25

Group Administered Interventions and Cluster Designs

Randomization is too important to be left to chance

- J. D. PETUCELLI

INTRODUCTION

Cluster randomized trials randomly assign clusters of individuals to treatment versus control conditions rather than randomly assigning individuals per se to these conditions. For example, I might randomly assign classes within a school to treatment or control conditions, with the effect that every student within a class is assigned to one of the two conditions. In this case, each class is viewed as a "cluster" and is subject to the random assignment process. As another example, I might administer a therapy to each of 50 small groups of individuals, with each group consisting of 5 members. My control group also might have 50 groups of 5 individuals each, with the members of the control group engaging in a group activity unrelated to the intervention topic. This is a cluster randomized design where the different groups are conceptualized as "clusters" and the clusters are randomized to one of the two the treatment conditions. Finally, a researcher might randomly assign 30 clinics to either a treatment or control condition and then randomly sample 100 clients from each clinic to participate in the study. This also is a cluster randomized trial with clinics as clusters.

When we analyze data at the individual level, the standard independence assumption

is that the disturbance scores for the outcome variable are uncorrelated. However, in clustered designs such independence cannot be assumed. If a group of individuals or a class of students contains a particularly disruptive individual, then the outcome scores for all members of that cluster are likely to be affected — inducing a dependency among their observations. However, members of other clusters will not be directly affected since the other groups are not exposed to the disruptive member. Alternatively, a group might have a particularly good group therapist/teacher/leader, and all of the members of that group would benefit accordingly. In these cases, the presence of such clustering creates dependency among the disturbances of the outcome variables, and if the dependencies are strong enough then adjustments need to be made in statistical estimation and testing to accommodate them. Applying conventional structural equation modeling (SEM) to individual-level analyses of clustered data but ignoring such clustering can result in downwardly biased standard error estimates. The net result could be rejecting null hypotheses inappropriately, which could include rejecting valid models based on global chi-square tests as well as rejecting null hypotheses associated with causal effects that are estimated within a model.

In this chapter, I consider issues in the analysis of cluster randomized RETs. I first distinguish clustering in the context of respondent sampling versus experimental designs. I then consider two orientations towards clustering: either as a nuisance or as something that is theoretically and/or substantively interesting. I next consider the hierarchical structure of clusters in order to introduce two-level and three-level clustering. I then tie these concepts to the specification of multilevel equations. I introduce intraclass correlation coefficients and design effects, followed by a numerical example that I use throughout the chapter to illustrate analytic methods for RETs with clusters. I consider pseudo-maximum likelihood, multilevel SEM, Bayesian methods, and specialized tools for the analysis of RETs with a small number of clusters. I conclude with a discussion of methodological issues that can arise with clustered designs.

The literature on cluster randomized trials is vast and I can't cover here all the ins and outs of designing and analyzing such studies. For summaries, see Turner, Prague, Gallis, Li and Murray (2017) and Turner et al., (2017). The current chapter is long and not easily processed in a single sitting. You can read it in parts over time.

SAMPLING/EXPERIMENTAL DESIGN AND CLUSTERING

You will encounter statements about the importance of cluster adjustments when analyzing data, but decisions to adjust for clustering and how to do so can be complex. One of the most common reasons given for making such adjustments is that individuals within the

same cluster might be subject to a shared, cluster-focused "random shock" (e.g., a disruptive member) that then creates score dependencies among people in the same cluster. A core dilemma is how researchers justify making cluster distinctions on some population partitions but not others. If a researcher conducts a study in a school and applies traditional statistical methods to evaluate a model, then they implicitly assume the model disturbances are independent. But surely it is possible that the behavior of one student in the school can affect the behavior of another student in that class, creating disturbance dependencies.

To provide additional context, consider that there are often naturally occurring cliques of students within a school (e.g., athletes) that are ignored in single-school studies, raising the possibility of ignored dependencies. In a multi-school study, failing to adjust for the existence of such cliques may be more consequential than failing to adjust for membership in a given school per se. Yet, we often cluster-analyze data using different schools as clusters while ignoring potentially consequential clusters within schools. The reality is that dependencies are likely a fact of life in most data modeling. The key question is not whether dependencies exist but rather whether the operative dependencies are sufficiently large and consequential to lead us astray in the inferences we make. I discuss below ways of gaining perspectives on this question.

Abadie et al. (2017) conceptualize decisions to adjust for clustering as a matter of sampling design, experimental design, or both. Clustering is a matter of **sampling design** when sampling formally follows a two-stage process in which, at the first stage, a subset of clusters is randomly sampled from a population of clusters (e.g., middle schools from all possible middle schools in the United States) and, in the second stage, individuals are randomly sampled from these randomly selected clusters (e.g., a random sample of students is taken from each of the selected schools). Clustering is a matter of **experimental design** when clusters of lower-level units rather than units per se (e.g., classes as opposed to individual students) are randomly assigned to treatment versus control conditions, but analyses are pursued at the lowest level of analysis. Abadie et al. (2017) argue that cluster adjustments usually are needed when clusters are a formal part of the sampling design or the experimental design. Other sources of disturbance dependencies are seen as resulting from more traditional modeling matters, such as omitted variable bias or specification error, that should be addressed by more standard modeling methods for doing so. Cluster randomized trials typically (but do not always) involve clustering vis-a-vis experimental design, so such cases are my primary focus in the current chapter.

CLUSTERS AS A NUISANCE OR AS THEORETICALLY MEANINGFUL

In some randomized trials, clustering is not of substantive interest per se. Rather, the

clusters are viewed as nuisances whose impact must be taken into account when calculating p values and confidence intervals for making statistical inferences. In such cases, one analyzes data across individuals with appropriate clustering corrections, potentially including differences in cluster means that could otherwise confound results. In other scenarios, the clusters are of substantive interest in their own right; one seeks to build models that incorporate characteristics of the clusters into the research questions. For example, in a randomized trial that assigns classes of students to treatment versus control conditions, a researcher might want to know if class size affects treatment versus control mean differences on the study outcome. In this case, the researcher tests moderation of intervention effectiveness as a function of a cluster level characteristic, class size (i.e., a class size X intervention interaction). As will become apparent, one's approach to data can differ depending on whether clusters are seen as nuisances or theoretically meaningful.

HIERARCHICAL STRUCTURES OF CLUSTERS

In survey research that uses clustered designs, the clusters often are called **primary sampling unit**s (PSUs) because they are the primary target of random selection. Individuals within each cluster are called **elements** or **secondary sampling units** (SSUs). In the literature on multilevel modeling, clusters and cluster-level measures often are referred to as **Level-2 data** whereas data collected on individuals within clusters are referred to as **Level-1 data**.

There are cluster designs where clustering has complex hierarchical structures. I might randomly sample census tracts from the United States (one level of clustering) and within each census tract, randomly sample schools (a second level of clustering); then, I randomly sample, say, 100 students from each selected school for purposes of conducting an RET on student achievement but where I randomly assign schools as opposed to students per se to either a treatment or control condition. In this design, the individuals represent Level 1-data that are nested within schools (i.e., each student occurs in a different school), the schools represent Level-2 data that are nested within census tracts, and the census tracts and data we collect on them represent **Level-3 data**. Cluster randomized trials can incorporate such hierarchical structures but it is not possible for me to address the many variants of such three or four level cluster designs here. My focus in this chapter will be on two level cluster randomized trials.

MULTILEVEL MODELS

In this section, I describe how researchers often represent multilevel data in equation form. I then use the equations in later sections to frame the analysis of cluster randomized trials.

To develop core concepts, I use a hypothetical study of 50 high schools in which the outcome variable of interest is youth physical activity. Adolescence is often associated with declines in physical activity. School-based interventions have been developed to slow the decline. In general, it is recommended that high school aged youth engage in about 60 minutes of moderate-to-vigorous physical activity (MVPA) per day. The trial randomly selected 50 schools from a broader population of schools and randomly assigned 25 of these schools to an intervention condition and 25 schools to a treatment as usual (TAU) control group. The outcome variable was the number of minutes of MVPA per day engaged in by a student as measured for two weeks at the end of the school year using an accelerometer for 20 randomly selected students from each school. The clusters are schools and the elements within the cluster are the random sample of students within each school. Indices of school characteristics reflect Level-2 data and measures of individual student physical activity and other such variables represent Level-1 data. Keep in mind that the population of schools I am working with is much larger than the 50 sampled schools and that the total number of students in each sampled school is more than the 20 students sampled from each school

When introducing population parameters that researchers often reference when invoking multilevel models, I will make several simplifying assumptions for the sake of pedagogy. I relax many of them later. Also, when estimating population parameters from sample data, we need to invoke statistical theories that allow us to take into account different forms of sampling error. I elaborate these statistical theories below.

I begin by representing mathematically selected facets of the population from which the 50 schools and 20 students per school are randomly selected.. In theory, I can express the within-cluster (Level-1) MVPA minutes per day in the population data as an intercept-only equation, as follows:

$$Y_{ij} = \alpha_j + \varepsilon_{ij} \qquad [25.1]$$

where Y is the number of minutes of MVPA per day for individual *i* in school *j*, α_j is the intercept for school *j* and ε_{ij} is the disturbance or error score for individual *i* in school *j*. When there are no predictors in a linear equation, as is the present case, it turns out the intercept will equal the mean of the outcome for a given school, so in this case, α_1 is simply the average number of minutes of MVPA for all students in School 1, α_2 is the average number of minutes of MVPA for all students in School 2, α_3 is the average number of minutes of MVPA for all students in School 3, and so on.[1] Suppose the average number of minutes of MVPA for students in School 1 is 42 minutes per day. If I encounter a student

[1] For now, I assume no measurement error. I relax this assumption later in the Chapter

from this school, I would predict that his or her physical activity is 42 MVPA minutes per day, which is the value of the school mean as reflected by the intercept in Equation 25.1.

Suppose that the student, who I will refer to as Student 1, engages in 40 minutes of MVPA per day. My prediction would be off by 2 minutes and this is reflected in the error or disturbance term ε_{ij} in Equation 25.1. The value of ε_{11} is 42 - 40 = 2. In theory and at the population level, I can calculate such error scores for every student in School 1. I might then calculate the standard deviation of the error scores across all the students in School 1 and find that it equals 23.0. This reflects the amount of variability in MVPA minutes per day *within* School 1. I also can calculate the standard deviation of the within school errors across all schools/clusters to obtain an index of the within school variation in MVPA minutes across schools. I symbolize this parameter as σ_ε. Note that there is no subscript *j* for this expression because the index is calculated across all schools. Suppose σ_ε equals 25.0. This means that MVPA scores typically vary within the schools by about 25 minutes per day. The population level **within-cluster variance** is the square of this value, σ_ε^2, which is $25^2 = 625$. The σ_ε sand σ_ε^2 are important indices because they provide us with a sense of how much within-cluster or, in this case, within-school variability exists. They reflect how different the students are from each other after eliminating the differences between the schools. Researchers who use multilevel models often seek to estimate the value of this parameter from the data they collect.

Shifting to a Level-2 perspective, I might ask what is the average Level-1 intercept, α_j, across the *j* schools/clusters. I can express this as a Level-2 linear equation as follows (using γ_0 as the intercept on the right side of the equation to distinguish it from Level-1 coefficients):

$$\alpha_j = \gamma_0 + u_j \qquad [25.2]$$

In this equation, α_j is the within-school intercept for school *j* and γ_0 , also an intercept but calculated across schools, is conceptualized as the statistical expectation (or mean) of the α_j across clusters/schools. Suppose γ_0 equals 47.0. This means that the average minutes of MVPA per day across all clusters is 47.0. If someone asks me what is the average amount of time students engage in MVPA per day in School 1, my answer *based on Equation 25.2* would be 47 minutes. Recall, however, that the actual value of α_j for School 1 was 42.0 minutes. My answer based on Equation 25.2 is in error by 47.0 – 42.0 = 5.0 minutes and this error is captured in the u_j term in Equation 25.2 (technically, it is $\alpha_j - \gamma_0$). As with the Level-1 data, I can calculate the standard deviation of the *u* errors across all clusters, which I symbolize as σ_u. This statistic gives me a sense of the across-school variability in the mean MVPA per day across the schools. Suppose it equals 20. This means that the school means differ from the grand mean γ_0, on average, by 20 minutes. The **between-**

school/cluster variance is then $\sigma_u^2 = 20^2 = 400$. Researchers who use multilevel models also often estimate the value of this parameter from the data they collect.

Note that in a cluster randomized trial, the between-school variance includes the effects of the intervention in it. I can take this into account in the population Level-2 equation by adding a dummy variable for the treatment condition (0 = control, 1 = intervention) to Equation 25.2:

$$\alpha_j = \gamma_0 + \gamma_1 T_j + u_j \quad [25.3]$$

γ_1 is a regression/path coefficient and equals, like any dummy variable, the difference in the mean α_j for the intervention schools minus the corresponding mean for the control schools. The γ_1 coefficient is of primary interest to program evaluators. Note that in Equation 25.3, the standard deviation σ_u now reflects the variability in α_j holding constant (or removing the effects of) the treatment versus the control group.

The above exposition allows me to identify key concepts in the analysis of cluster randomized trials. In the absence of any Level-1 and Level-2 predictors, σ_ε^2 is the **within-cluster variance** of the outcome in the RET and σ_u^2 is the **between-cluster variance** of the outcome. It can be shown mathematically that the total variation in the outcome, Y, is an additive function of these two variances:

$$\sigma_Y^2 = \sigma_\varepsilon^2 + \sigma_u^2 \quad [25.4]$$

When a Level-2 predictor is included in the model, such as a dummy variable for the treatment condition or any other Level-2 covariate, σ_u^2 is the between-cluster variance holding constant or after removing the variance explained by the added Level-2 predictors. When a Level-1 predictor is included in the model, σ_ε^2 is the within-cluster variance holding constant or after removing the explained variance by those Level-1 predictors. I make use of these concepts below.

THE INTRACLASS CORRELATION COEFFICIENT

A statistical index often used in studies with clustering is the **intraclass correlation coefficient** (ICC). It often is used to provide perspectives on the need to adjust for clustering, to provide insights into dependency structures, and to justify pursuit of multilevel modeling. There are many types of ICCs (some of which go by the name **interclass correlation** rather than intraclass correlations) which can lead to confusion about the properties of ICCs. My discussion here focuses on intraclass correlation coefficients in forms that are most relevant to the analysis of cluster randomized trials.

One instantiation of the ICC uses it to document between-cluster mean differences

on the outcome relative to the total variability in the outcome; that is, it is the ratio of between-cluster variance for the outcome relative to the total variance of the outcome, or, when there are no Level-1 or Level-2 predictors:

$$ICC = \sigma_u^2 / \sigma_Y^2 = \sigma_u^2 /(\sigma_u^2 + \sigma_\varepsilon^2) \quad [25.5]$$

The ICC is often thought of as the proportion of variance in Y due to across cluster dynamics. One minus the ICC is the proportion of variance in Y due to within-cluster dynamics. Under this conceptualization, if the ICC is 1.0, this means that each score within a cluster is identical – that variation in Y is completely due to across cluster dynamics and there is no within-cluster variability in scores. As the ICC becomes increasingly less than 1.0, the within-cluster scores become less similar. When the ICC equals zero, then within-cluster variability dominates the total variability in Y; between-cluster variability is zero. A large ICC often raises red flags to statistically adjust for error dependencies because across cluster dynamics are at play.

In RCTs, usually half of the clusters are assigned to the intervention group and half to the control group, so a low ICC also is indicative of a modest or trivial intervention effect. To gain additional perspectives on error dependencies independent of this effect, researchers often evaluate the ICC for the intervention and control groups separately or after the effects of the treatment condition have been removed from Y, per Equation 25.3.

A common point of confusion in the literature is the reference to the ICC as a correlation coefficient when Equation 25.5 portrays it as a proportion of explained variance, which traditionally is viewed as a squared correlation. The ICC *is* a correlation coefficient but it is a special type. I elaborate this point here given common misunderstandings of ICCs. Uninterested readers can skip the next three paragraphs.

Consider a hypothetical example that consists of eight clusters with two members/elements in each cluster. The third column of Table 25.1 presents the Y scores (which have a 0 to 10 metric) for each individual in each cluster:

Table 25.1: Clustered Data Example

Cluster	Individual	Y
1	1	5
1	2	6
2	1	3
2	2	2
3	1	7
3	2	9

4	1	2
4	2	2
5	1	3
5	2	5
6	1	6
6	2	9
7	1	4
7	2	2
8	1	8
8	2	7

To calculate estimates of the variance components σ_u^2 and σ_ε^2, I conducted a one way random effects analysis of variance where the independent variable or factor is the cluster (I did this using SPSS but you can use other statistical packages). The factor has 8 levels because there are 8 clusters. I used a random effects analysis as opposed to the traditional fixed effects analysis (see Chapter 5) because the 8 clusters are assumed to represent a random sample of clusters from a broader population of clusters, a conceptualization I comment more on below. I symbolize the sample estimate of σ_u^2 as $\hat{s}_u^2$ and, from the analysis of variance, I find that it equals 4.50. I symbolize the sample estimate of σ_ε^2 as $\hat{s}_e^2$ and the analysis of variance yields a value for it of 1.50. From Equation 25.5, I find

$$ICC = \hat{s}_u^2 / (\hat{s}_u^2 + \hat{s}_e^2) = 4.50 / (4.50 + 1.50) = 0.75$$

The estimated intraclass correlation is 0.75, which is large. Variation in Y is dominated by between-cluster variance, which is consistent with inspection of the data in Table 25.1.

Adopting a correlational perspective, I restructure the data in Table 25.1 to represent each pair of scores in a cluster in a single row, with one member of the cluster in one column and the second member of the cluster in a second column, like this:

Cluster	Y.1	Y.2
1	5	6
2	3	2
3	7	9
4	2	2
5	3	5
6	6	9
7	4	2
8	8	7

If I calculate the traditional Pearson correlation between the two columns to document

score similarity within clusters, I find the correlation is 0.82, which does not equal the previously computed ICC. This occurs because there is a problem with using Pearson's correlation in this way. In the above data table, the cluster member who I designate as member 1 and who I designate as member 2 is arbitrary; For example, I could just as easily flip the row scores within cluster 1 to be 6 and 5 instead of 5 and 6 given that who I designate as Y.1 and who I designate as Y.2 is arbitrary. If I calculate the traditional Pearson correlation after enacting such a switch for just row 1, I obtain a result different from 0.82. This is because Pearson's correlation presumes that scores are meaningfully identified within their respective columns. For example, If I correlate height and weight with height in the first column and weight in the second column, then the person's height must go in the first column and weight in the second column. I can't arbitrarily interchange numbers between columns for any given row. Studies of twins that correlate scores for one twin member with scores for the other twin member also encounter this arbitrary pairing issue; which twin do you designate as member 1 and which twin as member 2? Sir Ronald Fisher proposed a solution to the arbitrary pair problem using what he called the intraclass correlation: Have every pair be included twice, in both orders, and then compute the Pearson correlation. Here is the reshaped data matrix:

Cluster	Y.1	Y.2
1	5	6
1	6	5
2	3	2
2	2	3
3	7	9
3	9	7
4	2	2
4	2	2
5	3	5
5	5	3
6	6	9
6	9	6
7	4	2
7	2	4
8	8	7
8	7	8

If I calculate the correlation between the two columns, I obtain 0.75, which equals the value of the variance based ICC. The ICC is indeed a correlation coefficient but it is a correlation for the case of "unpaired" data in the same cluster. If there were more than two individuals

per cluster, the logic is the same but there are more ways of creating pairs of individuals within clusters. The ICC is the correlation between all possible pairings but unordered.

I do not want to get sidetracked on details, but, in theory, the ICC ranges from -1 to +1 and can take on negative values. Some methodologists view negative ICCs in cluster randomized trials as artifacts of statistical bias corrections, arguing that such ICCs should be treated as if they are very small and positive or zero. Other methodologists argue that negative ICCs can be meaningful if they reflect "shocks" to clusters that encourage heterogeneity rather than uniformity on the outcome. One example is the 'fixed pie' case where there is a fixed amount of a resource within a group, such as speaking time during a meeting, in which case more time for one person means less time for another (Islam & Zyphur, 2005). The issue is moot in most cluster randomized trials because the ICC is rarely negative in them. Also, keep in mind that there are special variants of the ICC for the case where the variables involved are binary or ordinal (see Ridout, Demétrio & Firth, 1999; Eldridge, Ukoumunne & Carlin, 2009).

DESIGN EFFECTS

The intraclass correlation by and of itself gives us some sense of whether dependencies are present but it does not inform us how much the dependencies might disrupt statistical inference. A concept that reflects the latter is known as a design effect. The **design effect** is the ratio of the magnitude of the standard error one would observe for a dependency-corrected analysis compared to the standard error one would observe under simple random sampling. If the design effect is 2.0, then this means the adjusted standard error is twice as large as the standard error one obtains under random sampling of independent replicates. The larger the design effect, the greater the need to adjust for bias due to dependencies, everything else being equal. Estimates of the magnitude of the design effect depend on the parameter of interest and the method used to adjust for dependencies. Generally, design effects larger than 2.0 are deemed worthy of statistical remediation but this standard is arbitrary and context dependent. The reason some analysts are reluctant to correct for dependencies is because it typically reduces the power of statistical tests and the precision of estimates. The argument in favor of corrections is that ignoring adjustments can be misleading because unadjusted data underestimate the role of sampling error.

When quantifying design effects, some methodologists use the standard errors to form the relevant ratio whereas other researchers use the square of the standard errors (which are the standard errors expressed as variances). It is important to know which index is used because they imply different design effect magnitudes given the same standard errors. For example, suppose the unadjusted standard error is 2.0 and the adjusted standard

error is 4.0. The design effect index using the standard errors directly is 4.0/2.0 = 2.0. If I use instead the variances, the design effect index is $4.0^2/2.0^2 = 4.0$. The design effect as applied to parameter variances (i.e., squared standard errors) is often symbolized by DEFF whereas the design effect as applied to the standard errors directly is symbolized by DEFT, although there is some inconsistency in such uses in the literature.

Design effects can vary in magnitude within the same study depending on the variables analyzed, how those variables are distributed, and the type of analysis being applied to the variables. Design effects can be less than 1.0, in which case, they increase statistical precision and power, 1.0, in which case they have no effect on statistical precision and power, or they can be greater than 1.0, in which case, they decrease statistical precision and power; see Park & Lee, 2004; Vierron & Giraudeau, 2009.

If the ICC is zero, the design effect typically will equal 1.0. Also, in many (but not all) scenarios, smaller cluster sizes lead to smaller design effects. This is because the dependencies will then be limited to a small number of individuals in each group/cluster. If the dependencies extend to a large number of individuals within a cluster, then this is potentially more problematic. Later, I show how the DEFT is used by some to adjust for clustering effects.

CLUSTER POPULATIONS

The analytic methods I discuss below make different assumptions about the cluster population in one's study. Classic two step random sampling of clusters occurs when we first delineate the population of clusters (e.g., clinics or schools) that we seek to make inferences about and then we secure a formal random sample of clusters from that population. In Chapter 4, I introduced the concept of a meta-population that turns this process on its head: The researcher identifies a set of, say, clinics or schools that will participate in his or her study as a matter of convenience but then construes these clinics as a random sample from a broader meta-population of clinics. In both cases, we deal with a random sample of clusters from a broader population of clusters but the process of sampling is different. In the latter approach, the researcher's job is to make a convincing case about who the meta-population is, i.e., the population of clusters that the sample of studied clusters can be construed as a random sample from. This approach justifies the use of inferential statistics on sample level data even with convenience samples. However, ambiguities occur because we are not always certain who the relevant target population is.

When specifying populations, there is an important distinction to be made. Suppose I conduct a study in which my clusters are different clinics. I might have in my final sample some clinics that serve 300 clients per year, some that serve 400 clients per year, some that

serve 500 clients per year and some that serve 600 clients per year. One way of construing the broader population of clusters that the clinics represent is as those clinics having these particular clinic sizes. When such one-to-one correspondence exists between-cluster values on a variable in the sample and the cluster values on that variable in the population, statisticians refer to the variable values as being **fixed** in character

A different way of construing the broader population of clusters on any given variable is that the population consists of clusters with the particular values of the variable in my sample, but also other values as well. In my clinic size example, the particular clinic sizes that occur in my sample are construed as representing a random sample of values from the broader population of clinics I am studying. Stated another way, I seek to generalize not just to the fixed clinic sizes in my sampled clinics, but to the broader population of clinics that contain more varied clinic sizes than those that happen to show up in my sampled data. Statisticians often refer to such variable values as being **random** in character when they are associated with a random distribution and a random sampling scheme. It turns out that the way one approaches these two scenarios analytically, i.e., the presence of fixed versus random predictors, can differ because in the random predictor case, one must take into account the additional random error introduced by not having sampled all the predictor values in the population. I will draw on these distinctions later in this chapter.

Gelman (2005) notes that the terms ‘random’ and ‘fixed’ have been used in many different ways in statistics to refer to different concepts in different ways, indeed sometimes even in contradictory ways. Gelman prefers to avoid the terms altogether and I will tend to follow his lead in this chapter. However, know that these terms are used in different ways in the broader literature on cluster randomized trials and it can be confusing.

My key point is that many cluster randomized trials sample clusters (e.g., clinics, schools) and study participants in ways that reference meta-populations rather than existing populations. Researchers need to think about if the values of the variables they are studying have one-to-one correspondence to those in the broader meta-population that is the focus of generalization or if the values represent random samples from variable values in the meta-population. Researchers also need to address issues of just who the meta-populations are that the sampled clusters and participants in a study are thought to represent—this latter issue is common to all research involving statistical inference, but arguably becomes more important in clustered designs because two types of populations are being considered.

WORKED EXAMPLES

I use two numerical examples of cluster randomized explanatory trials to illustrate key points for analysis. The first example is the previously discussed school-based intervention

to increase the amount of moderate-to-vigorous physical activity (MVPA) per day on the part of high school youth. The trial randomly assigned 25 schools to the intervention condition and 25 schools to a treatment as usual (TAU) control group. The outcome variable was the number of minutes of MVPA per day as measured for two weeks at the end of the school year using an accelerometer for 20 randomly selected students from each school. The clusters are schools and the elements within the cluster are students within each school. Indices of school characteristics reflect Level-2 data and measures of individual student physical activity represent Level-1 data. The intervention sought to affect two mediators, (1) educating youth about the benefits or advantages of MVPA, and (2) teaching youth how to engage their friends in helping them receive peer support for engaging in MVPA. Each mediator was measured on a multi-item self-report scale at posttest that ranged from -5 to +5 with 0 as a neutral point; more negative scores reflect low levels of perceived advantages and peer support for MVPA and more positive scores reflect high levels of perceived advantages and peer support for MVPA. The scales were multi-item composites based on the individual's average response to disagreement or agreement with statements about the respective construct (-5 = strongly disagree, -3 = moderately disagree, -1 = slightly disagree, 0 = neither agree nor disagree, 1 = slightly agree, 3 = moderately agree, 5 = strongly agree). In order to keep the example simple for purposes of pedagogy, I do not include covariates but these would normally be included in such an experiment. Their inclusion is straightforward, although one must be concerned about possible bias if one uses a lagged outcome as a covariate.

The goal of the intervention was to increase MVPA by a minimum of 20 minutes per week (or an average of about 3 minutes per day). Any effect size below this magnitude was deemed too weak to be meaningful. For the mediators, consultation with experts suggested a mean difference of half a scale unit, 0.50, as a reasonable meaningfulness standard for the intervention minus control effect on the mediator. Finally, a 3 unit change in the MVPA measure for every one unit increase in the mediator was judged by the experts to be a reasonable meaningfulness standard for the effects of each mediator on MVPA.

The second example is an RET in which individuals are randomly assigned to an intervention or control condition, with individuals in the intervention then randomly assigned to small groups of 10 members each for a group administered interactive intervention on wearing masks during the COVID pandemic. The control group, instead, receives an interactive intervention on eating nutritious foods, also in the same small group format. The primary outcome variable is the intention to regularly wear a mask in the future, measured on a multi-item -5 to +5 scale with higher scores indicating a more positive intention. The intervention targeted as mediators the individuals' attitudes toward wearing masks (i.e., the perceived advantages of wearing a mask and the perceived

disadvantages of not wearing one) and perceived norms surrounding mask use. Measures of mask wearing attitudes and perceived norms to regularly wear a mask in the future were obtained after the intervention/control activities were completed. Each construct was assessed on a multi-item scale whose scores on individual items ranged from -5 (strongly disagree) to +5 (strongly agree), per the earlier described metric. The total score for all measures was the average item response across the items, with appropriate reverse scoring. The higher the score, the more positive the intention to wear a mask regularly, the more positive the attitude towards doing so, and the more supportive the perceived norm. There were 50 small groups in the intervention condition and 50 in the control condition.

The goal of this intervention was to increase intentions to regularly wear masks by a scale unit of 0.20. Any effect below this magnitude was deemed too weak to be meaningful. For the mediators, consultation with experts suggested a mean difference of one third a scale unit, 0.33, was a reasonable meaningfulness standard for the intervention minus control conditions on the mediator. Finally, a 0.33 unit change in mask wearing intentions for every one unit increase in the mediator was judged by the experts as a reasonable meaningfulness standard for the effect of each mediator on mask wearing intent.

Figure 25.1 presents the model structure for both studies. The treatment condition is scored 1 = intervention group, 0 = control group. The model assumes no direct effect of the treatment condition on the outcome over and above the two mediators. The correlated disturbances between the mediators reflect the fact that there likely are other sources of the correlation between them than just the common cause of the treatment condition.

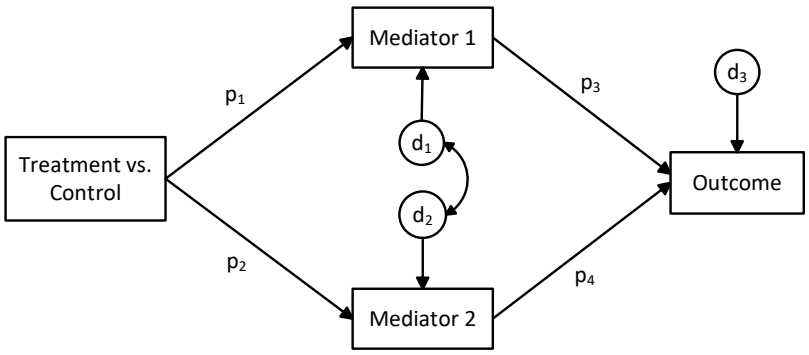


FIGURE 25.1. Model structure for two RETs

For both studies, there are three endogenous variables in the model, yielding three core causal equations, expressed here using sample notation:

$M1 = a_1 + p_1$ Treatment $+ d_1$

$M2 = a_2 + p_2$ Treatment $+ d_2$

$Y = a_3 + p_3\ M1 + p_4\ M2 + d_3$

where M1 is mediator 1, M2 is mediator 2, and Y is the outcome.

CLUSTERING AS A NUISANCE

Suppose clustering creates error or disturbance dependencies that are non-trivial in magnitude. It may be the case that the clusters per se are not of substantive interest to you but that they instead act as nuisance factors whose effect on dependency structures must be dealt with to make valid inferences. In this case, you will need to introduce cluster-based corrections for standard error estimation but beyond that, the analysis is pretty much the same as traditional SEM regression modeling.

Mplus uses a multivariate form of pseudo-maximum likelihood (PML) estimation to make adjustments to the standard errors in such cases. The mathematics of the approach are described in Asparouhov (2005) and Asparouhov and Muthén (2005). PML uses Taylor series linearization to adjust standard errors (Oberski, 2014). Forms of bootstrapping also are available (Asparouhov & Muthén, 2010e). When applying PML, you obtain the same regression/path coefficients as traditional maximum likelihood analysis but the estimated standard errors typically differ. The unadjusted standard errors usually will be underestimates (i.e., too small). The Mplus approach is tied to asymptotic theory and requires sufficiently large sample sizes to satisfy the required asymptotics. It also assumes there is no non-trivial omitted variable bias and that the clusters are not part of a higher level hierarchical structure that introduces non-trivial dependencies at the cluster level. The Mplus output is similar in format to traditional classic single level models without cluster adjustments. However, the adjustments have indeed been implemented on the output. Outcome variables can be continuous, censored, binary, ordinal, nominal, or counts and all results are interpreted much as in traditional SEM.

PML requires that the number of clusters be relatively large. About 50 clusters is generally considered to be sufficient (Angrist & Pischke, 2008; Donner & Klar, 2000; Kahan et al., 2016; Leyrat, Morgan, Leurent & Kahan, 2018) but some methodologists argue that a smaller number can be used in certain contexts. For example, the authors of the Mplus program suggest that 20 clusters might be feasible in some contexts (Muthén, 2014; see also Huang, 2016, 2018). I discuss in Chapter 28 how you can conduct simulations to determine if the number of clusters you have or plan to have in your study

is problematic.

The PML approach works best when clusters tend to be of roughly equal size (Nichols & Schafer, 2007). It handles missing data seamlessly via maximum likelihood methods per Chapter 26. It also readily lends itself to mediation and moderation analyses.

Mplus Analysis with Nuisance Cluster Adjustments

For the physical activity example, Table 25.2 presents the relevant Mplus syntax for treating clustering as a nuisance. Most of the syntax should be familiar to you from prior chapters. I number the lines for reference, but the numbers do not appear in the code when you program in Mplus.

Table 25.2: Mplus Nuisance Based Analysis

```
1.  TITLE: Cluster as nuisance ;
2.  DATA:
3.  FILE IS mvpa.dat ;
4.  VARIABLE:
5.  NAMES ARE
6.    mvpa peers advant treat school  ;
7.  USEVARIABLES ARE
8.    mvpa peers advant treat ;
8a. MISSING ALL (-9999) ;
9.  CLUSTER is school ;   !identify cluster variable
10. ANALYSIS:
11. TYPE = COMPLEX ;    ! specify complex design option
12. !BOOT = 5000 ;
13. MODEL :
14. mvpa ON advant peers  ;
15. peers on treat ;
16. advant ON treat ;
17. advant with peers ;
18. MODEL INDIRECT:
19. mvpa IND treat ;
20. OUTPUT: SAMP STANDARDIZED RESIDUAL MOD(ALL 4)
21. CINTERVAL TECH4 ;
22. !CINTERVAL(BOOTSTRAP) TECH4 ;
```

Lines 5 and 6 input the data, including the cluster id variable called `school` which is characterized by an integer between 1 and 50 (inclusive) that signifies which cluster the individual is in. There are 50 such clusters, each with 20 members. Lines 7 and 8 tell Mplus which variables to use in the causal model, leaving out the cluster variable because it is not a formal part of the defining equations. Line 9 identifies the cluster id variable and Line 11

uses the COMPLEX option to inform Mplus that this is a clustered design. The remaining syntax is straightforward based on material I have covered in prior chapters. I explain the commented out lines (Lines 12 and 22) later. I do not specify the estimation method because Mplus invokes its robust algorithm by default in models that use COMPLEX.

Because I have described Mplus output formats in priori chapters, I present the output here with limited exposition. The results for model global fit are as follows:

```
Chi-Square Test of Model Fit

          Value                              0.392*
          Degrees of Freedom                     1
          P-Value                           0.5315
          Scaling Correction Factor         7.6667
            for MLR

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.000
          90 Percent C.I.                0.000  0.071
          Probability RMSEA <= .05           0.844

CFI/TLI

          CFI                                1.000
          TLI                                1.000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.011
```

All of the indices suggest satisfactory model fit. The localized fit indices (modification indices, standardized residuals) also point to satisfactory fit:

```
           Standardized Residuals (z-scores) for Covariances
              MVPA          PEERS         ADVANT        TREAT
              ________      ________      ________      ________
 MVPA           0.001
 PEERS          0.000         0.000
 ADVANT         0.002         0.000         0.000
 TREAT         -0.553         0.000         0.000         0.000
```

```
MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index     4.000

                                  M.I. E.P.C. Std E.P.C. StdYX E.P.C.

No modification indices above the minimum value.
```

To address the first question of whether the intervention affects the outcome, Mplus provides an analysis of the total effect of the intervention on MVPA in the section labeled TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from TREAT to MVPA				
Total	7.348	2.310	3.182	0.001

There was a statistically significant effect of the intervention on MVPA (the difference between treatment and control groups = 7.35 minutes, margin of error (MOE) = ±4.62, critical ratio (CR) = 3.18, $p < 0.05$). The lower limit of the 95% confidence interval was 2.73 minutes of MVPA per day. This result is just below the meaningfulness standard of 20 minutes per week, or 2.85 minutes per day. Thus, I am 95% confident that the program effect is non-zero, but technically, I cannot assert that the program yields a meaningful effect with the same degree of confidence. To be sure, my best guess of the program effect is the observed sample mean difference, which is to increase MVPA by 7.35 minutes per day. But if I acknowledge that sampling error is present in this estimate, it is possible that the error is sufficiently large that my confidence is not strong (i.e., 95% strong) that the effect is indeed meaningful.

The second core question of interest in an RET asks whether the intervention has a meaningful effect on each of the mediators, which in the present example is defined as a mean difference of 0.50 scale units for each mediator. Here are the results for the estimated effects of the intervention on the mediators:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PEERS ON				
TREAT	0.776	0.319	2.436	0.015
ADVANT ON				
TREAT	0.733	0.306	2.395	0.017

The estimated effect of the intervention on perceived advantages was to increase perceived advantages, on average, by 0.73 units on its -5 to +5 metric, which was statistically significant (MOE = ±0.62, CR = 2.40, $p < 0.05$). The lower limit of the 95% confidence interval was 0.16 which overlaps the meaningfulness standard. Thus, I am again in a situation where I can conclude the intervention effect on the mediator is non-zero but U can't be confident it is meaningful. My best guess of the intervention's effect on the perceived advantages mediator is to raise perceived advantages by 0.73 units, but after taking into account sampling error, I can't say with confidence that the effect is meaningful.

The estimated effect of the intervention on the peer support mediator on its -5 to +5 metric also was statistically significant (coefficient = 0.78 ± 0.64, CR = 2.44, $p < 0.05$), but I also cannot conclude with confidence that the effect is meaningful after taking into account sampling error.

In terms of the standardized effect size of the intervention on the mediators, you can convert results on the Mplus output to any of the effect size indices discussed in Chapter 10 (e.g., Cohen's *d*, probability of exceptions to the rule), something I leave to you as an exercise. As an example, I can calculate Cohen's *d* for the perceived advantages mediator as the path coefficient 0.733 divided by the square root of the unstandardized residual variance for perceived advantages. The value of the residual variance from the Mplus output was 2.05 (not shown above). This yields $d = 0.512$.

The third question in an RET focuses on estimating the effects of each mediator on the outcome. Here are the results from the Mplus output for these estimated effects:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
MVPA ON				
ADVANT	4.696	0.577	8.142	0.000
PEERS	5.035	0.609	8.264	0.000

The estimated path coefficient for the effect of perceived advantages on MVPA was 4.70 ±1.13, CR = 8.14, $p < 0.05$; for every one unit that perceived advantages increases on its -5 to +5 metric, the mean MVPA is predicted to increase 4.70 minutes holding constant peer support. The lower limit of the confidence interval for this coefficient is 3.56 which exceeds the a priori meaningfulness standard for it of 3.0. The perceived advantage mediator has meaningful impact on MVPA.

The estimated path coefficient for the effect of peer support on MVPA was 5.04 ±1.22, CR = 8.26, $p < 0.05$; for every one unit that peer support increases on its -5 to +5 metric, the mean MVPA is predicted to increase 5.04 minutes, holding perceived attitudes constant. The lower bound confidence interval for it was 3.82, which also exceeds its

meaningfulness standard. Both mediators appear to be relevant determinants of MVPA.

Here is the estimated squared multiple correlation for MVPA as predicted from perceived advantages and peer support:

```
R-SQUARE

    Observed                                        Two-Tailed
    Variable        Estimate       S.E.    Est./S.E.    P-Value

    MVPA               0.432      0.052      8.252      0.000
```

Perceived advantages and peer support account for 43% of the variance in MVPA.

Mplus also provides an analysis of the omnibus indirect effect of each mediator on MVPA based on Line 19 of Table 25.2. Here is the relevant output:

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

                                                    Two-Tailed
                    Estimate       S.E.    Est./S.E.    P-Value

Effects from TREAT to MVPA

  Specific indirect 1
    MVPA
    PEERS
    TREAT              3.908      1.652      2.366      0.018

  Specific indirect 2
    MVPA
    ADVANT
    TREAT              3.440      1.500      2.294      0.022
```

The overall mediational effect through peer support was statistically significant (coefficient = 3.91 ±3.24, CR = 2.37, $p < 0.05$) and this also was true for perceived advantages (coefficient = 3.44 ±3.00, CR = 2.29, $p < 0.05$). These results are consistent with what one would conclude using the joint significance test. I generally prefer the joint significance test coupled with a link-by-link analysis of the strength of the effect for each link in the respective mediational chain as opposed to the above omnibus tests, per my discussions in Chapters 10 and 17. However, some researchers prefer the omnibus test.

As noted in previous chapters, the omnibus indirect effect tests often can be improved by using bootstrapping instead of relying on the MLR estimator of Mplus. I can use bootstrapping by commenting out Line 21 in Table 25.2 and uncommenting Lines 12 and 22. In this particular instance, the results of the two analyses were comparable, so I do not

show the bootstrapping output.

An advantage of the above "cluster as nuisance" strategy is that all of the traditional methods for mediation and moderation directly apply to it (McNeish, Stapleton & Silverman, 2017) as do the interpretation of standardized coefficients. A disadvantage of the approach is that one cannot address certain kinds of substantive questions that may be of interest at the level of clusters which I elaborate below when I present multilevel structural equation modeling (MSEM). The "cluster as nuisance" strategy conflates sources of between-cluster and within-cluster variability and if unconflated analyses are theoretically important to you, the approach is suboptimal. Having said that, I show you later in the chapter how to bring Level-2 predictors into the above modeling strategy to tease out between-cluster and within-cluster effects. Although this helps, there are still things you can do in MSEM that you cannot do with the "nuisance" strategy and I will elaborate those later.

Generalized Estimation Equations

An alternative approach to dealing with clustering that treats clustering as a nuisance is a method known as **generalized estimating equations** (GEE). GEE uses a different statistical theory than PML. It makes use of an a priori, user-specified working correlation structure for observations within a cluster. This presumed correlation structure then informs the estimation of the regression coefficients and their standard errors in the specified regression model (for details, see Liang & Zeger, 1986; Zeger & Liang, 1986; Zeger, Liang, & Albert, 1988). GEE does not handle missing data as well as PLM. Specifically, GEE assumes missing data are MCAR whereas PLM assumes missing data are MAR (see Chapter 26). This is because GEE is not a likelihood-based method (Ghisletta & Spini, 2004). Although this property of GEE is disconcerting, Fitzmaurice, Laird, and Rotnitzky (1993) found in simulations that the bias of GEE when applied to data that are MAR tends to be small unless the amount of missing data is large (near 50%). Weighted GEE methods have been developed to make GEE more flexible for handling missing data but these require further evaluation (Chen, Yi, & Cook, 2010; Lipsitz, Ibrahim, & Zhao, 1999; Robins, Rotnitzky, & Zhao, 1995).

GEE is popular in epidemiology and public health. It is not as flexible as traditional multilevel frameworks but it can offer useful perspectives on panel data and clustered regression. I discuss GEE panel analysis in Chapter 16 but return to it later in this chapter after I have developed selected foundational concepts. For a non-technical tutorial on using GEE for cluster adjustments, see Huang (2022). The document on the *Resources* tab of my webpage for Chapter 25 also discusses and applies GEE to clustered binary outcomes. My website has a program for conducting it, called *GEE cluster regression.*

Cluster Level Dummy Variables

A third approach to adjusting for clustered data is to treat the clusters as a nominal variable and then to introduce dummy variables for them vis-a-vis traditional OLS regression analysis. This strategy removes or controls for all between-cluster variation in the outcome, which is tantamount, more or less, to adjusting for error dependencies as a function of the clusters. This approach is problematic for clustered randomized trials, however, because the intervention versus control group manipulation occurs at the cluster level. Given this, you are unable to compare the two treatment conditions because you have removed those differences when you include the cluster dummy variables. A contrast strategy to circumvent this limitation has been proposed by McNeish and Stapleton (2016) but their approach needs greater exploration and statistical justification. As well, statisticians have shown that in certain contexts, the dummy variable approach is not sufficient to guarantee the removal of within-cluster error dependencies (see, for example, Abadie et al., 2017).

Little et al. (2022; see also Niolon et al., 2019) suggest a two stage multiple group SEM strategy for cluster randomized trials that makes use of cluster-based dummy variables. One of the groups is the intervention condition and the other group is the control condition. In the first stage, model variables are regressed within each group onto the set of cluster defined dummy variables and each of the per subject unstandardized residuals are saved for use in stage two. In the second stage, the unstandardized residuals are used as indicators of variables for standard SEM multiple group analyses that test (a) within group causal relationships between mediators and with the outcome, (b) across group differences in the within group causal coefficients, and (c) between group differences in means of the residual indicators. Little et. al's approach has the advantage of relying on well-known multiple group SEM analytic structures, but it has not been subjected to formal simulation evaluations in the context clustered randomized trials. At this point, it is a provocative but understudied analytic strategy that needs further exploration.

In sum, if clusters are nothing more than a nuisance and you have a sufficient number of clusters then the PML approach offered by Mplus often is a reasonable analytic strategy for analyzing mediation or moderation in a clustered RET.

MULTILEVEL SEM

Multilevel SEM is a second major strategy for analyzing clustered RETs but where you seek to make statements about cluster dynamics per se, i.e., the clusters are more than just a nuisance. For example, you might want to formally test if the effects of a group-administered intervention on an outcome are moderated by the experience levels of the group facilitators/therapists leading the groups, with more experienced facilitators bringing

about more group-level change than less experienced facilitators. MSEM allows you to test for such group-level effects in ways that the nuisance approach cannot. MSEM is distinct from traditional multilevel modeling that uses software like HLM and MLwin as well as a method commonly known as mixed effects modeling. To be sure, these methods can be shown to be special cases of MSEM but MSEM generally is superior to them because it can incorporate latent variables into the analysis and it can deal with complex structural relationships between variables in ways that are not possible in the more traditional methods.

When applied to sample data, the significance tests for the coefficients in an MSEM model require reasonable estimates of σ_ε and σ_u described earlier. Estimating σ_ε usually is straightforward but estimating σ_u requires that a sufficient number of Level-2 clusters be included in the study. If there are, for example, 8 clusters in the study, then we effectively estimate σ_u based on only 8 data points, which can be problematic. In general, a minimum of about 40 to 50 clusters seem necessary to produce reasonable results for estimating σ_u for many multilevel methods (Angrist & Pischke, 2009; Carter, Schnepel & Steiger, 2012; McNeish & Stapleton, 2016b) but sometimes even more are needed. Disparate cluster sample sizes also can create estimation difficulties. For example, MacKinnon and Webb (2013) found that with heterogeneous cluster sizes, as many as 100 clusters were needed in some contexts to yield valid estimates of σ_ε and σ_u. I revisit these issues below and discuss how to deal with scenarios where the number of clusters in your study is small.

MSEM can be applied using traditional maximum likelihood or it can use Bayesian estimation. Because of its complexity, MSEM based on maximum likelihood often encounters convergence and estimation challenges, especially as the number of clusters, the sample size within clusters, and the intraclass correlations all become smaller (Depaoli & Clifton, 2015; Li & Beretvas, 2013; Ludtke et al., 2011; Meuleman & Billiet, 2009). Bayesian estimation is a better alternative in many of these scenarios (Asparouhov & Muthen, 2012; Depaoli & Clifton, 2015; Hox et al., 2014). Mplus has made significant advances in Bayesian estimation for MSEM, so I will rely on Bayes modeling here.

Level-1 and Level-2 Variables Revisited

As noted, in multilevel designs Level-1 variables focus on characteristics of individuals within clusters whereas Level-2 variables focus on characteristics of the clusters to which individuals belong. In the MVPA example, peer support and perceived advantages are Level-1 variables. The treatment condition that a school/cluster is assigned to is a Level-2 variable. If I want to evaluate if the intervention for improving MVPA is more effective in private versus public schools, then I would need to obtain a measure of whether each school is public or private and this measure then would be used as a Level-2 variable in the context

of multilevel moderated regression. In MSEM as applied to clustered RETs, we typically are interested in theorizing about between-cluster variables or causes of variation across clusters. In the nuisance variable approach, we instead usually are more interested in variation across individuals ignoring clusters or taking into account only the treatment status, intervention or control, of the cluster.

Global versus Contextual Level-2 Variables

Researchers often make distinctions between two types of Level-2 variables. **Contextual variables** are obtained via the aggregation of Level-1 data. In the MVPA example, I might calculate the average posttest peer support in each cluster and then treat that average as a Level-2 variable that predicts or has implications for other Level-2 variables. This Level-2 variable has the same value for each member of a cluster because it is the cluster mean but the values can differ across clusters. A common Level-2 contextual variable used in school-based interventions is the average SES of students in the school. We often want to know, for example, if an intervention is more effective in schools that tend to serve lower income students as opposed to schools that tend to serve higher income students. The average SES of students in a school represents an indicator of school student income levels. The second type of Level-2 variable is a **global or integral variable**. Such variables are measured directly at Level 2 and cannot be broken down or seen as an aggregate of Level-1 scores on that variable. Being in a public versus private school in the MVPA study is such a variable. It also is constant for everyone in the school/cluster. The distinction between global and contextual Level-2 variables is important and I return to it later.

When an aggregate of a within-cluster variable for study participants is used to represent a between-cluster contextual variable, there is an important qualification to keep in mind. Suppose in the MVPA example I obtain a measure of SES for each student in the study and then use the aggregate (average) SES for participants attending the same school as an index of the school-level SES. The problem with this strategy is that the sample average SES for a given school is only an estimate of the population mean SES for *all* students at that school. We know there will be sampling error associated with this estimate and ideally our statistical analyses takes this sampling error into account. Otherwise, coefficient bias can result. The bias is referred to as **Lüdtke's bias**; see Lüdtke et al. (2008) and Asparouhov and Muthén (2019). One feature that sets Mplus-based MSEM apart from other multi-level modeling strategies is that it adjusts for Lüdtke's bias.

The amount of Lüdtke bias that is consequential depends, among other things, on (a) the model being evaluated (b) the magnitude of the ICC, (c) the sampling ratio of the number of individuals in a given cluster relative to the size of the population of that cluster from which the individuals are sampled, and (d) the absolute number of individuals in a

cluster. Generally speaking, if one ignores sampling error for cluster averages when sampling ratios are greater than 0.20 and the absolute sample size within a cluster is reasonably large (say, n > 30 or more), then Lüdtke bias often will not be problematic (Lüdtke et al., 2008; Kush, Konold & Bradshaw, 2022; Marsh et al., 2012; Preacher, Zhang & Zyphur, 2011). As one falls below these rough guidelines, then Lüdtke bias can be problematic and conclusions must be more tentative. Some methodologists suggest a 50-50 rule that argues for at least 50 clusters with 50 individuals per cluster, but, in my opinion, this is too crude and overly discouraging. In the final analysis, one probably should use the localized simulation strategies discussed in Chapter 28 to gain perspectives on the viability of a given research design. Parenthetically, contextual variables that ignore sampling error or deem it not applicable are referred to as **manifest between-level variables**.

Variable Decomposition

It can be shown that the total variability of any variable in a multilevel model is an additive function of its within-cluster variability and its between-cluster variability. For MVPA, for example,

$$var_{TOTAL}(MVPA) = var_{BC}(MVPA) + var_{WC}(MVPA)$$

where var_{BC} is an index of the between-cluster variance of a variable and var_{WC} is an index of the within-cluster variance of that variable, expressed using sample notation. Stated another way, the total variability in MVPA across all data points is a function of how the average MVPA for a school/cluster varies across schools/clusters and also how much MVPA varies within each school/cluster. The same is true for the peer support and perceived advantages variables:

$$var_{TOTAL}(Advant) = var_{BC}(Advant) + var_{WC}(Advant)$$

$$var_{TOTAL}(Peers) = var_{BC}(Peers) + var_{WC}(Peers)$$

Note that for a global variable like whether a school is public or private, there is no within-cluster variability; its variability is completely determined by var_{BC}. But for variables like MVPA, perceived advantages of MVPA, and peer support, both var_{BC} and var_{WC} contribute to the overall variability of those variables. Between-cluster variability for any given variable is assumed to be uncorrelated with its within-cluster variability; knowing the variability of the mean MVPA across schools/clusters does not allow us to say how much variability in MVPA there is within specific schools/clusters.

Focus for the moment on an outcome and a predictor that have both within-cluster and between-cluster variability, say, MVPA and the predictor perceived advantages of

MVPA. Within clusters, I can estimate the path coefficient for the impact of perceived advantages on MVPA for each school/cluster separately by regressing MVPA onto perceived advantages. I symbolize this path by p_{WC} because it is a within-cluster statistic. There are multiple p_{WC}s for a given predictor, one for each cluster. We often are interested in the average value of the p_{WC} for a given predictor across clusters so as to document the "typical" within-cluster effect for the predictor, in this case, of perceived advantages on MVPA. We also might be interested in how much the p_{WC} coefficient varies across clusters.

Between clusters, I can estimate a corresponding path coefficient but now I calculate it not using individual scores within each cluster but rather using estimates of the Level-2 cluster means for the respective variables. In the MVPA example, my sample size for the calculation of such paths is 50 because I have 50 clusters and the estimated cluster mean values on MVPA and the cluster mean values on perceived advantages are used to calculate the path coefficient, regressing one set of means onto the other set of means.[2] I call these path coefficients p_{BC} and there is only one such coefficient for each predictor. It also is of theoretical interest because it estimates causal effects at the between-cluster level. An attractive feature of MSEM is that it allows you to construct a causal model for your variables at the between-cluster level using cluster level indices and a separate causal model for your variables at the within-cluster level. Usually the models will be the same but this will not always be nor need it be the case. Each model can include mediation and moderation dynamics (or not) which is not true of traditional multilevel models.

Traditional multilevel software reports a single coefficient for a predictor that is a weighted average of p_{WC} and p_{BC} i.e., it conflates the values of these distinct parameters. The mathematics of this conflation are described in Preacher et al. (2010). A useful feature of MSEM is that it unconflates the coefficients and provides unambiguous estimates of the within-cluster predictor effect on an outcome and the between-cluster effect of that predictor on the outcome. For a global Level-2 variable such as whether the school is public or private, MSEM only calculates p_{BC} because there is no within group variability for it.

Varying Slopes versus Non-Varying Slopes

Another feature of both multilevel modeling and MSEM is that they allow us to evaluate if slopes (and intercepts) vary meaningfully across clusters. In the MVPA study, I can, in theory, calculate the path coefficient for each school when I regress MVPA onto perceived advantages and peer support within schools (I refer to these two coefficients as $p1_{WC}$ and $p2_{WC}$, with the p representing a path coefficient and WC indicating it is "within cluster"). Note that for purposes of the present discussion, I use sample notation, for reasons that will be apparent shortly. I might record the values of the coefficients for each cluster, like this:

[2] The process is more complex but this conveys the spirit of what is being done.

Cluster	$p1_{WC}$	$p2_{WC}$
1	3.84	4.84
2	5.14	8.93
3	1.32	10.82
.	.	.
.	.	.
49	6.22	3.20
50	6.02	6.38

If you scan the column for the path coefficient for perceived advantages ($p1_{WC}$), you can see there is variability in the values of $p1_{WC}$ across the different clusters. Some of this variability is due to sampling error and the question becomes whether the observed variability reflects the case where the true population values of $p1_{WC}$ are, in fact, equal in every cluster and the variability you see is just sampling error. Or, alternatively, do the population within-cluster coefficients truly vary across the clusters? If the true population coefficients for the predictor are all equal, then the coefficient is said to be **non-varying**. If the true population coefficients differ across clusters, then the coefficient is said to be **varying**. In MSEM you can evaluate these properties. Later, I show you how to do so.[3]

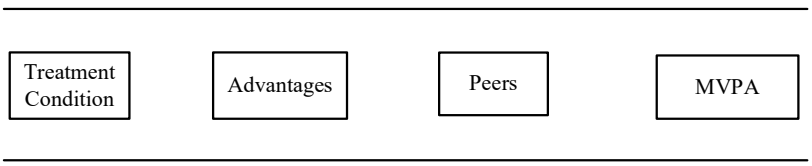
If you conclude the within-cluster coefficients for a given predictor truly differ across clusters, then the question becomes why is this the case? You might formulate hypotheses about the sources of such variability and then measure the hypothesized causes of coefficient variability, testing the relationship between these presumed causes and coefficient variability. Such tests are called **cross-level moderation** in the multilevel modeling literature because the Level-2 between-cluster variable moderates the impact of the within-cluster predictor on the within-cluster outcome.

Influence Diagrams for Multilevel SEM

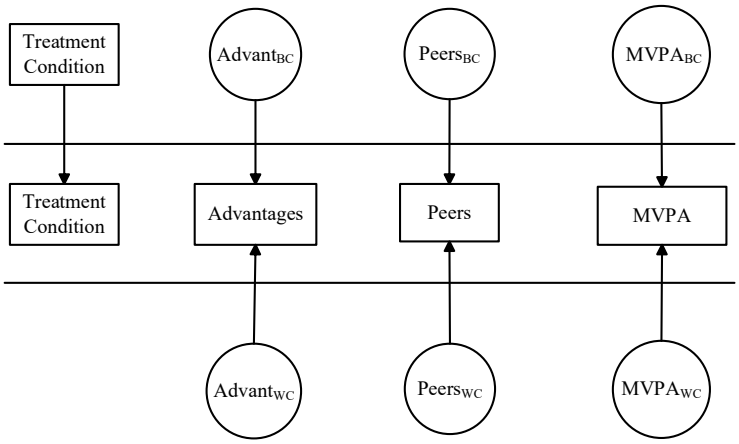
Like SEM, MSEM makes use of influence diagrams but the diagrams are more complicated than traditional SEM diagrams. There is no one set of accepted diagramming conventions and you will encounter different representations by different authors. In this section, I outline one set of graphing conventions for MSEM using the MVPA example.

One begins by drawing boxes of all the observed variables in your model and placing them from left to right between parallel horizontal lines, like this:

[3] In the MSEM literature, a non-varying coefficient is often referred to as being **fixed** and varying coefficients are said to be **random**, yet another use of these terms.



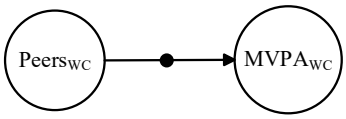
Next, I "decompose" the observed variables into their between-cluster and within-cluster latent variables by drawing representations of the between-cluster components above the top horizontal line and the within-cluster components below the bottom horizontal line. Ultimately, I will rearrange these variables to reflect a causal model among concepts that capture between-cluster and within-cluster causal dynamics. Here is the new diagram:



Note that I use circles to indicate the components because technically they are latent (unmeasured) representations whose values are inferred from the observed variables. The exception is the treatment condition (scored 1 = intervention, 0 = control) because it is a global Level-2 variable that only occurs at the between-cluster level; it has no within-cluster variability. Nor is it latent because there is a one-to-one correspondence between the measure and the construct it represents, namely assignment to treatment condition. Next, I arrange the top and bottom portions of the figure to form the hypothesized causal models at the between-cluster and within-cluster levels, yielding Figure 25.2.

The causal depictions at each level follow traditional influence diagram conventions. Each endogenous variable in the model is presumed to have an intercept. The intercepts calculated on within-cluster data typically are assumed to vary across clusters as this is necessary to statistically adjust for cluster-induced dependencies. However, if one wants and it seems substantively and statistically justifiable to do so, one can tell Mplus to set any intercept to be non-varying. There are no hypothesized varying path coefficients in the above model, but if there was, this would be signified using a black dot on the target path.

For example, suppose I thought that the magnitude of the effect of peer support on MVPA meaningfully varies across the different clusters independent of sampling error. I would signify this in the Figure as follows:



MSEM influence diagrams can become quite complex sometimes losing their heuristic value. Researchers often deal with the complexity by breaking the diagram into pieces and showing different parts of the overall model in different figures.

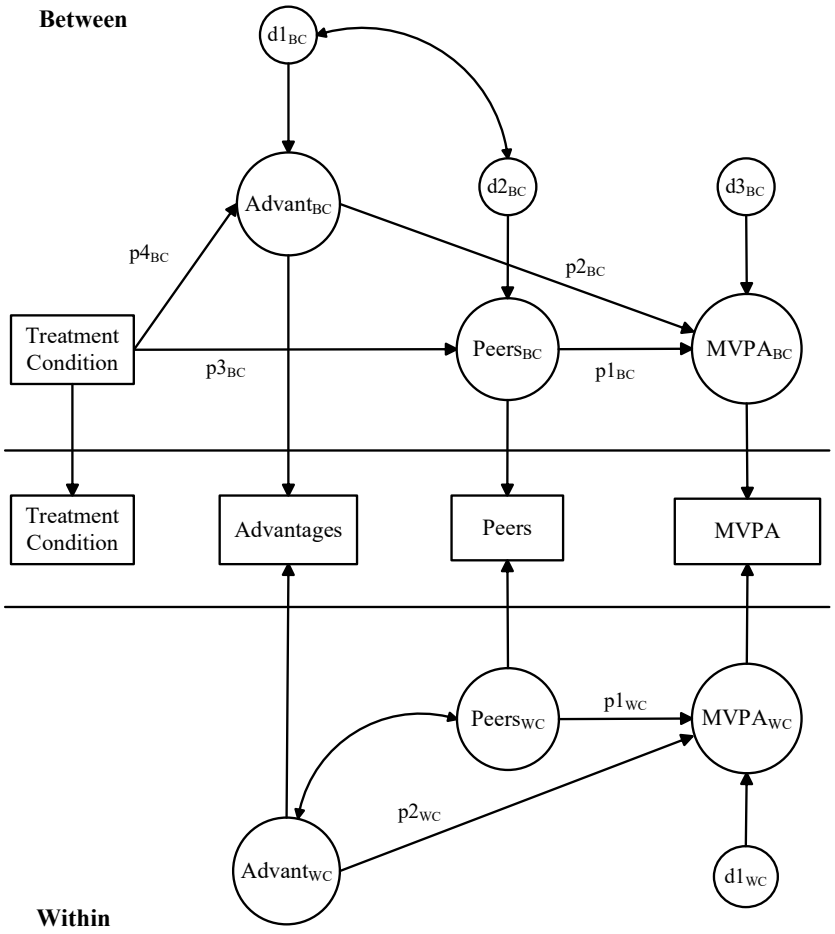


FIGURE 25.2. MSEM influence diagram

MSEM Analysis of the Numerical Examples

In this section, I apply the aforementioned concepts to the two numerical examples. I show you MSEM programming using Bayesian methods of analysis because these are becoming the method of choice (see Chapter 8 for an introduction to Bayesian SEM).

RET for School Intervention for Moderately Vigorous Physical Activity

The syntax for the program for the MVPA example appears in Table 25.3.

Table 25.3: MSEM Syntax for MVPA Example

```
1.  TITLE: MSEM analysis ;
2.  DATA: FILE IS mvpa.dat ;
3.  VARIABLE:
4.  NAMES ARE
5.    mvpa peers advant treat school  ;
6.  USEVARIABLES ARE
7.    mvpa peers advant treat ;
8.  CLUSTER is school ;
9.  BETWEEN IS treat ;          ! specify global/integral level 2 variables
10. ANALYSIS:
11. TYPE = TWOLEVEL ;
12. ESTIMATOR = BAYES ;
13. BITERATIONS=100000 (50000); BCONVERGENCE =.01;
14. MODEL :
15. %WITHIN%                   ! specify within cluster model
16. mvpa ;                   ! estimate within disturb var of outcome
17. mvpa ON advant peers (p1wc p2wc); ! regress mvpa on within mediators
18. peers WITH advant ;               ! correlate predictors
19. %BETWEEN%                           ! specify between cluster model
20. [mvpa] ;                          ! estimate outcome intercept
21. [advant] ;  [peers] ;             ! estimate mediator intercepts
22. mvpa ;                            ! estimate disturb var of outcome
23. advant ;  peers ;                 ! estimate disturb var of mediators
24. mvpa ON advant peers (p3bc p4bc) ; ! regress outcome onto mediators
25. advant ON treat (p1bc) ;          ! regress advant onto treatment
26. peers ON treat (p2bc) ;           ! regress peers onto treatment
27. advant WITH peers;                ! correlate disturbances
28. MODEL CONSTRAINT:                   ! define contrasts
29. NEW (medadv medpeer tot con1 con2) ; ! give names to contrasts
30. medadv = p1bc*p3bc ;                ! omnibus mediation for advant
31. medpeer = p2bc*p4bc ;               ! omnibus mediation for peers
32. tot = medadv + medpeer ;            ! total effect of treatment
33. con1 = p3bc-p1wc ;                  ! context effect 1
34. con2 = p4bc-p2wc ;                  ! context effect 2
```

```
35. OUTPUT: STAND(STDYX) RESIDUAL CINTERVAL(HPD) TECH4 TECH8 ;
```

Lines 1 through 8 should be familiar and do not need further comment. Line 9 specifies variables that are global or integral Level-2 variables; they have no within-cluster variability and do not represent aggregates of Level-1 variables. The analysis type on Line 10 is specified as two level; Mplus also offers a three level option. Lines 12 and 13 are the standard syntax I use to invoke Bayesian SEM and that you have encountered in prior chapters. Line 15 tells Mplus you will specify the within-cluster model and Line 19 tells Mplus the ensuing lines will be for the between-cluster model. For the within-cluster model, I tell Mplus to estimate the disturbance variance for the outcome (Line 16), the two within-cluster path coefficients (Line 17) and to allow the two predictors on Line 17 to be correlated. Line 20 tells Mplus to estimate the between-cluster intercept of the outcome. Line 21 tells Mplus to estimate the between-cluster intercepts for the two endogenous mediators. Lines 22 to 23 tell Mplus to estimate the disturbance variances for the three between-cluster endogenous variables. Lines 24 to 26 specify the between-cluster path coefficients (with labels) and Line 27 allows for correlated disturbances between the between-cluster mediators. Lines 28 to 34 use the Mplus `MODEL CONSTRAINT` feature to calculate the omnibus mediation tests for the between-cluster model for the effect of the treatment on the outcome using syntax I have covered in prior chapters. I explain Lines 33 and 34 below in the context of the Mplus output. Line 35 specifies the desired output.

Model Fit. I discussed fit indices for Bayesian SEM in Chapter 8. The potential scale reduction (PSR) ratio should be less than 1.1 (some prefer a standard of 1.05) to indicate adequate convergence. A separate PSR is calculated for each model parameter with the results being reported in the `TECH8` output section. Here is the output at the last iteration:

```
                POTENTIAL        PARAMETER WITH
ITERATION    SCALE REDUCTION       HIGHEST PSR

50000            1.000                3
```

The largest PSR at the final iteration was 1.00, which suggests the model converged. Mplus also conducts a Kolmogorov-Smirnov (KS) test of convergence which should be statistically non-significant (Mplus only prints the KS test result if $p < 0.05$ for it). For the current model, the KS result was not printed, also suggesting convergence.

In place of the p value for the traditional chi square test, Mplus reports a posterior predictive p-value; a questionable model fit is suggested by a p value < 0.05. Mplus also provides a 95% confidence interval for the difference between observed and replicated chi-square values based on replicated data sets of the same size as the original data during the iterative process. A good fitting model will produce a value of zero close to the middle of

the confidence interval; if zero is not in the confidence interval, it suggests a poor model fit. Here is the relevant Mplus output, which is consistent with good model fit:

```
Bayesian Posterior Predictive Checking using Chi-Square

          95% Confidence Interval for the Difference Between
          the Observed and the Replicated Chi-Square Values

                              -17.902            18.391

          Posterior Predictive P-Value               0.489
```

The above fit indices do not distinguish whether model fit is acceptable at both the between-cluster level and within-cluster level. In many studies, within-cluster sample sizes (the total N) are much larger than the between-cluster sample size (the number of clusters) so that overall model fit is dominated by the fit of the within-cluster model. One approach for segregating the respective model fits is called **partially saturated modeling**; see Ryu & West, 2009. In the current case, the within-cluster model is saturated so any ill fit is between-cluster in nature.

For localized fit, Bayes MSEM does not produce modification indices nor residual tests, but it does provide predicted correlations between the variables for the between-cluster model and for the within-cluster model; these can be compared visually with the between-cluster correlations and within-cluster observed correlations on a cell-by-cell basis per Chapter 8. However, to obtain the respective observed correlations, you need to execute specialized syntax separately that for the current example appears in Table 25.4:

Table 25.4: Syntax for Descriptive Statistics in Two Level Designs

```
1. TITLE: Two level descriptives ;
2. DATA: FILE IS mvpa.dat ;
3. VARIABLE:
4. NAMES ARE
5. mvpa peers advant treat school  ;
6. USEVARIABLES ARE
7. mvpa peers advant treat ;
8. CLUSTER is school ;
9. BETWEEN IS treat ;         ! specify global/integral level 2 variables
10. ANALYSIS:
11. TYPE = TWOLEVEL BASIC ;
12. OUTPUT:  ;
```

The first 10 lines are identical to the syntax in Table 25.3. Line 11 tells Mplus to calculate

descriptive statistics for the two level design. Here is the output that reports the within-cluster and between-cluster observed correlations:

```
   ESTIMATED SAMPLE STATISTICS FOR WITHIN

           Correlations
              MVPA          PEERS         ADVANT
              ________      ________      ________
MVPA            1.000
PEERS           0.475         1.000
ADVANT          0.478         0.279         1.000


    ESTIMATED SAMPLE STATISTICS FOR BETWEEN

           Correlations
              TREAT         MVPA          PEERS         ADVANT
              ________      ________      ________      ________
TREAT           1.000
MVPA            0.255         1.000
PEERS           0.335         0.582         1.000
ADVANT          0.330         0.505         0.135         1.000
```

The above correlations do not take into account prior distributions or any Bayesian concepts; indeed, they are maximum likelihood estimates of the correlations. Some would argue that such correlations are not appropriate for evaluating model fit in a Bayesian context but, again, as crude flags of potential model problems, I have found the comparison of predicted and observed correlations in this fashion to be helpful.

Here are the predicted correlations for the model from the `RESIDUAL OUTPUT` section of the main multilevel Bayes analysis:

```
     WITHIN LEVEL

           Correlations
              MVPA          PEERS         ADVANT
              ________      ________      ________
MVPA            1.000
PEERS           0.475         1.000
ADVANT          0.478         0.279         1.000

    BETWEEN LEVEL

           Correlations
              TREAT         MVPA          PEERS         ADVANT
              ________      ________      ________      ________
TREAT           1.000
MVPA            0.299         1.000
PEERS           0.313         0.578         1.000
ADVANT          0.309         0.500         0.122         1.000
```

The within-cluster predicted and observed correlations perfectly match one another

because this portion of the model is just-identified. The between-cluster predicted and observed correlations match reasonably well suggesting satisfactory model fit.

The program in Table 25.4 also reports the intraclass correlations for the mediators and for the outcome. Here is the relevant output:

```
Estimated Intraclass Correlations for the Y Variables

              Intraclass              Intraclass              Intraclass
    Variable  Correlation   Variable  Correlation   Variable  Correlation

    MVPA         0.461      PEERS        0.565      ADVANT       0.562
```

There is substantial between-cluster variability in the variables. Technically, these ICCs are based on maximum likelihood estimation. For large N, the results typically will be close in value to ICCs based on Bayes estimation and I often rely on them accordingly. If you want the formal Bayes estimates of the ICCs, you can use the program provided on my website in the document called *Computation of Bayes ICCs*.

Coefficients. I primarily am interested in the between-cluster coefficients because the treatment variable occurs only at the between-cluster level. Here is the edited output for the relevant unstandardized coefficients:

```
MODEL RESULTS

                                    Posterior    One-Tailed          95% C.I.
                    Estimate         S.D.       P-Value    Lower 2.5%  Upper 2.5% Sig

Between Level

 MVPA       ON
    ADVANT             4.601       1.191      0.000       2.291       6.973      *
    PEERS              5.303       1.145      0.000       3.089       7.604      *

 PEERS      ON
    TREAT              0.775       0.343      0.013       0.097       1.449      *

 ADVANT     ON
    TREAT              0.731       0.329      0.014       0.078       1.376      *

New/Additional Parameters
    MEDADV             3.204       1.784      0.014       0.190       7.079      *
    MEDPEER            3.961       2.052      0.013       0.245       8.256      *
    TOT                7.344       2.719      0.002       2.412      13.067      *
    CON1              -0.272       1.243      0.412      -2.681       2.189
    CON2               0.674       1.196      0.286      -1.643       3.058
```

As noted in Chapter 8, Mplus uses the 95% credible interval (labeled `95% C.I.`) to conduct significance tests relative to a null hypothesis of zero effect. If the credible interval does not contain the value zero, the effect in question is declared statistically significant

(the * in the last column of the output signifies a statistically significant result). Mplus also reports a one tailed p value for the parameter in question. For a positive estimate, the p-value is the proportion of the posterior distribution that is below zero; for a negative estimate, the p-value is the proportion of the posterior distribution that is above zero. The idea is that the reported p value maps onto a one-sided p value for the test that the parameter equals zero; one can obtain a rough analog of a two-sided p value by doubling it.

The total effect of the intervention on MVPA is shown in the contrast called `tot` in the `New/Additional Parameters` section of the output. The estimated between-cluster mean difference in MVPA between the treatment and control groups was 7.34 (95% credible interval (CI) = 2.41 to 13.07); the intervention increased MVPA on average by just over 7 minutes per day. The lower margin of error for the estimate is 2.41 - 7.33 = -4.92 and the upper margin of error is 13.07-7.33 = 5.74. The group difference is statistically significant, $p < 0.05$, but the lower limit of the credible interval is below the meaningfulness standard of 2.85, meaning I can't say with confidence that the intervention produced a meaningful result. The results are similar to what I find when I treated clusters as nuisance variables using PML estimation.

The estimated between-cluster effect of the intervention on perceived advantages was 0.73 units (± 0.65, 95% CI =0.08 to 1.38). The effect is statistically significant, $p < 0.05$. As with my prior analysis of this RET facet, the lower limit of the credible interval was less than the meaningfulness standard of 0.50, so that even though the sample mean difference estimate is promising (0.73), I cannot confidently conclude the intervention produced a meaningful effect on the perceived advantages mediator, as also was the case in the PML analysis.

The estimated between-cluster effect of the treatment condition on peer support was 0.78 (95% CI = 0.10 to 1.45). The lower margin of error for the coefficient is -0.68 and the upper margin of error is 0.65. The effect is statistically significant, $p < 0.05$, but like the perceived advantages mediator, the lower limit of the confidence interval did not exceed the meaningfulness standard.

Finally, the estimated between-cluster effect of the perceived advantages mediator on MVPA was 4.60 (95% CI = 2.29 to 6.97); for every one unit that the across cluster perceived advantages increases, the mean between-cluster MVPA per day is predicted to increase by 4.60 minutes. The lower margin of error is -2.31 and the upper margin of error is 2.37. The effect is statistically significant, $p < 0.05$ but the lower limit of the credible interval was smaller than the meaningfulness standard of 3.0. The estimated between-cluster effect of the peer support mediator on MVPA was 5.30 (95% CI = 3.09 to 7.60). The lower margin of error is -2.21 and the upper margin of error is 2.30. The effect is statistically significant, $p < 0.05$ and is meaningful because the lower limit of the credible

interval is larger than the meaningfulness standard.

Using the joint significance test, both perceived advantages and peer support mediate some of the effect of the intervention on MVPA across clusters; neither mediational chain exhibited a broken link across the links of the chain because both links were statistically significantly different from zero.

I can further evaluate the strength of the between-cluster effects of interest using the principles discussed in Chapter 10. I leave this as an exercise for you. Here are the estimated squared multiple correlations for the across cluster endogenous variables in the analysis:

```
R-SQUARE

Between Level

                          Posterior  One-Tailed         95% C.I.
    Variable        Estimate       S.D.     P-Value   Lower 2.5%  Upper 2.5%

    MVPA               0.531       0.106      0.000       0.315       0.723
    PEERS              0.099       0.075      0.000       0.000       0.248
    ADVANT             0.096       0.074      0.000       0.000       0.246
```

The estimated between-cluster squared correlation predicting MVPA from perceived advantages and peer support was 0.53 (lower MOE = -.22, upper MOE = 0.19). The estimated squared correlation reflecting the effect of the intervention on perceived advantages was 0.10 and this also was true for peer support.

The omnibus mediation effects appear in the `New/Additional Parameters` section of the output. The estimated between-cluster effect of the intervention on MVPA through the perceived advantages mediator was 3.20 (95% CI = 0.19 to 7.08). The effect is statistically significant, $p < 0.05$. The estimated between-cluster effect of the intervention on MVPA through the peer support mediator was 3.96 (95% CI = 0.24 to 8.26). The effect also is statistically significant, $p < 0.05$.

In sum, the intervention had a non-zero effect on the mean MVPA per day. The total effect mean difference (7.34) was suggestive that the effect was meaningful, but the width of the credible interval was sufficiently wide that I could not conclude this was the case with confidence. Both perceived advantages and peer support had non-zero effects on MVPA but only the latter had a sufficiently narrow credible interval to conclude the effect was meaningful. The intervention had non-zero effects on both of the targeted mediators but the wide credible intervals for them did not permit strong conclusions of meaningfulness.

Additional Analyses. A phenomenon of interest in some clustered randomized trials

is that of context effects. In the present example, the question of context effects focuses on whether the effect of a mediator on the outcome at the between-cluster level differs from the effect of that mediator on the outcome at the within-cluster level. If the respective path coefficients are different, then this suggests that there is something about the cluster/school context that enhances (if the path coefficient is stronger at the between level than at the within level) or mutes (if the path coefficient is weaker at the between level than at the within level) the effect of M on Y at the individual level as reflected by the within level analysis. To be sure, there are other ways that context effects are defined (e.g., Asparouhov & Muthén, 2019; Diez-Roux, 2002; Lüdtke et al., 2008; Raudenbush & Bryk 2002), but the above contrast is often key. In the MVPA syntax, Lines 33 and 34 of Table 25.3 request from Mplus the between versus within level contrasts for each of the two MVPA mediators in the contrasts I called `con1` (for perceived advantages, the path coefficient at the between-cluster level minus the path coefficient at the within-cluster level) and `con2` (for peer support). I reported above the path coefficients for the effects of the mediators on the outcome at the between-cluster level. I repeat them here for convenience coupled with the within-cluster results:

	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	Upper 2.5%	Sig
Between Level						
MVPA ON						
ADVANT	4.601	1.191	0.000	2.291	6.973	*
PEERS	5.303	1.145	0.000	3.089	7.604	*
Within Level						
MVPA ON						
ADVANT	4.874	0.354	0.000	4.180	5.558	*
PEERS	4.634	0.341	0.000	3.962	5.299	*

Here are the results from the output in the `New/Additional Parameters` section:

	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	Upper 2.5%
New/Additional Parameters					
CON1	-0.272	1.243	0.412	-2.681	2.189
CON2	0.674	1.196	0.286	-1.643	3.058

Neither mediator showed evidence of a statistically significant contextual effect given the presence of 0 within both of the 95% credible intervals.

Another analysis of potential interest addresses whether the path coefficient reflecting

the effect of a given mediator on the outcome varies meaningfully across clusters/schools. For example, the model tests I conducted above assumed that the population within-cluster/school path coefficients reflecting the effect of perceived advantages on MVPA were the same across all clusters/schools and this also was true for peer support. Is this a reasonable assumption or could the population coefficients vary across the clusters/schools? The issue is relevant because if the coefficients vary but I treat them as non-varying then the fitted model is mis-specified and this can undermine coefficient estimates, their standard errors, their p values and/or their credible intervals. Similarly, if the coefficients do not vary but I model them as varying, this can create estimation issues.

The Mplus syntax for analyzing the MVPA model with varying cluster coefficients for the perceived advantages predictor is shown in Table 25.5. The syntax is very similar to that of Table 25.3. I highlight using a red font the most relevant changes.

Table 25.5: MSEM Syntax for MVPA Example with Varying Slopes

```
1. TITLE: MSEM analysis ;
2. DATA: FILE IS mvpa.dat ;
3. VARIABLE:
4. NAMES ARE
5. mvpa peers advant treat school  ;
6. USEVARIABLES ARE
7. mvpa peers advant treat ;
8. CLUSTER is school ;
9. BETWEEN IS treat ;          ! specify global/integral level 2 variables
10. ANALYSIS:
11. TYPE = TWOLEVEL RANDOM;
12. ESTIMATOR = BAYES ;
13. BITERATIONS=100000 (50000); BCONVERGENCE =.01;
14. MODEL :
15. %WITHIN%                          ! within cluster model
16. mvpa ;
17. vs1 | mvpa ON advant ;          ! specify varying slope for advant
18. mvpa ON peers (p2wc) ;          ! specify non-varying slope for peers
19. peers WITH advant ;             ! allow predictors to be correlated
20. %BETWEEN%                         ! between cluster model
21. [mvpa] ;                        ! estimate outcome intercept
22. [advant] ; [peers] ;            ! estimate mediator intercepts
23. [vs1] ;                         ! estimate average of the varying slope
24. mvpa ;                          ! estimate disturb var of outcome
25. peers ; advant                  ! estimate disturb var of mediators
26. vs1 ;                           ! estimate across cluster var of path
27. mvpa ON advant peers (p3bc p4bc) ; ! regress outcome onto predictors
28. advant ON treat (p1bc) ;          ! regress advant onto treatment
29. peers ON treat (p2bc) ;           ! regress peers onto treatment
30. advant WITH peers;                ! correlate disturbances
```

```
31. OUTPUT: STAND(STDYX) RESIDUAL CINTERVAL(HPD) TECH4 TECH8 ;
32. PLOT: TYPE = PLOT3 ;
```

On Line 11, I specify the model type by including the word `RANDOM` to indicate there will be one or more varying effects. On Line 17, I define the varying effect using the label `vs1`, the symbol | which stands for "defined as" and to the right of it I indicate the path coefficient of MVPA on perceived advantages. Wherever I use the label `vs1` in the syntax, I am referencing this varying path coefficient. In the process of adding this syntax line I simultaneously tell Mplus to regress MVPA onto perceived advantages within each cluster. Line 23 tells Mplus to calculate the mean of `vs1` across clusters and Line 26 tells Mplus to calculate the variance of `vs1` across clusters. On Line 32 I add a plot command to produce a plot of the posterior distribution of the vs1 variances.

Here is the output for the DIC fit index for this model:

```
Deviance (DIC)                        13224.021
```

For the original model with the perceived advantages → MVPA path set to non-varying status, the DIC was 13220.45. The difference in the DICs is minor (see Chapter 8), and if anything, they favor the non-varying effects model (the lower the DIC, the better the model-data correspondence). I conclude based on this test that there is not strong evidence for allowing the perceived advantages → MVPA path to vary across clusters.

Here are the results for the mean and estimated variance of `vs1` across clusters for the model that allowed for non-varying coefficients:

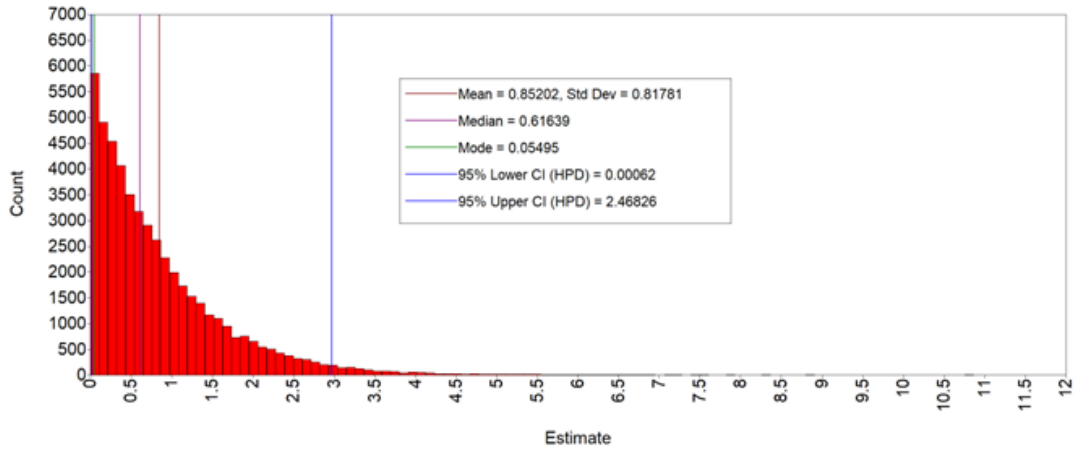
```
                                Posterior   One-Tailed          95% C.I.
                    Estimate       S.D.     P-Value    Lower 2.5%  Upper 2.5%   Sig
Means
    VS1                4.862      0.382       0.000        4.150       5.640      *

Variances
    VS1                0.616      0.818       0.000        0.001       2.468      *
```

The estimated average coefficient across clusters for the effect of perceived advantages on MVPA was 4.86 with an across-cluster variance estimate of 0.616. The variance seems subjectively small to me, supporting the treatment of the coefficients as non-varying. The data for this example are hypothetical and when I created the population data for it, I made the two mediator-to-outcome coefficients non-varying. The above results are consistent with the data generating population model.

Another diagnostic is to examine the plot of the posterior distribution for the across cluster variance estimates using the Mplus PLOT command. To obtain this plot, I chose the menu option on the plots tab of Mplus called "Bayesian Parameter Posterior

Distributions" and then choose "Parameter 14 %BETWEEN% vs1). Here is the plot



Note the stacking of the distribution near zero. Mplus also offers a significance test in the `TECH16` option that evaluates a null hypothesis of a zero variance, but its accuracy is not well known.

Overall, across these many forms of analysis, a reasonable conclusion seems to be that the coefficient does not meaningfully vary across clusters.

Parenthetically, if both of the path coefficients in a given mediational chain are treated as varying in your final model, then the formulae for computing the omnibus mediation effect for that chain and the total effect in the `MODEL CONSTRAINT` command in Lines 28 to 32 of Table 25.3 no longer hold; we need to incorporate into the commands the covariance between the two slopes. This rarely occurs in clustered RETs because the T→M link in the mediational chain only occurs at the between-cluster level so the within-cluster effect of T→M cannot vary across clusters. For a discussion of such adjustments, see Bauer, Preacher and Gil (2006).

Numerous simulation studies have been conducted to determine ways you might be misled if your model is misspecified relative to the variation of coefficients across clusters. LaHuis et al. (2020) found that treating the coefficient variance across clusters for a predictor as zero when the variance is, in fact, non-zero creates more bias in between-cluster standard errors than treating the coefficients as varying when they should be treated as uniform. Type I errors were more likely when the coefficient was treated as non-varying but it should have been treated as varying. Type II errors were more likely when the coefficient was treated as varying but it should have been treated as non-varying (see also the results of Algina & Swaminathan, 2011; Barr et al. 2013; Bell et al. 2019; Heisig & Schaeffer 2019; Hoffman & Walters, 2022; and Ye & Daniel 2017).

To lessen the possibility of Type I errors, some researchers suggest making all

coefficients varying by default in the initial program, letting the data speak for itself regarding the presence of coefficient variability (e.g., Barr et al. 2013; Heisig & Schaeffer 2019). However, this strategy often results in more model nonconvergences and it usually lowers statistical power (Park et al. 2020). Some researchers instead recommend making the decision to treat a predictor as having varying coefficients based on the magnitude of coefficient variability observed in the data (Hoffman & Walters, 2022) or on the basis of the results of the above significance tests and/or DIC indices. However, the above tests bring their own statistical baggage to data analysis and ultimately can undermine multilevel statistical theory about sampling distributions, as I discussed in Chapter 11 for preliminary tests more generally. They are best used with caution and should be theory driven.

RET for Group Level Intervention to Increase Pandemic Mask Wearing

I next analyze data for the second example, namely a group interactive intervention (n = 10 people per group) designed to influence mask wearing during the COVID pandemic. The primary outcome variable is the intention to regularly wear a mask in the future as measured on a -5 to +5 multi-item scale. The control group also receives an interactive intervention but it is on eating nutritious foods, a topic irrelevant to the outcome variable. The intervention targets as mediators the individuals' attitudes toward wearing masks (i.e., the perceived advantages of wearing a mask and the perceived disadvantages of not wearing one) and norms surrounding mask use, with each assessed on a multi-item scale whose total score ranges from -5 (strongly disagree) to +5 (strongly agree). There were 50 small groups in the intervention condition and 50 such groups in the control condition.

What sets this study apart from the MVPA study is that the 10 members in a given group are **not** conceptualized as a random sample from a larger cluster/group. In the MVPA study, I selected a sample of 20 students from each of the schools in the study. The 20 students were a random sample from their respective school and I used these 20 students to estimate the mean perceived advantages of MVPA and the mean peer support in a school. The means for a given school, of course, are subject to sampling error because the sample mean of the 20 students from a given school will not perfectly represent the true mean of all students for that school. Mplus takes this sampling error into account when applying MSEM. By contrast, in the mask wearing study, the 10 members from a group are not a random sample from a larger group; the group is what it is and the mean attitude toward wearing a mask regularly for the 10 members of a group is indeed the mean of that group (absent measurement error) as is the case for norms. In this study, we do not necessarily want to correct for sampling error for a given group mean as representing a broader group mean because there is no sampling error for it. To be sure, each group *is* conceptualized as having been randomly selected from a larger population of groups, but the members within

a group define the characteristics of that group in a formative sense (Lüdtke et al., 2008). In contrast to the MVPA example, we need to instruct Mplus not to correct for sampling error for the various cluster means as representing the broader cluster; instead we treat the means as global Level-2 variables, per my earlier discussion of global and contextual variables. There are different ways one can approach such data, namely whether to treat the groups as formative or treat them as a sample from a larger population.

In the current example, I believe the arguments of Lüdtke et al. (2008) favoring formative aggregation are compelling. Formative aggregation for a variable using Level 1 data (also called **compilation aggregation**) is such that variation in group members on that variable can be thought of as a substantively important group characteristic in its own right rather than merely representing a subsample whose mean contains sampling error relative to the mean of a larger population. If a researcher seeks to evaluate the biological sex composition of students in each of a large number of different classes and has information for all students within each class, then it is not unreasonable to adopt a formative Level 2 index, such as the % of females in the class as a sampling error free Level 2 measure. If a particular class happens to have a disproportionate number of females, this feature of the class reflects a true characteristic of that class. In such cases, the percentage of females in a class is a formative construct for purposes of multilevel data analysis. Similar points have been made by Bliese (2000) and by Kozlowski and Klein (2000). I approach my analysis of group-based cluster variables accordingly.

The relevant syntax for the analysis is presented in Table 25.6.

Table 25.6: MSEM Syntax for Group Intervention

```
1. TITLE: MSEM analysis for group interactive intervention ;
2. DATA: FILE IS group.dat ;
3. DEFINE:
4.  matt = CLUSTER_MEAN (att);
5.  mnorm = CLUSTER_MEAN (norm);
6. VARIABLE:
7. NAMES ARE
8.  intent norm att treat group ;
9. USEVARIABLES ARE intent treat matt mnorm ;
10. CLUSTER is group ;
11. BETWEEN IS mnorm matt treat ; ! specify global level 2 vars
12. ANALYSIS: TYPE = TWOLEVEL ;
13. ESTIMATOR = BAYES ;
14. BITERATIONS=100000 (50000); BCONVERGENCE =.01;
15. MODEL :
16. %WITHIN%                        ! no within cluster variables
17. %BETWEEN%                       ! between cluster model
18. [intent] ; [matt] ;  [mnorm] ;  ! estimate intercepts
```

```
19. intent ;  matt ;  mnorm ;       ! estimate disturbance variances
20. intent ON matt mnorm (p3 p4) ;  ! estimate mediator paths
21. matt on treat  (p1) ;           ! estimate treat effect on att
22. mnorm ON treat (p2) ;           ! estimate treat effect on norm
23. matt WITH mnorm ;               ! allow for correlated disturbances
24. MODEL CONSTRAINT:               ! estimate mediation and total effects
25. NEW (medatt mednorm tot) ;
26. medatt = p1*p3 ;
27. mednorm = p2*p4 ;
28. tot = p1*p3 + p2*p4 ;
29. OUTPUT: STDYX RESIDUAL CINTERVAL(HPD) TECH4 TECH8 ;
```

All of the syntax should be self-explanatory except Lines 4 and 5. The raw data file for Mplus takes the form of a traditional Mplus data file with individuals listed as rows and variables as columns. I need to analyze the *observed* predictor means for each cluster as global Level-2 variables, so I create a new set of variables using the `DEFINE` command and the `CLUSTER_MEAN` transformation within it. The transformation creates a variable that is the average of the values of the individual-level target variable for each cluster separately. I label these variables with an *m* in front of them, but you can use any label. Because the `treat` variable already is global in character, I do not need to calculate its per cluster mean. These transformed variables must be listed at the end of the `USEVARIABLES` line.

Model Fit. Here is the output for the last iteration of the Bayesian analysis:

```
              POTENTIAL          PARAMETER WITH
ITERATION    SCALE REDUCTION      HIGHEST PSR

     50000           1.000            5
```

The largest PSR at the final iteration was 1.00, suggesting convergence. No Kolmogorov-Smirnov (KS) tests of convergence were printed, which also suggests convergence.

Here are the chi square statistics and the posterior predictive p-value for the model, all suggesting adequate fit:

```
Bayesian Posterior Predictive Checking using Chi-Square

          95% Confidence Interval for the Difference Between
          the Observed and the Replicated Chi-Square Values

                              -15.170           15.704

          Posterior Predictive P-Value              0.499
```

There are only between-level covariance and correlation matrices with more than one element in this analysis, so any ill fit should show up in disparities between predicted

versus observed correlation matrices at the between level. I used the syntax in Table 25.7 to calculate the "observed" correlations between the variables (again, these are maximum likelihood estimates, which I am using informally).

Table 25.7: Syntax for Descriptive Statistics

```
1. TITLE: MSEM analysis for group interactive intervention ;
2. DATA: FILE IS group.dat ;
3. VARIABLE:
4. NAMES ARE
5. group intent norm att treat ;
6. USEVARIABLES ARE
7. intent att norm treat ;
8. CLUSTER is group ;
9. BETWEEN IS intent norm att treat;
10. ANALYSIS:
11. TYPE = TWOLEVEL BASIC;
12. OUTPUT:
```

Here is the output that reports the between-cluster observed correlations:

```
     ESTIMATED SAMPLE STATISTICS FOR BETWEEN


           Correlations
              INTENT        NORM          ATT           TREAT
              ________      ________      ________      ________
 INTENT         1.000
 NORM           0.470         1.000
 ATT            0.388        -0.015         1.000
 TREAT          0.177        -0.075         0.401         1.000
```

Here are the predicted between-level correlations for the model from the `Residual` output for the original syntax in Table 25.6 based on Bayesian estimation :

```
           Correlations
              INTENT        NORM          ATT           TREAT
              ________      ________      ________      ________
 INTENT         1.000
 NORM           0.511         1.000
 ATT            0.421        -0.014         1.000
 TREAT          0.130        -0.072         0.390         1.000
```

The between-cluster predicted and observed correlations match reasonably well.

Coefficients. Here is edited output for the key unstandardized coefficients of interest:

```
Between Level

                                  Posterior   One-Tailed          95% C.I.
                    Estimate       S.D.      P-Value    Lower 2.5%  Upper 5%   Sig

INTENT     ON
    MATT               0.527       0.110       0.000       0.314       0.744      *
    MNORM              0.558       0.096       0.000       0.369       0.746      *

 MATT       ON
    TREAT              0.481       0.114       0.000       0.257       0.702      *
 MNORM      ON
    TREAT             -0.102       0.141       0.235      -0.372       0.185

New/Additional Parameters
    MEDATT             0.247       0.081       0.000       0.102       0.414      *
    MEDNORM           -0.055       0.081       0.235      -0.222       0.100
    TOT                0.195       0.115       0.040      -0.033       0.421
```

The estimated total effect of the intervention on mask wearing intentions is shown in the contrast `tot` in the `New/Additional Parameters` section. The estimated mean difference between the treatment and control groups was 0.195 (95% credible interval (CI) = -0.03 to 0.42) which is just below the meaningfulness standard of 0.20. In traditional null hypothesis testing, the intervention effect was not statistically significant (the one tailed p value, doubled, was 0.082 and the value of zero occurred within the credible interval). However, as discussed in Chapter 9, the joint significance test logic suggests the presence of a total effect because there is at least one non-broken mediational chain linking the treatment to the outcome, as you will see shortly. Given this, I am inclined to declare an effect is present, but the mean difference is slightly less than the meaningfulness standard of 0.20 and certainly does not sustain the more conservative evaluation of meaningfulness that takes into account sampling error via the lower limit of the credible interval.

The estimated effect of the intervention on attitudes as reflected by the intervention vs. control group difference was 0.48 (±0.23, 95% CI =0.25 to 0.70). The effect is statistically significant, $p < 0.05$. The meaningfulness standard was 0.33 and the lower limit of the credible interval is not greater than it. Thus, I can't say with confidence that the effect is meaningful but the difference of 0.48 is suggestive. The estimated effect of the treatment condition on norms was -0.10 (95% CI = -0.38 to 0.17). The effect is not statistically significant and the mean change is weak and could be zero. It represents a broken link in the chain from the treatment condition to the outcome through norms.

I can evaluate the strength of the effects using principles discussed in Chapter 10, which I leave for you as an exercise. The estimated squared multiple correlations are:

```
R-SQUARE

Between Level
```

Variable	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	Upper 2.5%
INTENT	0.386	0.077	0.000	0.233	0.532
MATT	0.152	0.061	0.000	0.037	0.273
MNORM	0.007	0.019	0.000	0.000	0.054

The estimated between-cluster squared correlation predicting intention to wear a mask from attitudes and norms was 0.39 (± 0.15). The estimated eta squared reflecting the effect of the intervention on perceived advantages was 0.15 (± 0.12) and for norms it was 0.007.

The estimated between-cluster effect of the attitude mediator on the intention to wear a mask regularly was 0.53 (95% CI = 0.31 to 0.74); for every one unit that the cluster attitude increases, the cluster mean intent to wear a mask regularly is predicted to increase by 0.53 units. The effect is statistically significant, $p < 0.05$ but the lower limit of the 95% credible interval (0.31) was just slightly below the meaningfulness standard of 0.33.

The between-cluster effect of norms on the cluster mean intent to wear a mask regularly was 0.56 (95% CI = 0.37 to 0.75). The effect is statistically significant, $p < 0.05$. The lower limit of the 95% credible interval (0.37) is larger than the meaningfulness standard of 0.33, so we judge the effect to be meaningful with strong confidence (95% confidence).

Although they are of lower priority for RETs for program evaluation, the omnibus mediation effect for each mediator appears in the `New/Additional Parameters` section of the output. The estimated effect of the intervention on intent through the attitude mediator was 0.25 (95% CI = 0.10 to 0.41). The effect is statistically significant, $p < 0.05$. The estimated effect of the intervention on intent through the norms mediator was -0.06 (95% CI = -0.22 to 0.10). The effect is not statistically significant due to the intervention failure to meaningfully impact norms.

Additional Analyses. There is no reason to test for context effects in the current example because there are no within-cluster path coefficients to compare with corresponding between-cluster path coefficients. Some methodologists might argue that one use traditional MSEM with both within and between-cluster representations of the intention, attitude and norm variables but this is controversial because Mplus then makes Lüdtke's bias adjustments that assumes the individuals within a cluster are a random sample from a larger population of cluster members. Perhaps one can justify such a conceptualization in some cases. Another possibility is to treat the groups as nuisance variables and analyze the data using PML methods that adjust for clustering.

Moderation Analyses in Multilevel SEM

Moderation analyses in multilevel SEM can be straightforward but they also can be nuanced and challenging. I describe examples in Section III of my book on moderation analyses. The core logic of the underlying statistical theory is described in Asparouhov and Muthén (2020). I provide an example here that uses Bayesian estimation applied to a single mediation model to keep matters simple. The model is shown in Figure 25.3. The figure shows a single mediator, single outcome RET model with three interchangeable indicators for the mediator and the outcome. I make my main points using this model.

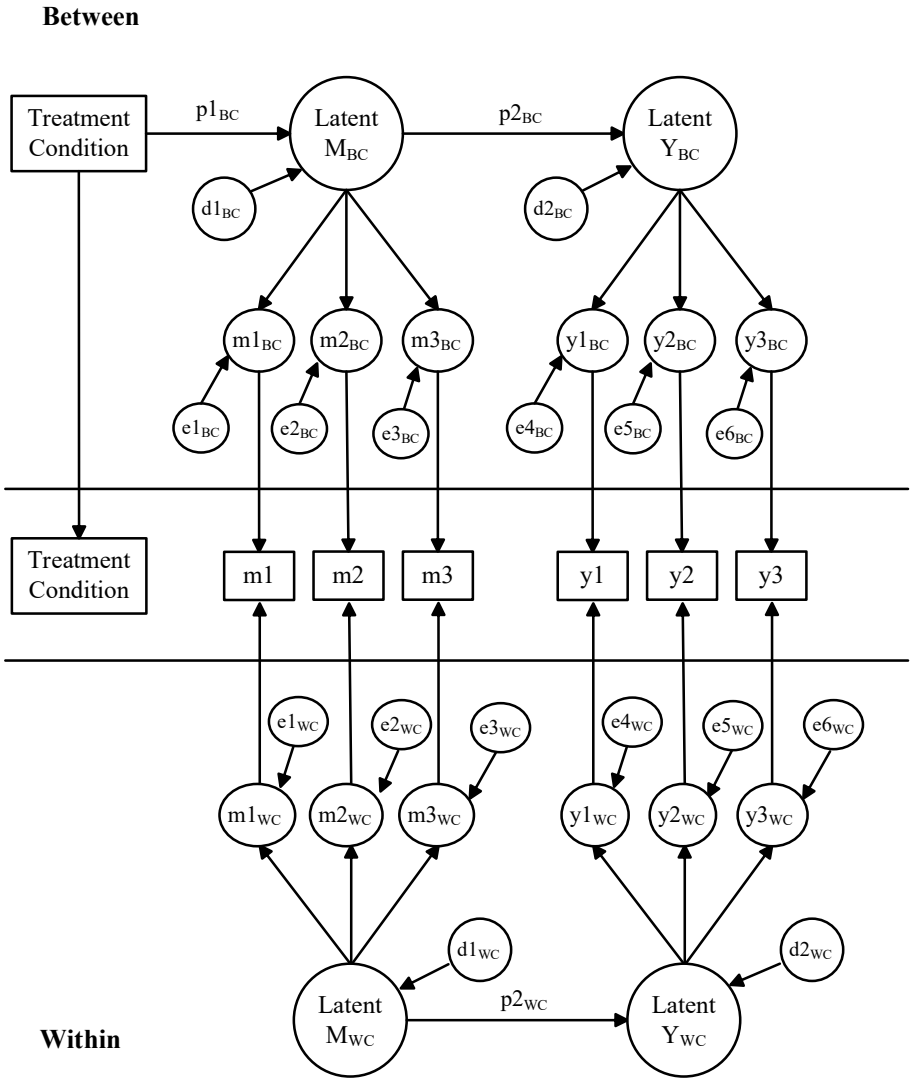


FIGURE 25.3. MSEM influence diagram with measurement multiple indicators

I seek to test if the treatment condition moderates the effect of the latent mediator on the latent outcome, i.e., I test for a treatment-mediator interaction. I make use of the `XWITH`

feature in Mplus for purposes of representing the interaction effect. This method can be used to model interactions or moderated relationships between an observed variable and a latent variable but it technically assumes the latent mediator and the observed variable both are normally distributed (see Chapters 15 and XX). This is clearly not the case when the treatment condition is binary with approximately equal sample sizes in the two conditions. Gonzalez & Valente (2023) evaluated eight different approaches to testing treatment by latent mediator moderation/interactions, one of which was Bayesian estimation with the `XWITH` command for a model similar to that in Figure 25.3. They varied in a Monte Carlo study a wide range of factors including sample size, effect size, distributional properties of the latent mediator indicators. The Bayes method produced reasonable estimation, power, and credible interval coverage across all conditions. The approach seems viable despite the non-normal binary variable in the interaction. Table 25.8 shows the Mplus syntax for the model in Figure 25.3, most of which you are familiar with:

Table 25.8: Syntax for Test of Moderation in MSEM

```
1. TITLE: MSEM invariance analysis ;
2. DATA: FILE IS moderation.dat ;
3. VARIABLE:
4. NAMES ARE
5.   y1 y2 y3 m1 m2 m3 treat school ;
6. USEVARIABLES ARE
7.   y1 y2 y3 m1 m2 m3 treat ;
8. CLUSTER is school ;
9. BETWEEN IS treat ;  ! specify global/integral level 2 variables
10. ANALYSIS:
11. TYPE = TWOLEVEL RANDOM ;
12. ESTIMATOR = BAYES ;
13. BITERATIONS=100000 (50000); BCONVERGENCE =.01;
14. MODEL :
15. %WITHIN%  ! specify within model
16. y1 ; y2 ; y3 ;  ! estimate resid var of outcome indicators
17. m1 ; m2 ; m3 ;  ! estimate resid var of mediator indicators
18. lyw by y1 y2 y3 (pw1-pw3) ; ! define measurement model for y
19. lmw by m1 m2 m3 (pw4-pw6) ; ! define measurement model for m
20. lmw ; ! estimate var of latent mediator
21. lyw ; ! estimate disturbance variance of latent outcome
22. lyw ON lmw (pw7) ; !regress latent outcome on latent mediator
23. %BETWEEN% !specify between model
24. y1 ; y2 ; y3 ; ! estimate resid var of outcome indicators
25. m1 ; m2 ; m3 ; ! estimate resid var of mediator indicators
26. lyb by y1 y2 y3 (pb1-pb3); ! define measurement model for y
27. lmb by m1 m2 m3 (pb4-pb6); ! define measurement model for m
28. lyb ; ! estimate disturbance variance of latent mediator
```

```
29. lmb ; ! estimate disturbance variance of latent outcome
30. lmb on treat (pb7) ; ! regress latent med onto treatment
31. int| treat XWITH lmb ; !define interaction term
32. lyb ON lmb treat int (pb8-pb10) ; ! regress latent outcome onto
33.                                   ! interaction and component parts
34. MODEL CONSTRAINT: ! conduct simple effects
25. NEW (streat scontrol diff) ; ! give labels to contrasts
36. streat = pb8 + pb10 ; ! simple effect of ly on lm for treat grp
37. scontrol = pb8 ; ! simple effect of ly on lm for control grp
38. diff=streat-scontrol ! check on interaction
39. OUTPUT: STDYX Cinterval(hpd) TECH4 TECH8 RESIDUAL ;
```

Line 31 creates the moderation/interaction term using the `XWITH` keyword. As review from Chapter XX, the label for the moderation/interaction term is placed to the left of `|` and to the right of it are the two variables to be used in the term separated by `XWITH`. At least one of the variables must be a latent variable. Line 32 specifies the linear equation predicting the latent outcome from the moderator/interaction term and its component parts. The `MODEL CONSTRAINT` commands starting on Line 34 define the two simple effects for the path coefficient reflecting the effect the latent mediator on the latent outcome, one for the treatment group (Line 36) and the other for the control group (Line 37). The underlying algebra was discussed in Chapter XX and the expressions rely on the Mplus labels for the relevant path coefficients. Line 38 calculates the difference between the two simple effects and should equal the coefficient for the `int` variable in Line 32.

When I analyzed the data, I found that the largest PSR on the final iteration was 1.000, suggesting the model converged. No global fit indices are reported by Mplus for the model, which is a limitation of the Bayesian approach in this particular case. I can compare the predicted and observed correlations for the observed variables as an informal check on model fit, per my earlier discussion in this chapter. I do not show the matrices here but the model estimated correlations were reasonably close to the observed correlations at both the between-cluster and within-cluster levels. I discuss in the document on my webpage *Moderation Analyses in Multilevel SEM* approaches to use for moderation models.

To save space, I do not report results for the measurement model (the results for it were reasonable) and focus only on the moderation analyses. Here is the relevant output for the unstandardized coefficients:

```
MODEL RESULTS

                                    Posterior   One-Tailed         95% C.I.
```

```
                    Estimate       S.D.      P-Value    Lower 2.5%  Upper 2.5%  Sig

Between Level

 LYB        ON
    LMB                0.224      0.187      0.112      -0.144       0.587
    INT                0.432      0.250      0.038      -0.047       0.936
    TREAT             -0.336      0.277      0.106      -0.878       0.222

New/Additional Parameters
    STREAT             0.657      0.168      0.000       0.338       1.000      *
    SCONTROL           0.224      0.187      0.112      -0.144       0.587
    DIFF               0.432      0.250      0.038      -0.047       0.936
```

The moderator/interaction term (`INT`) is not statistically significant because the 95% credible interval for it contains zero (-0.047 to 0.936). The coefficient for it was 0.432, which estimates the difference between the latent variable M→Y path coefficient for the intervention group minus the corresponding path for the control group. The estimates for the component individual path coefficients are in the section `New/Additional Parameters`. The estimate for the intervention group is 0.657 (95% CI = 0.338 to 1.000) and for the control group it is 0.224 ((95% CI = -0.144 to 0.587). The estimated squared correlation for the between-cluster latent Y regressed onto the between-cluster latent mediator, treatment effect and the moderation/interaction term was:

```
R-SQUARE

Between Level
                              Posterior   One-Tailed          95% C.I.
    Variable        Estimate       S.D.      P-Value   Lower 2.5%  Upper 2.5%

    LYB                0.290      0.092      0.000       0.119       0.476
```

or 0.290 (95% credible interval = 0.119 to 0.476).

Assumptions

Like most statistical methods, multilevel models make statistical assumptions. The major ones are: (1) the model is correctly specified (all the predictors associated with the outcome and relevant random effects are part of the model), (2) the functional form of relationships is correct (e.g., linearity), (3) Level-1 disturbances are independent and normally distributed, (4) Level-2 disturbances are independent and normally distributed, (5), disturbances at Level 1 and Level 2 are unrelated, and (6) predictors at one level are not related to disturbances at another level.

Assumptions about normality and variance homogeneity are less pressing when robust or certain forms of Bayesian estimation are used in multilevel modeling.

Specification error is always of concern. With multilevel models, the issue of mistakenly treating a predictor as having varying coefficients across clusters when they are, in fact, non-varying or vice versa has received considerable attention. As noted, simulations suggest that the effects of such violations depend on how severe the misspecification is: Treating a varying coefficient predictor as non-varying is not particularly problematic if the amount of variation in the coefficients is modest. Treating a non-varying coefficient predictor as varying across clusters also is not particularly problematic because the lack of coefficient variability will reveal itself during the modeling process and will be taken into account accordingly. Most of the other assumptions can be addressed using methods outlined in Chapter 11 for preliminary analyses with continuous outcomes. I usually create two sets of scores in my wide-format data set, scores that are cluster mean centered and that map onto the Level 1 data and scores that are either cluster means that map onto the Level 2 data. I then apply the methods for preliminary analyses outlined in Chapter 11 to each data set to gain perspectives on potential problems.

The Use of Covariates in MSEM

The examples I have considered do not include covariates for confounder control. It is reasonably straightforward to include covariates to adjust for confounds for either the within-cluster model, the between-cluster model, or both; you just include the covariate in one or both of the `%WITHIN%` and `%BETWEEN%` statements of the Mplus code as you would in any standard regression model. If lagged outcomes are included, bias can result given the presence of random effects, so care must be taken accordingly (see Chapter 16).

An important nuance in covariate control is if the covariate is nominal (such as biological sex or ethnicity) and you want to include it in both the within-cluster and between-cluster models. At the within-cluster level, you would dummy code the variable using traditional 0-1 dummy coding and enter into the equation the dummy variables for all groups except the reference group. This also would be true for the between-cluster model. Using Bayesian MSEM, you would not mention any of the dummy variables on the `WITHIN` and `BETWEEN` subcommands under the `VARIABLE` command, thereby telling Mplus to invoke latent variables for them. The coefficients for the dummy variables for the within-cluster model are interpreted as in traditional regression – they represent the average cluster mean difference between the group scored 1 on the dummy variable and the reference group. However, the coefficient for the between-cluster model takes on a different meaning because the dummy variable now reflects the *mean* of the 0-1 scores for the dummy variable in question in each cluster, This mean equals the proportion of individuals in each respective cluster that have a score of 1 on the dummy variable. Yaremych, Preacher & Hedeker (2023) suggest interpreting the between-cluster coefficient by dividing it by 10.

Let me make this concrete for you. Suppose I code biological sex as 0 = female and 1 = male and include the dummy variable (which I call biosex) in both the within-cluster and between-cluster models as covariates. Suppose my outcome variable is the number of minutes of vigorous physical exercise per day (MVPA). If the coefficient for biosex in the within-cluster model is 3.0, this means that within-clusters, males are predicted to engage in MVPA, on average, three minutes more per day than females, i.e., it is the mean for the group scored 1 on the dummy variable minus the reference group, holding the other variables in the equation constant. Suppose the coefficient for biosex in the between-cluster model is 20.0. If I divide this by 10, I obtain 2.0. I interpret the result as follows: For every 10% increase in the percentage of males in a cluster (school), the cluster average number of MVPA minutes per day is predicted to increase by 2.0, holding constant the other predictors in the between-cluster equation.[4] By dividing by 10, this allows me to use the phrase "for every 10% increase."

As another example, suppose the nominal variable is ethnicity with three levels, (a) non-Hispanic White, (b) Black, (c) Latinx. I create two dummy variables with dummy coding, one for Blacks and the other for Latinx, with Whites being the reference group. I include D_{BLACKS} and D_{LATINX} as covariates in the within-cluster model and also in the between-cluster model. Suppose the coefficient for D_{BLACKS} was 1.0. This means that within-clusters, Blacks are predicted to engage in MVPA, on average, one minute more per day than Whites. Suppose the coefficient for D_{BLACKS} in the between-cluster model was 15.0. If I divide this by 10, I obtain 1.50 and interpret the result as follows: For every 10% increase in the percentage of Blacks in a cluster (school), the cluster average number of MVPA minutes per day is predicted to increase by 1.5, holding constant the other predictors in the between-cluster equation.

There is some controversy about how best to implement the treatment of nominal covariates in multilevel modeling. Asparouhov and Muthén (2018b) argue that the above approach is widely applicable and has the advantage of explicitly dealing with Lüdtke bias. For binary nominal variables, Enders and Tofighi (2007) suggest instead using observed group mean centering rather than latent variable centering, but this approach is subject to Lüdtke bias and is not as widely applicable (see Asparouhov & Muthén, 2018b, for details).

Concluding Comments on Multilevel SEM

[4] In the unusual case where each cluster contains entirely one category of the nominal variable (i.e., the categorical predictor has no within-cluster variability), the coefficient for the variable will equal the mean difference on y between that group and the reference group and be interpreted as the mean difference on the outcome when moving from a cluster composed entirely of the reference group to a cluster composed entirely of the group scored 1 on the dummy variable.

Multilevel modeling is an approach for analyzing clustered RET data when you seek to gain perspectives on the clusters themselves. MSEM is distinct from traditional multilevel modeling as implemented by software like HLM and MLwin and what is commonly known as *mixed modeling*, although these strategies as well as SEM more generally can be thought of as special cases of MSEM. MSEM is superior to traditional multilevel modeling because it can incorporate latent variables into the analysis, it can deal with complex structural relationships that are not easily evaluated in traditional multilevel models, and it addresses often overlooked conflation of effects. In the current chapter, I focused on the case of MSEM with continuous outcomes for single indicator models. However, it also can be applied to binary and ordinal outcomes as well as count outcomes (for an example of an application to a binary outcome, see Cho, Preacher & Bottge, 2015). Extensions to binary outcomes/mediators are reasonably straightforward but complications do arise (Huang, 2023; Hayes, 2024). I include a document on my webpage called *Multilevel Structural Equation Modeling with Binary Outcomes* that considers such cases in more depth and provides detailed syntax for them. I also include a document on the use of multiple indicator latent variables on my webpage, titled *Latent Variables in MSEM*.

COMPARISON OF MSEM AND CLUSTERS-AS-NUISANCE APPROACHES

The cluster-as nuisance approach to the analysis of cluster randomized trials is rooted in traditional regression modeling for individuals but standard errors are adjusted for dependencies using sandwich-based corrections. Unlike multilevel modeling, there is no desire to characterize random variances nor are we interested in the variance of slopes across clusters, at least in the sense that multilevel models do. One assumes the same slope value applies to each cluster, just as we do in standard regression for individuals. To be sure, their likely will be some cluster differences in slopes but the assumption is that the variability is not meaningful and reflects sampling error. If this is not the case, then one models the variability using product terms rather than the MSEM random slopes approach.

At the outset of this chapter I illustrated the cluster-as nuisance approach but did not include Level-2 predictors with the exception of the assignment of individuals to treatment versus control conditions, a global Level-2 variable. It turns out that one can include both global and contextual Level-2 predictors in the cluster-as-nuisance approach. If you do so, then clustering probably is no longer viewed strictly as a nuisance because you are now exploring how cluster characteristics impact the outcome. Given this, I henceforth refer to the approach as the **cluster robust standard error approach** (CRSE) and distinguish it from MSEM by its use of specialized sandwich type standard errors to specify both within-cluster and between-cluster effects on outcomes in a more traditional SEM context. To be

sure, MSEM has more analytic flexibility for exploring non-varying slopes across clusters and to adjust for Lüdtke's bias but it also comes with a host of assumptions that may be unrealistic in some settings. McNeish, Stapleton & Silverman (2017) provide a useful and detailed comparison of the CRSE and traditional multi-level approaches. Sometimes, I find it helpful to analyze Level-1 and Level-2 predictors of an outcome using a CRSE approach rather than MSEM.

Recall that global Level-2 variables assign the same value to all individuals in a cluster (such as the size of a school representing a cluster) while contextual Level-2 variables assign the mean value for the cluster for all individuals in that cluster (such as the mean SES of individuals in a school or the percent of students in the school who are non-white if schools). Including a global Level-2 variable in the CRSE approach is straightforward because there is no within-cluster variability associated with it; its variation is completely determined by between-cluster differences on the predictor. For Level-2 predictors that are the average of a Level-1 predictor, however, one needs to include both the Level-2 predictor as well as the Level-1 predictor in the linear equation(s) in order to make the interpretation of their respective coefficients meaningful. For the Level-1 predictor, you enter it into the equation using one of two versions depending on your research question (Snijders & Bosker, 2012; Antonakis et al., 2021).

The first version of the Level-1 contextual predictor is to cluster-mean center it before entering it into the equation. In this case, you subtract the mean of the cluster from each cluster members' original score and then include the transformed score as your predictor. You also include as a separate predictor the Level-2 version of the variable, which is simply the cluster mean assigned to each individual in their respective cluster. The resulting path coefficient for the Level-1 predictor will then estimate the within-cluster effect of the predictor on the outcome and the coefficient associated with the average of the Level-2 predictor will estimate the between-cluster effect of the predictor on the outcome.

To illustrate this approach using the MVPA example, I adapt the cluster-as-nuisance syntax I showed you at the beginning of the chapter (see Table 25.2) to the syntax in Table 25.9, where I now include between-cluster mean vectors for perceived advantages and peer support (Lines 5 and 6) after which I cluster mean center the level 1 raw scores (Line 7).

Table 25.9: Syntax for CRSE Approach: Version 1

```
1. TITLE: CRSE Version 1 ;
2. DATA:
3.   FILE IS mvpa.dat ;
4. DEFINE:
5.   madvant = CLUSTER_MEAN (advant);  ! define level 2 var of cluster means
6.   mpeers = CLUSTER_MEAN (peers);
```

```
7. CENTER advant peers (GROUPMEAN) ;   ! cluster mean center level 1 vars
8. VARIABLE:
9. NAMES ARE
10.  mvpa peers advant treat school ;
11. USEVARIABLES ARE
12.  mvpa peers advant treat madvant mpeers ;  ! list new vars last
13. CLUSTER is school ;                        !identify cluster variable
14. ANALYSIS: TYPE = COMPLEX ;                 ! specify complex design option
15.  !BOOT = 5000 ;
16. MODEL :
17.  mvpa ON advant peers madvant mpeers ; ! regress Y onto level 1 and 2 vars
18.  mpeers on treat ;                     ! regress level 2 vars onto treat
19.  madvant ON treat ;
20.  madvant with mpeers ;  ! allow correlated disturbances for level 2 vars
21. MODEL INDIRECT:
22.  mvpa IND treat ;
23. OUTPUT: SAMP STANDARDIZED(STDY) RESIDUAL MOD(ALL 4)
24. CINTERVAL TECH4 ;
25. !CINTERVAL(BOOTSTRAP) TECH4 ;
```

I highlight in red the syntax you should be sure to note. Lines 5 and 6 create the Level-2 variables `madvant` and `mpeers` that assign the respective cluster mean value to each individual. Line 7 enacts the cluster mean centering transformation by subtracting an individual's cluster mean from his or her raw score on the variables. It is important that this transformation occurs after I define `madvant` and `mpeers`. Line 15 is commented out because I am not using bootstrapping in this initial run but I might do so in later runs for sensitivity purposes. This also is true for Line 25. Line 17 adds `madvant` and `mpeers` as predictors of MVPA and Lines 18 and 19 regress the Level-2 instantiations of the mediators onto the Level-2 dummy variable for the treatment condition. Line 20 allows for correlated disturbances between `madvant` and `mpeers` due to unmeasured common causes.

In the interest of space, I do not report the model fit indices but all of them were favorable. I instead focus on the unstandardized path coefficients that emerged in the analysis that illustrate my earlier points:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
MVPA ON				
ADVANT	4.873	0.341	14.273	0.000
PEERS	4.630	0.317	14.607	0.000
MADVANT	4.615	0.921	5.014	0.000
MPEERS	5.298	0.993	5.333	0.000

The coefficients for the first two predictors of MVPA (`advent` and `peers`) are the

estimated within-cluster effects of the two mediators on MVPA. The coefficients for the second two predictors of MVPA (`madvant` and `mpeers`) are the estimated between-cluster effects of the two mediators on MVPA. Let's compare them to their counterparts in the Bayesian MSEM analysis. Here are the between level effects from the Bayesian MSEM:

```
                                  Posterior  One-Tailed        95% C.I.
                    Estimate       S.D.      P-Value    Lower 2.5%  Upper 2.5% Sig

Between Level

 MVPA       ON
    ADVANT             4.601      1.191      0.000       2.291       6.973      *
    PEERS              5.303      1.145      0.000       3.089       7.604      *
```

You can see that the results are comparable to the CRSE analyses. Here are the Bayesian coefficients for the within level analysis:

```
Within Level

 MVPA       ON
    ADVANT             4.874      0.354      0.000       4.180       5.558      *
    PEERS              4.634      0.341      0.000       3.962       5.299      *
```

The CRSE results for the within-cluster coefficient also map closely onto these values.

Here are the between level results for the CRSE analysis that regress the mediators onto the treatment effects in order to isolate the effect of the treatment on them:

```
                                                    Two-Tailed
                    Estimate       S.E.    Est./S.E.    P-Value
MPEERS   ON
    TREAT              0.776      0.319      2.436      0.015

 MADVANT  ON
    TREAT              0.733      0.306      2.395      0.017
```

and here are the corresponding results from the Bayesian MSEM analysis:

```
                                  Posterior    One-Tailed        95% C.I.
                    Estimate       S.D.      P-Value    Lower 2.5%  Upper 2.5% Sig

Between Level

 PEERS      ON
    TREAT              0.775      0.343      0.013       0.097       1.449      *
 ADVANT     ON
    TREAT              0.731      0.329      0.014       0.078       1.376      *
```

Again, the results are comparable.

Result compatibility to the more complex Bayesian SEM is an attractive feature of the CRSE approach because it shows that one can use it to address key substantive questions about within-cluster and between-cluster effects in a cluster randomized trial and avoid shifting to a Bayesian framework, which might comport better with some audiences.

In MSEM, I noted its ability to test for contextual effects by formally testing differences in the effects of the mediators on the outcome at the between-cluster level with those same effects at the within-cluster level. It turns out I can also accomplish such tests in the CRSE approach by using version 2 coding of the Level-1 mediators. In this version, I execute the same syntax as in Table 25.9 but I do not cluster mean center the Level-1 scores for the mediators, i.e., I remove or comment out Line 7 from the Table 25.9 syntax. This step will not change the (within-cluster) coefficients for the Level-1 predictors but it will change the (between-cluster) coefficients for the Level-2 predictors. Specifically, the Level-2 coefficients will now estimate and test the context effects for each mediator. Here are the results for the key coefficients from the CRSE analysis when I do this:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
MVPA ON				
ADVANT	4.873	0.341	14.273	0.000
PEERS	4.630	0.317	14.607	0.000
MADVANT	-0.257	0.940	-0.274	0.784
MPEERS	0.667	1.040	0.642	0.521

Note that the within-cluster coefficients remain the same but the coefficient for MADVANT is the context effect, namely the between-cluster coefficient for MADVANT→MVPA minus the within-cluster coefficient ADVANT→MVPA or 4.601-4.873 = -0.257, which was statistically non-significant (margin of error = 1.88, critical ratio = 0.274, p = 0.784). The result maps well onto the corresponding contrast I performed in the Bayesian MSEM and this was also true for the peer support context effect. Thus, we also have the ability to use the CRSE approach to formally test for context effects.

In sum, we can use the CRSE method to analyze clustered randomized trial data and obtain perspectives on much of what we can with MSEM but all in a more traditional SEM framework. To be sure, MSEM has more flexibility in its ability to characterize and explore non-varying slopes and its ability to adjust for Lüdtke's bias. However, often CRSE is more audience friendly. Of course, both analytic frameworks can incorporate latent variables to address measurement error.

General Estimating Equations Revisited

The above CRSE strategies can also be used in conjunction with GEE modeling to obtain estimates of within-cluster, between-cluster, and context effects but from a population averaging perspective and without distributional assumptions associated with random intercepts. I introduced basic concepts for GEE estimation in Chapter 16 for panel regression but I review those concepts here for clustered randomized trials in the interest of continuity.

GEE adjusts for non-independence of within cluster scores using an estimation strategy known as **quasi-maximum likelihood**. It is an extension of generalized linear models with different link functions, so it can be applied to clustered data with continuous outcomes (using the identity link), with binary outcomes (using either logit or probit binary links) or with count outcomes. GEE works with fixed rather than random intercepts.

GEE estimates population-averaged coefficients rather than traditional conditional regression coefficients (see Chapter 16). Its parameter estimates adjust for covariates more akin to the way that average marginal effects do so per my discussion in Chapter 5. GEE estimates how the average outcome changes across the population for a unit change in the predictor taking into account the covariate distributions as a whole. The more traditional conditional models estimate the effect of a predictor on an outcome holding constant the covariates at specific predictor values. Often the results of the two analytic methods are quite similar but sometimes not. The relevant statistical theory and underlying mathematics are described in Hardin and Hilbe (2012) and Hoover, Shi, Burstyn and Anastos (2019). Estimation is often pursued using a form of robust sandwich estimation but this tends to not fare all that well with small N, in which case a jackknife method seems to be better (Hardin & Hilbe, 2012).

Suppose you have conducted an RCT with 100 groups of 5 individuals each with the groups being randomly assigned to treatment or control conditions. Group membership is treated as the cluster variable because the behavior of one group member might affect the behavior and outcomes for another member of the same group. Traditional statistical tests assume the scores of the five individuals in each group are independent but this may not be the case. The task for the GEE analyst is to make an educated guess about the dependency structure among the scores for the five group members of each group. This often is operationalized in the form of a correlation matrix. An **independence dependency structure** has 1s on the diagonal and 0s on the off diagonals of the 5X5 cluster membership matrix, indicating that the outcomes of the individuals within a cluster are not related to each other. In multilevel modeling terminology, this corresponds to an intraclass correlation coefficient of zero. Although you can express any matrix form, three commonly used dependency structures are the exchangeable structure, the first order autoregressive structure, and the unstructured structure, which are represented as follows using the

correlation ρ to represent the dependency:

Independence

	Y1	Y2	Y3	Y4	Y5
Y1	1	0	0	0	0
Y2	0	1	0	0	0
Y3	0	0	1	0	0
Y4	0	0	0	1	0
Y5	0	0	0	0	1

Exchangeable

	Y1	Y2	Y3	Y4	Y5
Y1	1	ρ	ρ	ρ	ρ
Y2	ρ	1	ρ	ρ	ρ
Y3	ρ	ρ	1	ρ	ρ
Y4	ρ	ρ	ρ	1	ρ
Y5	ρ	ρ	ρ	ρ	1

First order autoregressive

	Y1	Y2	Y3	Y4	Y5
Y1	1	ρ	ρ^2	ρ^3	ρ^4
Y2	ρ	1	ρ	ρ^2	ρ^3
Y3	ρ^2	ρ	1	ρ	ρ^2
Y4	ρ^3	ρ^2	ρ^3	1	ρ
Y5	ρ^4	ρ^3	ρ^2	ρ	1

Unstructured

	Y1	Y2	Y3	Y4	Y5
Y1	1	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{15}
Y2	ρ_{21}	1	ρ_{23}	ρ_{24}	ρ_{25}
Y3	ρ_{31}	ρ_{32}	1	ρ_{34}	ρ_{35}
Y4	ρ_{41}	ρ_{42}	ρ_{43}	1	ρ_{45}
Y5	ρ_{51}	ρ_{52}	ρ_{53}	ρ_{54}	1

The exchangeable structure assumes the dependencies between all the Ys are identical in magnitude; that each cluster member is affected by other members in the same way. The first order autoregressive structure assumes the dependencies follow a classic autoregressive/simplex pattern. It is commonly used for longitudinal panel models as discussed in Chapter 16, but it can applied to cross-sectional like settings under certain circumstances. For example, suppose within a cluster we can order individuals in terms of how proximal they are to one another on a dimension of interest. It might be that the dependencies for individuals who are more proximal to one another are stronger than the dependencies for individuals who are less proximal to one another in a way that follows autoregressive dynamics. In this case, to apply an autoregressive structure, one would want to be sure to order individuals within a cluster in the data set in accord with their proximity to one another. The most flexible structure is the unstructured pattern, where the dependencies can take on any pattern. Simulations suggest that the unstructured form coupled with the sandwich estimator is, in general, a good choice for a wide range of scenarios (Hardin & Hilbe, 2012) but it also tends to reduce statistical power and can lead to convergence issues. For this reason, some methodologists prefer the exchangeable structure in cross-sectional like studies considered in this chapter (Hoover et al, 2019).

GEE estimation is unique because it uses the *a priori* specified dependency structure to derive estimates, standard errors and significance tests of the core regression parameters. Simulation studies suggest that misspecifying the dependency structure often does not create bias in model estimates but it can lower efficiency of the parameter estimates in a strict statistical sense. Hoover et al. (2019) describe several scenarios where misspecification can affect bias. Some methodologists suggest it is good practice to estimate your model under different reasonable working dependency structures and determine sensitivity or robustness of conclusions to doing so accordingly.

The GEE approach does not require distributional assumptions in the way more traditional approaches do because estimation of the population-average model depends primarily on correctly specifying few features of the data-generating distribution (e.g., correct specification of the mean model and the working dependency structure), not the entire joint distribution.

The GEE method can be used in the spirit of the GLM based within-between frameworks described by using the same centering regression strategies. Specifically, for contextual Level 2 variables, one includes in the linear equation both the Level-2 predictor comprised of the mean value of the predictor for each cluster as well as the Level-1 predictor, either in cluster centered form or in its original untransformed form. The former strategy yields a coefficient for the Level 1 predictor that reflects the within-cluster effect while the coefficient for the predictor means reflects the between-cluster effect. Using the original untransformed score for the Level 1 predictor isolates contextual effects, as discussed above. For details, see McNeish, Stapleton and Silverman (2017). For a worked example, watch the video on my website associated with the program that executes GEE cluster regression.

In sum, there are a variety of ways you can pursue analyses of within-cluster and between cluster effects in clustered randomized trials, including MSEM, the CSRE approach, and GEE strategies. Each has its strengths and weaknesses and you can draw on these for your particular needs accordingly.

STRATEGIES WHEN THERE ARE FEW CLUSTERS

It is well known that many methods for analyzing clustered data are questionable when the number of clusters is small. For the CRSE approach, research suggests that standard errors can be biased downward when the number of clusters is small which, in turn, leads to higher Type I error rates (Cameron, Gelbach, & Miller, 2008; Cameron & Miller, 2015; Imbens & Kolesar, 2016; MacKinnon & Webb, 2017). As noted, about 50 clusters is generally considered to be sufficient, with 20 or so clusters being reasonable in some

contexts. For multilevel models, an early rule of thumb was known as the 30-30 rule that argued for at least 30 clusters of 30 observations each (Kreft, 1996), but this rule has fallen into disrepute. Hox and Maas (2001) conducted a simulation study to explore sample size requirements for MSEM and found that 50 clusters generally sufficed for low intraclass correlations and clusters of equal size. Under more general conditions, they recommended 100 clusters or more were likely needed (see also Hox, Maas & Brinkhuis, 2010). Usually, the best way to determine a reasonable number of clusters and sample size for your study is via computer simulation, which I address in Chapter 28.

There are several ad hoc methods that have been suggested for analyzing data with few clusters. One early approach analyzed data using standard methods that ignore clustering but then multiplies the resulting standard errors by a design effect correction, DEFT, before forming a z value or t ratio for purposes of significance testing. In cases where the probability of selection of each member of the population of clusters is equal and the cluster sizes are equal, the correction factor in many cases is

$$\text{DEFT} = \sqrt{1 + \text{ICC}(1 - n_c)}$$

where ICC is the intraclass correlation coefficient and n_c is size of each cluster. If the cluster sizes are unequal but close in magnitude, some researchers use the harmonic mean of the cluster sizes in place of n_c. For example, if the observed margin of error based on a 95% confidence interval of a statistic ignoring clustering is 3.0 and DEFT is 2.0, the cluster adjusted margin of error would be (3.0)(2.0) = 6.0. If the z value for a significance test for a parameter of a mean difference or regression coefficient is 1.50 when ignoring clustering, then the cluster adjusted z value would be (1.50)/(2.0) = 0.75. The DEFT correction approach is crude and tends to be conservative. Thomas and Heck (2001) suggest using more liberal alpha levels if applying the approach but do not provide guidelines for how to choose an alpha level. Hedges (2007, 2015) provides a correction multiplier to traditional z or t ratios that can be applied to studies that mistakenly ignore the effect of clustering by analyzing the data as if it were from a simple random sample. These corrections can be helpful when conducting meta-analyses. In the final analysis, better approaches to adjusting for cluster effects with few clusters are available. I discuss two such methods here, one called bias-reduced linearization (BRL; Bell & McCaffrey, 2002) and the other based on Bayesian modeling.

Bias-Reduced Linearization (BRL)

Several reasonably effective small sample corrections exist for clustered data (Manor & Zucker, 2004; Skene & Kenward, 2010a, 2010b; Zucker, Liberman, & Manor, 2000,

Kenward & Roger, 1997, 2009). One popular method is the Kenward-Roger correction (Bell et al., 2014; McNeish & Stapleton, 2014) which frequently is used in traditional multilevel models. I focus here instead on the **bias-reduced linearization** (BRL) method that uses the CRSE framework (see Bell & McCaffery, 2002). The BRL approach must be used on an equation by equation basis in limited information SEM contexts. For coefficient estimation, the method adopts ordinary least squares regression (or maximum likelihood estimation) but uses a cluster-robust standard error based on what is known as a CR2 estimator. The CR2 estimator is a generalization of the heteroskedasticity-consistent HC2 estimator proposed by MacKinnon and White (1985). It is a sandwich estimator. Unlike Mplus that relies on asymptotic theory, the method invokes empirically based degrees of freedom to define p values and confidence intervals. The degrees of freedom are called df_{BM} and are data-based. The degrees of freedom can be non-integer and tend to yield more conservative critical values than conventional methods (Pustejovsky & Tipton, 2018). Several simulations have affirmed the utility of the approach for models with as few as 10 to 20 clusters (e.g., Bell & McCaffery, 2002; Huang & Li, 2021; Imbens & Kolesar, 2016; Pustejovsky & Tipton, 2018). Huang et al. (2023) have extended the BLR approach to the analysis of binary outcomes for logistic or probit regression contexts. I provide a program for the method on my website and a video that walks you through the program.

The BLR approach only can evaluate complex SEM models using limited information estimation frameworks without latent variables. Level-2 variables can be included in the modeling per my above discussion of the CRSE approach. With Level-2 predictors, it does not address Lüdtke's bias. However, when faced with a small number of clusters, MSEM and the Mplus based CRSE approach become unfeasible; BLR can be a viable alternative, albeit one with limitations. Of course, the BLR approach does not solve the problem of low statistical power when analyzing few clusters in certain modeling contexts. The bottom line is that if you are going to conduct cluster randomized trials, you need to ensure you have plenty of clusters.

Bayesian Modeling with Informative Priors

Another approach to modeling clustered data with few clusters is to use Bayesian MSEM with informative priors. The present chapter emphasized Bayesian MSEM but using diffuse or noninformative priors. By contrast, if you use reasonably chosen informative priors, model estimation can yield smaller standard errors and less bias for parameter estimation. Zitzmann, Helm and Hecht (2021) suggest two strategies for defining informative priors for multilevel models with few clusters, (1) specify an informative prior for the cluster-level variance of cluster level predictors and (2) specify an informative prior for the path coefficient for one or more cluster level predictors. Zitzmann et al. discuss

factors to consider when using these strategies. In Chapter 28 on sample size decisions, I described how to incorporate informative priors into Bayesian frameworks in Mplus to address small sample sizes. I refer you to that chapter for an introduction to this topic and ways of implementing the suggestions of Zitzmann et al. (2021). Having said that, with diffuse prior distributions, Bayes methods with few clusters or with small sample sizes often perform no better and sometimes worse than traditional frequentist methods, so care must be taken in such cases (McNeish, 2016a; Depaoli & Clifton, 2015). The choice of informative priors is crucial.

Cluster Matching

When there are few clusters in a cluster randomized trial, some methodologists recommend using **cluster matching** for purposes of random assignment to treatment arm. The approach involves pairing clusters before randomization on key, theoretically relevant determinants of the outcome and then randomizing one member of the pair to each treatment arm. The idea is to reduce imbalance that can occur during randomization and to increase efficiency.

As discussed in Chapter 4, with randomization one expects baseline variable means and proportions to be equal but sometimes by virtue of random error, they are not, i.e., the treatment conditions are imbalanced. It usually is good scientific practice to adjust for such imbalance if possible but only for variables that matter, namely variables that impact the posttest outcome. One way of accomplishing this is through the introduction of covariates during data analysis. Another way is through matching during study design and execution.

Matching can vary from weak to strong. Strong matching occurs when all cluster pairs are perfectly matched on all target variables. Weak matching is when we are unable, for whatever reason, to form pairs with comparable scores on the target variables. Cluster matching tends to fail in the presence of weak matching (Chondros et al. 2021). I refer you to Chapter 4 for a more detailed discussion of imbalance, sample size, and matching. It is a strategy worth considering when analyzing data with few clusters. For analysis issues related to matched designs, see Diehr et al. (1995) and Martin et al. (1993).

POWER ANALYSIS/SIMULATIONS FOR CLUSTER RANDOMIZED TRIALS

There are numerous software packages available for power analysis for clustered randomized trials. I like the R package called PUMP by Porter et al., (2023a, b). Rutterford, Copas and Eldridge (2015) provide two simple formulae for estimating the per condition sample size needed to obtain a given level of power for a two arm treatment in a clustered randomized trial for (1) a mean difference between the intervention and control groups for a continuous outcome/mediator, and (2) a proportion difference between the intervention

and control groups for a binary outcome/mediator. Here is the formula for the continuous outcome:

$$m = \left[\frac{(Z_{1-\alpha/2} + Z_{\beta})^2)(2\sigma^2)}{\Delta^2} \right](1 + (n-1)\rho) \qquad [25.5]$$

where m = the group/condition sample size, $Z_{1-\alpha/2}$ is the z score of the standard normal distribution corresponding to $1-(\alpha/2)$, Z_{β} is the z score of the standard normal distribution corresponding to beta, i.e., the desired Type II error rate (power is one minus this value), Δ is the population mean difference of interest, σ^2 is the variance of the dependent variable, n is the cluster size, and ρ is the intra-class correlation. This formula assumes the cluster sizes are equal or close enough so to be reasonably represented by a single number.

As an example, if the alpha level is 0.05, then $Z_{1-\alpha/2}$ equals 1.96; if the desired power is 0.80, then Z_{β} equals 0.84. I might expect the typical cluster size to be 20 and the population intraclass correlation to be 0.05. I set the outcome variance to be 1.0 and the mean difference of interest to be 0.50, which maps roughly onto a Cohen's *d* of 0.50 given the variance equals 1.0. Substituting these values into Equation 25.5 yields an *m* value of approximately 91. I need about 90 individuals per condition, which will yield about 180/20 = 9 clusters, which is probably too few unless I use an analytic method appropriate for few clusters.

Here is the corresponding formula for a proportion difference:

$$m = \left[\frac{(Z_{1-\alpha/2} + Z_{\beta})^2)(P_1(1-P_1) + P_2(1-P_2))}{\Delta^2} \right](1 + (n-1)\rho) \qquad [25.6]$$

where P_1 is the target population proportion of responders in the intervention group, P_2 is the target population proportion of responders in the control group, Δ is $P_1 - P_2$ and all other terms are as previously defined. R code for executing Equations 25.5 and 25.6 is in the Appendix.

An interesting property of the above two equations is that the terms in the brackets yield approximations to the required per group sample size for the case of simple random sampling, i.e., they will be close in value to what you would get by conducting a traditional power analysis for a mean difference or a proportion difference. The result is then inflated by (i.e., multiplied by) an index of the anticipated design effect, in this case $1+((n-1)\rho)$, where n is the typical cluster size and ρ is the intraclass correlation. Some methodologists perform back-of-the-envelope estimates of sample size needs under clustering by

conducting a power analysis for sample size using traditional power analysis software for the case of simple random sampling and then multiplying the result by the above design effect expression. This requires, of course, that the design effect multiplier is appropriate for the type of analysis conducted; the multiplier can vary depending on the type of analysis performed and the nature of the clustering (see Rutterford et al., 2015).

Rather than rely on assumption bound approximations such as the above, I prefer to pursue power analysis through computer simulations. Although more work, the simulation strategy provides much more information and gives you more control over the power analysis than general power analysis software. Simulations also allow you to evaluate if the number of clusters you are analyzing is problematic and to gain perspectives on margins of error. I discuss how to construct such simulations in Mplus in Chapter 28 where I also provide an example for a clustered randomized trial.

METHODOLOGICAL ISSUES IN CLUSTER RANDOMIZED TRIALS

Partially Nested Designs

Cluster randomized trials sometimes use what are known as **partially nested designs**. These are designs where clustering occurs in some conditions but not others. For example, individuals in the intervention group might be given a treatment in small groups of 5 to 10 individuals but people in the control group are not given anything, hence there are no group clusters in the control condition. Some studies seek to formally compare group administered to individually administered interventions, in which case a partially nested design is used. These examples are distinct from **fully clustered designs** where clustering occurs in all study conditions. In partially clustered designs, it often is reasonable to assume that observations in the unclustered condition are independent but non-independent in the clustered condition.

The analysis of partially nested cluster trials has received considerable attention. Over half a dozen methods of analysis have been proposed (e.g., Bauer, Sterba & Hallfors, 2008; Baldwin, Bauer, Stice & Rohde, 2011; Sterba, 2017). Sterba et al. (2014) describe SEM methods for analyzing partially nested designs many of which rely on multiple group MSEM. I placed a link to the Sterba et al. article on the Resources tab of my webpage coupled with a link to their extensive Mplus syntax. Their programs use MSEM but with maximum likelihood estimation rather than Bayesian estimation. It is fairly straightforward to conduct multigroup SEM using maximum likelihood but this is not the case with Bayesian estimation. The latter requires the use of mixture modeling with known groups. I do not consider the analytic strategies discussed by Sterba et al. because their article and the accompanying syntax for it makes clear how to implement their approach. My own

preference is for simpler methods that work about as well.

One such option is a method known as **pseudo-clustering** (Candlish, Teare, Dimairo et al., 2018). The idea is to conduct traditional fully clustered analyses but to introduce arbitrary clustering into the non-clustered arm of the study, usually the control condition. One method of pseudo-clustering is to create artificial random clusters in the control group using the same number of clusters and cluster sizes as the intervention arm. This method introduces some bias in the estimation of the intraclass correlation because of the likely independence of within-cluster observations for controls. However, the bias may not be sufficiently strong to undermine the analysis. In their simulation studies, Candish et al. found the strategy yielded reasonable results. If you use this strategy it probably is wise to conduct a localized simulation in Mplus that maps onto your study conditions to ensure that Type I error rates are not inflated and that statistical power is adequate. I discuss how to construct such localized simulations in Chapter 28.

Therapists/Providers as Clusters

In many clinic-based randomized trials, therapies are administered by different therapists but in some trials, the same therapist individually treats multiple study participants. In such cases, therapists represent a form of clustering because patients are nested within therapists. There are different forms that such nesting can take. Sometimes the role of the therapist is inconsequential, such as when simply administering a shot or providing a medication for a patient to take, in which case clustering can be safely ignored. In other contexts, the role of the therapist is more substantial and can affect patient outcomes. In general, if patients in the respective treatment arms are assigned to therapists with different probabilities and if therapists can meaningfully affect outcomes, then therapist cluster effects may need to be taken into account. If patients in both treatment groups have an equal chance of being assigned to the same therapist, clustering may be ignorable.[5] If therapists treat patients in both arms, but are more likely to treat patients from a specific arm, clustering may need to be accounted for.

There are methods for dealing with multiple levels of clustering, such as patients within therapists within hospitals within counties. However, attempting to control for all levels of clustering can sometimes make complex analytic demands that may not work well in practice. Simpler analyses adjusting only for truly meaningful sources of clustering may be preferable.

[5] By ignorable, I mean that an unadjusted analysis will yield valid Type I error rates, but if the intraclass correlation is high, then statistical power can be adversely affected.

Multisite Designs

Cluster randomized trials are often undertaken at multiple sites. Multiple sites are sometimes used for non-substantive reasons, primarily as a means of increasing participant recruitment to achieve a large enough sample size for the study to be adequately powered. Site essentially operates as a nuisance variable in the broader context of the study. Other multisite studies are designed such that sites take on a theoretically meaningful role, such as to evaluate the effects of site-level variables on treatment effectiveness or to evaluate the generalizability of treatment effects across sites as a function of site characteristics. When the latter is the goal, researchers typically use a large number of sites and treat them as a higher level variable in multilevel analyses.

When a very small number of sites are viewed as nuisances, site effects usually are not treated as cluster variables in the ways I have described in this chapter. Rather, they are conceptualized as having fixed values with the sites being dummy coded and statistically controlled during the modeling process (see Feaster et al., 2011). Some researchers view such analyses as limiting the generalizability of the study to the particular sites that are studied but this is too narrow a conceptualization. Generalizations are indeed limited to the essential properties of the studied sites per my earlier discussion of cluster populations, but it is possible to select sites in ways that permit strong statements about generalizability. For example, if I select two sites that are highly disparate on their essential characteristics and show that my intervention has functionally the same effect in both sites, then this increases my confidence in the generalizability of the intervention to other types of sites as well. Per my discussion of meta-populations in Chapter 4, single site studies rarely randomly sample study participants yet study results typically are generalized to the types of participants that were included in the trial and not restricted to the study participants per se. Although this ultimately represents a form of nonstatistical generalization, it is a common practice in clinical trials and not unreasonable by fiat. The same is true for sites when treated as fixed effects.

CONCLUDING COMMENTS

Clustered randomized explanatory trials can be analyzed using SEM frameworks. The clusters can be treated either as nuisance variables or as meaningful units in their own right that we want to make inferences about. In either case, the clustering can create residual/error within-cluster dependencies that require specialized analytic structures. In Mplus, SEM models that treat clusters as nuisances typically are analyzed using maximum likelihood based sandwich estimators. Models that treat clusters as meaningful and that seek to explore effects at the cluster level use multilevel modeling strategies, most notably

MSEM. These strategies require a fairly large number of clusters in the study design. However, specialized small sample analytic strategies have evolved. The present chapter has only scratched the surface with respect to the analysis and design of cluster randomized trials.

There are many different types of cluster designs each with different analytic implications. For example, **step-wedged designs** sequentially transition clusters (such as schools, hospitals) from control to intervention conditions in a randomized order, thus representing a form of cross-over design but with clustering. No matter the design, it always is possible to bring mediators and moderators into these trials, thereby offering the power of randomized explanatory designs instead of outcome-only evaluations. Most current textbooks on cluster designs ignore mediation and moderation, which is unfortunate.

APPENDIX: R CODE FOR SAMPLE SIZE DETERMINATION

This appendix presents R code for executing Equations 25.5 and 25.6. For Equation 25.1, the R code is:

```
alpha <- .05 # specify alpha level of interest
beta <- 0.20 # specify desired Type 2 error or (1-power)
var <- 1.0   # specify variance of outcome
diff <- 0.50 # specify mean difference of interest
nclus <- 10  # specify typical sample size of cluster
icc <- 0.05  # specify intraclass correlation
zalpha <- qnorm(1-alpha/2)
zpower <- qnorm(beta)
unadjn <- (((zalpha-zpower)^2)*(2*var))/diff^2
groupn <- unadjn*(1+(nclus-1)*icc)
groupn # show result
```

and for Equation 25.6, the R code is

```
alpha <- .05 # specify alpha level of interest
beta <- 0.20 # specify desired Type 2 error or (1-power)
p1 <- .30    # specify proportion of responders for group 1
p2 <- .20    # specify proportion of responders for group 2
nclus <- 10  # specify typical sample size of cluster
icc <- 0.05  # specify intraclass correlation
diff = p1-p2
zalpha <- qnorm(1-alpha/2)
zpower <- qnorm(beta)
unadjn <- (((zalpha-zpower)^2)*(p1*(1-p1)+p2*(1-p2)))/diff^2
groupn <- unadjn*(1+(nclus-1)*icc)
groupn # show result
```