some weighted average of the two. Furthermore, different software packages compute the BIC differently. Some (e.g. SPSS) use the overall sample size, whereas others (e.g. SAS PROC MIXED) use the number of clusters (level-2 units). Therefore, even when different statistical packages produce identical –2LL and AIC values, the BIC value may differ. Hence, the choice of sample size to compute the BIC could potentially change the outcome(s) of the model selection process.

Regardless of the choice of sample size for BIC, the per parameter penalty for the BIC is higher than the per parameter penalty for the AIC. Generally, multilevel models

[i]The number 7 assumes that we are using the standard critical values for chi-square with $\alpha = .05$, not critical values that have been adjusted for boundary issues in the variances.

have at least 10 clusters, and for a sample size of 10, the penalty for the BIC is 2.3 times the number of parameters. (In fact, the sample size must be less than eight for the per parameter penalty for the BIC to drop below 2.) In contrast, the penalty for the AIC is 2 times the number of parameters. Therefore, whenever the AIC favours the simpler (less parameterised model), the BIC also favours the simpler (less parameterised) model. Whenever the BIC favours the more complex (more parameterised model), the AIC also favours the more parameterised model.

Not all software programs provide AIC and BIC measures in their output. However, it is easy to compute the AIC and the BIC from the deviance statistic. FIML is the most appropriate estimation method to use when computing information criteria (Verbeke & Molenberghs, 2000) compare two models that differ in terms of their fixed effects.

Using model fit criteria for model selection

Unfortunately, the AIC, BIC and LRT may differ in terms of which model they favour. Honestly, the differences in the model fit criteria can be a bit overwhelming, and the various criteria do not always favour the same model. Table 2.1 displays the total penalty imposed by each of the model fit criteria for models that differ by 1, 2, 3 and 4 parameters. For example, imagine that we want to compare two models that differ by one fixed effect. Our total sample size is 1000 people, nested within 50 clusters. The deviance of the model that includes the parameter is 3.9 points lower than the deviance of the model that did not. In such a scenario, the LRT and AIC would favour the model that includes the fixed effect parameter; both the BIC_2 and the BIC_1 would favour the model that eliminates the fixed effect parameter. In this situation, two different researchers, faced with the same results, might make different decisions about which model to favour.

Table 2.1 The total penalty imposed by each of the model fit criteria for models that differ by 1, 2, 3 and 4 parameters

	1 Parameter	2 Parameters	3 Parameters	4 Parameters
AIC	2	4	6	8
BIC (n = 10)	2.3	4.6	6.9	9.2
LRT (var)	2.7	5.14	7.05	8.76
LRT (trad)	3.84	5.99	7.82	9.49
BIC (n = 50)	3.91	7.82	11.73	15.64
BIC (n = 100)	4.61	9.22	13.83	18.44
BIC (n = 1000)	6.91	13.82	20.73	27.64
BIC (10,000)	9.21	18.42	27.63	36.84

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; LRT = likelihood ratio test.

We suggest some simple heuristics to help navigate the morass of model fit criteria. These are meant to provide guidance in using the various model fit criteria; they are not meant to supersede them. Because AIC and BIC do not require nested models, we can apply this strategy to both nested and non-nested models. Table 2.2 displays the *per-parameter* penalty imposed by each of the model fit criteria for models that differ by 1, 2, 3 and 4 parameters. Across all of the model fit criteria, the per-parameter penalty is always at least 2. Therefore, when comparing two models that differ by one or more parameters (in terms of the number of parameters estimated by the model), first compute the difference in the deviances of the two models. Then divide that number by the difference in the number of estimated parameters (Δ/p). If the ratio of the deviance difference to the number of parameters (Δ/p) is less than 2, the fit criteria favour the more parsimonious (less parameterised) model. This is the simplest scenario, given that all criteria always favour the more parsimonious model when Δ/p is less than 2. When this ratio is above 10, we recommend favouring the more parameterised model. Again, this is a staightforward decision, as all criteria suggest favouring the more complex model (except perhaps the BIC_1, but the total sample size needs to be almost 25,000 people for the BIC_1 to favour the simpler model, and even then, all other criteria favour the more parameterised model). In small- to moderate-sized samples, we tend to favour the more parameterised model when the ratio is above 4 because the AIC and LRT always favour the more parameterised model when the ratio is above 4. In very large sample sizes, this heuristic may not be appropriate, given that deviance is a function of sample size. Therefore, in very large samples, using the BIC_1 may be advisable.

Table 2.2 The per-parameter penalty imposed by each of the model fit criteria for models that differ by 1, 2, 3 and 4 parameters

	1 Parameter	2 Parameters	3 Parameters	4 Parameters
AIC	2	2	2	2
BIC (n = 10)	2.3	2.3	2.3	2.3
LRT (var)	2.7	2.57	2.35	2.19
LRT (trad)	3.84	3	2.61	2.37
BIC (n = 50)	3.91	3.91	3.91	3.91
BIC (n = 100)	4.61	4.61	4.61	4.61
BIC (n = 1000)	6.91	6.91	6.91	6.91
BIC (n = 10,000)	9.21	9.21	9.21	9.21

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; LRT = likelihood ratio test.

Δ/p ratios between 2 and 4 represent the 'grey zone', where some model fit criteria favour the simpler model and other model fit criteria favour the more parameterised model.

When Δ/p is between 2 and 4, compute the various model fit criteria. Then supplement the model fit criteria with Rights and Sterba's (2019) variance-explained measures (discussed in the next section). What proportion of within, between and/or total variance does this parameter explain? Does removing the parameter substantially reduce the predictive ability of the model? Also, given the purpose of the model, decide which would be a more grievous error: an error or omission or the inclusion of an unnecessary parameter. If omitting a potentially important parameter is more problematic, then we recommend favouring the more parameterised model. If including an unnecessary parameter is more problematic, then we recommend favouring the simpler model. Generally speaking, in the grey zone, we favour retaining potentially important *fixed effects* but eliminating unnecessary *random effects* (variance components).

If the difference in the number of parameters is 0 (which can happen when comparing two non-nested models), then we favour the model with the lower deviance. Why? We cannot use the LRT for non-nested models. If two models have the same number of parameters, then the model with the lower deviance always has the lower AIC, BIC_1 and BIC_2 (and in fact has the lower IC, regardless of which is IC chosen).

Model selection decisions should consider both the fit and the predictive ability of the multilevel model. Next, we turn our attention to quantifying explained variance in multilevel models.

Proportion of variance explained

In single-level regression models, an important determinant of the utility of a model is the proportion of variance explained by the model, R^2. In MLM, computation of the proportion of variance explained becomes far more complex. Variance components exist 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 varies as a function of the level-2 unit or cluster. Given that variance in the outcome variable is decomposed into multiple components, quantifying the variance explained by a set of predictors becomes more complicated than in the single-level case.

Conceptually, we could be interested in measuring variance explained within clusters, variance explained between clusters and/or total variance explained (both within and between clusters). For example, adding a cluster-level (level-2) variable to a multilevel model cannot possibly explain within-cluster variance. Similarly, cluster mean centred level-1 (within-cluster) variables cannot explain between-cluster variance. However, a cluster-level variable can explain between-cluster variance; and because it can explain between-cluster variance, it can also explain some of the total variance. Imagine a situation in which 5% of the variance in the outcome variable lies between clusters and 95% of the variance lies within clusters. Suppose we find a

variable that explains most (80%) of the between-cluster variance. This variable is a powerful predictor of the between-cluster variance, but it only explains 4% (80% * 5%) of the total variance. Therefore, deciding how to compute and report variance-explained measures in multilevel modelling requires explicit consideration of the context and goals of the research.

To this end, Rights and Sterba (2019) developed 'an integrative framework of R^2 measures for multilevel models with random intercepts and/or slopes based on a completely full decomposition of variance' (p. 309). To use Rights and Sterba's variance-explained measures to partition outcome variance into between-cluster outcome variance and within-cluster outcome variance, we group mean centre all level-1 predictor variables and add the aggregate level-1 variables into the level-2 model. (There is one exception: if the level-1 variable has only within-cluster variance, and has no between-cluster variance, then this is not necessary.) Then we can decompose the model-implied total outcome variance into five specific sources of variance: (1) variance attributable to level-1 predictors via fixed slopes (f_1), (2) variance attributable to level-2 predictors via fixed slopes (f_2), (3) variance attributable to level-1 predictors via random slope variation and covariation (v), (4) variance attributable to cluster-specific outcome means via random intercept variation (m)[ii] and (5) variance attributable to level-1 residuals (σ^2) (Rights & Sterba, 2019).

Assuming all level-1 variables have been group mean centred, three of the sources contain only within-cluster variance: (1) variance attributable to level-1 predictors via fixed slopes (f_1), (2) variance attributable to level-1 predictors via random slope variation and covariation (v) and (3) variance attributable to level-1 residuals (σ^2). Therefore, we can evaluate the proportion of within-cluster variance explained by level-1 predictors (f_1), and we can determine what proportion of within-cluster variance is accounted for by the variances and covariances of the randomly varying slopes (v) (Rights & Sterba, 2019).

Two of these sources contain only between-cluster variance: (1) variance attributable to level-2 predictors via fixed slopes (f_2) and (2) variance attributable to cluster-specific outcome means via random intercept variation (m). Therefore, we can determine the proportion of between-cluster variance that is explained by our level-2 predictors (f_2) and the proportion of between-cluster variance that is random intercept variance not explained by the level-2 predictors in our model (m) (Rights & Sterba, 2019).

In every model, it is possible to decompose the model-implied total variance into these five sources. Then, using Rights and Sterba's (2019) **integrative framework of R^2** measures in multilevel models, researchers can compute a variety of variance-explained

[ii]When all level-1 variables are cluster mean centred (and the aggregate is included at level-2), $m = \tau_{00}$.

measures, each of which provides potential insights into the model's predictive capabilities. Rights and Sterba (2019) show the correspondence between their integrative framework and other R^2 measures that have been used in MLM. Their integrative framework allows for easy computation of previously used variance-explained measures without needing to estimate multiple multilevel models. In addition, Shaw et al. (2020) developed an R package, *r2mlm*, that computes all measures in Rights and Sterba's integrative framework and provides graphical representations of the various measures. In Chapter 3, we return to this topic in greater detail, when we describe the process of fitting and evaluating multilevel models. There, we provide more details on Rights and Sterba's R^2 measures and provide concrete recommendations for using Rights and Sterba's integrative framework within the model building process.

Effect size

An *effect size* is a practical, interpretable, quantitative measure of the magnitude of an effect. As with any statistical analyses, it is important to report effect size measures for multilevel models. The R^2 measures described above can help researchers and readers to determine the impact that a variable or a set of variables has on a model, with respect to variance explained. In addition, researchers can compute Cohen's *d*-type effect sizes to describe the mean differences among groups. To calculate the equivalent of Cohen's *d* for a group-randomised study (where the treatment variable occurs at level 2), use the following formula (Spybrook et al., 2011):

$$\delta = \frac{\hat{\gamma}_{01}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}_{00}}} \tag{2.13}$$

Assuming two groups have been coded as 0/1 or –.5/+.5, the numerator of the formula represents the difference between the treatment and control groups. The denominator utilises the σ^2 and τ_{00} from the unconditional model, where the total variance in the dependent variable is divided into two components: (1) the between-cluster variance, τ_{00}, and (2) the within-cluster variance, σ^2. There are numerous ways to compute effect sizes in MLM (or any analysis), and not all effect sizes need to be standardised, especially when unstandardised metrics are commonly used and easily understood. We encourage you to present the results of your MLM as clearly as possible and to contextualise the parameters in practically meaningful and easily interpretable ways.

Now that we have introduced most of the fundamental concepts in MLM, let's turn our attention to an applied example so that we can provide concrete guidance on how to build and interpret multilevel models. Chapter 3 focuses on building, evaluating and interpreting multilevel models.

Chapter Summary

- In a multilevel model without predictors, each person's score on the dependent variable is composed of three elements: (1) the expected mean (γ_{00}), (2) the deviation of the cluster mean from the overall mean (u_{0j}) and (3) the deviation of the person's score from his/her cluster mean (r_{ij}). In this equation, γ_{00} is a *fixed effect*: γ_{00} is the same for everyone. The u_{0j} term is called a *random effect* for the intercept because u_{0j} randomly varies across the level-2 units (clusters).
- In multilevel models, *fixed effects* are parameters that are fixed to the same value across all clusters (or individuals), whereas *random effects* differ (vary) across clusters (or individuals).
- The variance in the outcome variable can be partitioned into within and between-cluster variance components. The ability to partition variance into within-cluster variance and between-cluster variance is one of MLM's greatest assets.
- Intercepts and slopes can vary across clusters in a multilevel model. We can build a regression equation at level 2 to try to explain the variation in these randomly varying intercepts and slopes.
- Model selection decisions should consider both the fit and the predictive ability of the multilevel model.
- Variance components exist at each level of the multilevel model. In addition, in random coefficients models, the relation between an independent variable and the dependent variable at level 1 varies as a function of the level-2 unit or cluster.
- Rights and Sterba (2019) developed 'an integrative framework of R^2 measures for MLM with random intercepts and/or slopes based on a completely full decomposition of the outcome variance'.

Further Reading

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

This is a classic book on the theory and use of hierarchical linear modelling and multilevel modelling. The book is a must read for researchers interested in diving further into the mathematical details of hierarchical linear models (e.g. estimation theory and multivariate growth models).

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138.

This article provides a detailed overview of grand mean centring and group mean centring in the context of two-level multilevel models. In addition to the expansive discussion of centring in multilevel models, it provides illustrative examples that should provide readers with a foundation to answering questions with their data.