Published in final edited form as:
Multivariate Behav Res. 2008 April 2; 43(2): 210–236. doi:10.1080/00273170802034810.

Evaluating Group-Based Interventions When Control Participants Are Ungrouped

Daniel J. Bauer,
University of North Carolina at Chapel Hill

Sonya K. Sterba, and
University of North Carolina at Chapel Hill

Denise Dion Hallfors
Pacific Institute for Research and Evaluation

Abstract

Individually randomized treatments are often administered within a group setting. As a consequence, outcomes for treated individuals may be correlated due to provider effects, common experiences within the group, and/or informal processes of socialization. In contrast, it is often reasonable to regard outcomes for control participants as independent, given that these individuals are not placed into groups. Although this kind of design is common in intervention research, the statistical models applied to evaluate the treatment effects are usually inconsistent with the resulting data structure, potentially leading to biased inferences. This article presents an alternative model that explicitly accounts for the fact that only treated participants are grouped. In addition to providing a useful test of the overall treatment effect, this approach also permits one to formally determine the extent to which treatment effects vary over treatment groups and whether there is evidence that individuals within treatment groups become similar to one another. This strategy is demonstrated with data from the Reconnecting Youth program for high school students at risk of school failure and behavioral disorders.

Methods for analyzing data from randomized experiments have been widely disseminated for the case where the unit of randomization matches the unit to which treatment is administered. Approaches for analyzing data in which *individuals* are randomly assigned to *individually* administered treatments (e.g., individual therapy) are found in standard univariate and multivariate texts (e.g., Maxwell & Delaney, 2004; Neter, Kutner, Nachtsheim, & Wasserman, 1996). Approaches for analyzing data in which preexisting (intact) *groups* (e.g., clinics, classrooms, or neighborhoods) are randomly assigned to *group*-administered treatments, as in cluster-randomized designs, are also readily available (see Murray & Blitstein, 2003; Murray, Varnell, & Blitstein, 2004; Raudenbush, 1997). These latter approaches account for lack of independence of observations within group to protect the nominal Type I error rate, either through adjustments of the test statistic and degrees of freedom (e.g., Baldwin, Murray, & Shadish, 2005) or by use of a mixed-effects (multilevel) model (e.g., Janega et al., 2004).

A third type of design is also common in practice yet has received comparatively little methodological attention. Under this design, randomization to treatment is done on an individual basis; however, the treatment is administered in a group setting so that multiple

Correspondence concerning this article should be addressed to Daniel J. Bauer, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270. dbauer@email.unc.edu.

individuals receive the treatment together. The groups are not preexisting but rather are formed by the investigator solely for the purpose of treatment provision. To avoid confusing treatment conditions and treatment groups, we use the term *arm* to refer to the treatment or control conditions and the term *group* to refer to a particular group of participants receiving treatment together (where many such groups may exist that receive the same form of treatment). For example, participants suffering from depression might be assigned to one of two study arms: cognitive-behavioral group therapy (CBT) or control. Individuals assigned to CBT are administered treatment within small groups. Control participants, in contrast, are not placed into groups and have no particular relationship to one another. We refer to the data structure generated by this kind of design as *partially nested* to indicate that participants are nested within groups in at least one arm of the study, whereas in another arm of the study they are not.

To ascertain the prevalence of this kind of design, we conducted a literature review of all randomized experiments (N = 94) in four representative public health and clinical research journals: the *American Journal of Public Health* (2003–2005), *Evaluation Review* (2004–2005), *Journal of Consulting and Clinical Psychology* (2004–2005), and *Prevention Science* (2003–2005). This review indicated that partially nested designs (N = 30; 32%) were *more* common than group-randomized (fully nested) designs (N = 26; 27%) and almost as common as individually randomized (nonnested) designs (N = 38, 40%). A prototypical example of a partially nested design is provided by Carey et al. (2004), who randomized drug rehabilitation outpatients to group therapy for HIV prevention or no-treatment control.

The grouping of participants within the treatment arm but not the control arm complicates the evaluation of treatment effects. On one hand, observations within the control arm can reasonably be assumed to be independent and hence do not require adjustment for grouping effects. On the other hand, observations within the treatment arm will likely be correlated within groups. This correlation could arise because group members promote preventative behaviors via social support, because the fidelity of the treatment implementation differs across groups, or because the effectiveness of treatment providers varies across groups. Some group interaction effects may even interfere with treatment; for instance, through contagion of tactics for needle sharing (see Weiss et al., 2005, for discussion of such iatrogenic effects). Thus when participants within the treatment arm are clustered into groups, the independence assumption of conventional statistical methods for individually randomized designs will be violated.[1] Yet, at the same time, models developed for fully nested (group-randomized) designs are also not optimal, given the lack of a grouping structure in the control arm of the study.

Despite their common occurrence, very few methodological papers have directly addressed how applied researchers can appropriately evaluate treatment effects when using a partially nested design. In an important exception, Hoover (2002) provided an adjustment for the independent samples t test for the case when one sample consists of individuals within groups and the other consists of ungrouped individuals. This method can be used to contrast outcome measures for treated and control participants, but it does not generalize straightforwardly to accommodate multiple treatment or control arms, pretest scores or other covariates, additional follow-up measures, or nonnormal outcomes. Nor does it provide direct information on the nature of the dependence in the data (e.g., variability in treatment effects, homogenization of group behavior, etc.). Drawing on another suggestion made by

[1] A separate but related issue is the need to account for correlated observations generated by shared therapist effects. This issue has been explored at length in a recent series of papers (see Crits-Cristoph & Mintz, 2001; Crits-Christoph, Tu, & Gallop, 2003; Serlin, Wampold, & Levin, 2003; Siemer & Joormann, 2003a, 2003b; Wampold & Brown, 2005; Wampold & Serlin, 2000) and can arise in either individually randomized trials, group randomized trials, or partially nested trials.

Hoover, a pair of papers recently appeared in *Clinical Trials* offering an alternative approach for the analysis of partially nested data using mixed-effects models (Lee & Thompson, 2005; Roberts & Roberts, 2005). The latter approach is more flexible, overcoming the limitations of the more specific analysis suggested by Hoover. Even in the latter two papers, however, several key issues that arise specifically with partially nested designs were left unaddressed (e.g., obtaining unbiased standard errors and appropriate degrees of freedom, what to do with group-level covariates, and the validity of causal inferences).

In light of the recency of these methodological contributions, it is unsurprising that our literature review identified no cases of partially nested data structures that were analyzed with either of the two aforementioned approaches. Of the 30 studies that fit this design, 87% used analyses appropriate for individually randomized (nonnested) designs and the other 13% used analyses developed for group-randomized (fully nested) designs. *None* of the studies in our sample reported analyses specifically tailored to reflect the partially nested study design. In contrast, 65% of the fully group-randomized studies in our review properly accounted for the grouping structure. The latter result corroborates Bland's (2004) finding that the proportion of group-randomized trials properly accounting for clustering increased sharply from 1993 to 2003. We submit that this increase reflects the publication of a number of methodological papers addressing the analysis of group-randomized trials (e.g., Murray, 1998; Raudenbush, 1997).

Although intervention researchers appear to be increasingly aware of the need to account for dependence of observations in group-randomized trials, they remain unaware of methods to account for dependence in partially nested designs. Indeed, under the impression that their partially nested design could not be analyzed properly, Fromme and Corbin (2004) lamented that "it was not possible to assess the impact of group composition on treatment outcomes because there was no group setting for control participants who did not complete the classes. However, it is possible that the group composition may have had an impact on the effectiveness of the intervention …" (p. 1046). Other authors have similarly echoed frustration at the gap between the complex experimental designs utilized in practice and the simplified experimental designs presented as examples in the methodological literature. For instance, Livert, Rindskopf, Saxe, and Stirratt (2001) remarked, "Although multilevel models are increasingly being utilized … actual application of such models to program assessment is complex and there are few examples (p. 155)."

In response, this article has four primary goals. Our first goal is to better explicate the logic of the mixed-effects (multilevel) modeling approach of Roberts and Roberts (2005) and Lee and Thompson (2005) for partially nested data. To do so, we juxtapose this approach with models that are currently being applied to partially nested data but that assume a parallel structure in both the treatment and control arms (either nonnested or fully nested). This allows us to clarify why models originally developed for nonnested or fully nested data are nonoptimal for partially nested data. Because both Roberts and Roberts and Lee and Thompson considered relatively simple analysis scenarios, our second goal is to extend this modeling approach to accommodate some of the more complex partially nested study designs that commonly occur in practice, such as those that include covariates at the individual *and* group levels, multiple treatment arms, and discrete outcome variables. In particular, the incorporation of group-level covariates into models for partially nested data has not been discussed elsewhere, despite its importance for elucidating sources of variability in treatment effects. Our third goal is to demonstrate the application of these models to partially nested data arising from an effectiveness study of the Reconnecting Youth program for adolescents at risk of school failure and behavioral disorders. Finally, our last goal is to provide a general discussion of the issue of causal inference in partially nested intervention studies, a topic that has heretofore been neglected in other papers concerned with this type of study design. We now address each of these goals in turn.

APPROACHES TO THE ANALYSIS OF PARTIALLY NESTED DATA

In this section we first discuss the limitations of the two ways that partially nested data are currently being analyzed, followed by an introduction to the basic model of Roberts and Roberts (2005) and Lee and Thompson (2005). We specifically define partially nested data to have the following structure: One subset of the data exhibits a hierarchical structure such that individuals are clustered into groups, whereas another subset of the data consists of independent individuals (with no clustering structure). Of interest is the particular case in which participants in the treatment arm of a study are placed into groups by the experimenter, but participants in the control arm of the study are not. To clarify the assumptions and limitations of the analysis approaches, it is necessary to present exemplar models in equation form. Very simple models, involving a single grouping variable (treatment vs. no treatment) and a single posttest outcome, will suffice for these purposes, but in the sections to follow we consider more complex (and more realistic) analysis scenarios. In the equations presented here and throughout the remainder of the article, we use notation consistent with Raudenbush & Bryk (2002), denoting individuals with the subscript *i* and denoting groups with the subscript *j*. For the subset of individuals who are not grouped, each individual comprises their own "group" of one.

Approach 1: Pretend No Observations Are Grouped

The first approach to analyzing partially nested data that we consider is a standard single-level regression model. A simple example of such a model might be

$$Y_{ij}=\beta_0+\beta_1 TREAT_{ij}+r_{ij} \tag{1}$$

where *Y* is the outcome variable at posttest; *TREAT* is an indicator variable scored 0 for members of the control arm and 1 for members of the treatment arm; β_0 is the regression intercept, interpretable as the mean of *Y* in the control arm; and β_1 is the regression slope, interpretable as the expected difference in *Y* associated with being a member of the treatment group (relative to the control group). The final term in the equation, the residuals *r*, are assumed to be independent and normally distributed with constant variance,[2] or

$$r_{ij} \sim N(0, \sigma^2).$$

This implies that the (conditional) variance of *Y* in both the treatment and control arms of the study is the same, with a value of σ^2. In particular, the independence assumption is highly problematic because the *Y* values for individuals within treatment groups will likely be positively correlated. Incorrectly assuming independence for the data will then lead to higher than nominal Type I error rates for tests of parameter estimates (e.g., treatment effects), increasing the risk of identifying spurious effects. Conversely, in the presumably less common situation that the observations are *negatively* correlated within groups (i.e., group members differentiate from one another), the Type I error rate will instead be too low, depressing the power to detect a true effect.[3]

[2] Note that under these assumptions this model is equivalent to a standard two-sample *t* test. Formulating this test within the general linear model will, however, facilitate the expression of later models.

Approach 2: Pretend All Observations Are Grouped

A second approach to analyzing partially nested data is to specify a multilevel model for the data. Multilevel models differ from traditional regression models in that they explicitly include sources of variability at both the individual and group level. Traditionally, model equations are written for each level of the data structure (i.e., individuals and then groups). Recall that, for modeling purposes, each participant in the control arm is viewed as being a member of their own "group" of one.

At Level 1 (the individual level), we specify a model similar to the one presented previously in Equation (1):[4]

$$Y_{ij}=\beta_{0j}+\beta_{1j}TREAT_{ij}+r_{ij}. \tag{2}$$

There is, however, one key difference between Equations (1) and (2): the regression intercept and slope have now been subscripted by *j*. This indicates that the values of these coefficients potentially differ across groups. We now express the potential variability in these coefficients across groups with the Level 2 (group-level) model:

$$\beta_{0j}=\gamma_{00}+u_{0j} \tag{3}$$

$$\beta_{1j}=\gamma_{10}. \tag{4}$$

Equation (3) indicates that the intercept of Equation (2) varies across groups. The average intercept is γ_{00} and the term u_{0j} indicates the extent to which group *j* deviates from this average. As we will see, it is the random variability in u_{0j} that will ultimately account for the correlation of observations within clusters. In contrast, Equation (4) includes no random deviation term, indicating that the treatment effect in this model is assumed to be constant (fixed) over treatment groups with a value of γ_{10}.

The grouping effect can also be seen in the combined model equation for the outcome, obtained by substituting Equations (3) and (4) into Equation (2):

$$Y_{ij}=\gamma_{00}+\gamma_{10}TREAT_{ij}+u_{0j}+r_{ij}. \tag{5}$$

The combined equation clarifies that this model posits two sources of unexplained variability. There are the usual individual level residuals (r_{ij}), but there is also a second disturbance due to groups (u_{0j}). This second disturbance implies that the individuals in some groups have generally higher or lower values for *Y* than in other groups.

As in the standard regression model, we must make assumptions about the nature of the unexplained variability in the model. Customarily, we assume that, within groups, the

[3] The models we discuss throughout are designed to account for positive correlations among group members. Multilevel models for observations that are negatively correlated are discussed in Kenny, Mannetti, Pierro, Livi, & Kashy (2002).

[4] Because treatment is assigned at the individual level, we treat this as an individual-level predictor, despite the fact that the value of this predictor is constant for all individuals within a particular group. This deviates from cluster-randomized designs in which treatment is assigned at the cluster level and treated as a cluster-level predictor.

individual residuals are normally distributed with constant variance over individuals (regardless of arm), or

$$r_{ij} \sim N(0, \sigma^2). \tag{6}$$

Similarly, the group-level disturbance is assumed to be normally distributed with constant variance over groups (regardless of arm), or

$$u_{0j} \sim N(0, \tau_{00}). \tag{7}$$

The variances of these two types of residuals are sometimes referred to as the variance components of the model. Finally, we assume that the two sources of unexplained variability are independent, that is, that there is no correlation between r_{ij} and u_{0j}.

Based on these assumptions, we can express the within-arm variance in Y as

$$V(Y|TREAT)=\tau_{00}+\sigma^2. \tag{8}$$

Further, the correlation in Y between any two members of the same group (intracluster correlation or ICC) can be expressed as

$$\mathrm{ICC}=\frac{\tau_{00}}{\tau_{00}+\sigma^2}. \tag{9}$$

The ICC thus captures the degree of similarity of participants who are members of the same group within the treatment arm. For controls, the ICC is irrelevant because each control participant is the sole member of their group.

Overall, this analysis approach appears to accomplish what we would like: the group-level variance component (τ_{00}) allows us to model the dependence of the observations within groups in the treatment arm, but the model continues to allow for independence in the control arm of the study because each control participant is a member of their own group. However, the model is inconsistent with the design of the study in an important way: the variance of Y within the control arm is decomposed in precisely the same way as the variance of Y in the treatment arm. Specifically, Equation (8) implies that

$$V(Y|TREAT=0)=\tau_{00}+\sigma^2 \tag{10}$$

$$V(Y|TREAT=1)=\tau_{00}+\sigma^2. \tag{11}$$

Equation (11) is sensible: variation in Y within the treatment arm is partly due to differences between treatment groups and partly due to differences among individuals within treatment groups. Equation (10), on the other hand, is not sensible—the same decomposition of variance does not apply because there is no grouping structure for participants in the control arm of the study.[5] As it turns out, this inconsistency will not matter much for testing the fixed effects (e.g., the test of the overall treatment effect) in the special circumstance that the variance in *Y* within the control and treatment arms is identical (as implied by Equations (10) and (11)). However, there is often reason to believe that this variance will differ between the treatment and control arms. If such heteroscedasticity is present, Roberts & Roberts (2005) have shown that this model will then generate biased tests of the treatment effect. What is needed is a more flexible approach that is more consistent with the study design.

Approach 3: Explicitly Model Partial Nesting Design

As an alternative to the two aforementioned approaches, Roberts and Roberts (2005) and Lee and Thompson (2005) provide a third approach to the analysis of partially nested data that better matches the data structure. Aside from being more consistent with the design that produced the data, this approach also provides model estimates that have appealing interpretations. Again, for purposes of model specification, participants within the control arm are viewed as the sole members of their own "groups."

The Level 1 model for this approach is identical to the preceding case:

$$Y_{ij}=\beta_{0j}+\beta_{1j}TREAT_{ij}+r_{ij}. \qquad (12)$$

The difference between Approach 2 and Approach 3 arises in the specification of the Level 2 model equations. Here we specify that the *slope* of this equation varies over groups and that the *intercept* term is constant over groups:

$$\beta_{0j}=\gamma_{00} \qquad (13)$$

$$\beta_{1j}=\gamma_{10}+u_{1j}. \qquad (14)$$

The combined model for the data, obtained by substituting Equations (13) and (14) into Equation (12), is now

$$Y_{ij}=\gamma_{00}+\gamma_{10}TREAT_{ij}+u_{1j}TREAT_{ij}+r_{ij}. \qquad (15)$$

The motivation for this model specification follows from several observations. The β_{0j} term of Equation (12) represents the group mean for participants in control group *j* (i.e., where *TREAT* = 0). However, in the control arm there is only one participant per group, so the group "mean" and the individual observation are identical. That is, we cannot separate group-level variation from individual-level variation, as these are one and the same. Because there is no need for a group-level residual for the control participants, no random component for β_{0j} is included in Equation (13). As such, in the previously mentioned model, variability

[5] This decomposition could still make sense even for groups with one member under certain circumstances where preexisting intact groups are used rather than groups formed during the treatment study. Consider, for instance, the case of data on siblings nested within families. In this case, the two variance components would correspond to variance due to unexplained child influences on *Y* and variance due to unexplained family influences on *Y*. Clearly family influences operate even on only children (though they could not be separately estimated from child influences unless multiple sibling families were also included in the analysis).

in the outcome variable for the control participants is sensibly decomposed into the mean of all participants in the control arm, captured by γ_{00}, and an individual residual $r_{ij,}$ as shown here:

$$Y_{ij}|(TREAT_{ij}=0)=\gamma_{00}+r_{ij}. \tag{16}$$

The β_{1j} parameter of Equation (12) reflects the difference between the mean for group j within the treatment arm and the overall mean for the control arm. As expressed in Equation (14), the present model permits the treatment group mean to vary across groups within the treatment arm through the inclusion of the term u_{1j}. Such differences reflect differential treatment outcomes across groups, due to the particular composition of the group, the fidelity of implementation of the treatment protocol, the effectiveness of the treatment administrator for the group, or other factors. Thus, for treated participants (when $TREAT$ = 1), Equation (15) can be rewritten as

$$Y_{ij}|(TREAT_{ij}=1)=\gamma_{00}+\gamma_{10}+u_{1j}+r_{ij}, \tag{17}$$

so that there are both individual- and group-level residuals. Comparing equations (16) and (17) we see that the average treatment effect across groups within the treatment arm is represented by the γ_{10} parameter, whereas differences across groups in the treatment effect are captured by the γ_{1j} term.

As in Approach 2, we assume that the individual- and group-level residuals are independent and normally distributed as

$$r_{ij} \sim N(0,\sigma^2) \tag{18}$$

$$u_{1j} \sim N(0,\tau_{11}). \tag{19}$$

Unlike Approach 2, however, the model-implied variance of Y now differs across the arms of the study. That is, the model explicitly accounts for potential heteroscedasticity across the two arms of the study. For control participants, this variance is simply

$$V(Y|TREAT=0)=\sigma^2, \tag{20}$$

whereas in the treatment arm it is

$$V(Y|TREAT=1)=\tau_{11}+\sigma^2. \tag{21}$$

Further, we can express the ICC within the treatment arm as

$$\text{ICC}_{TREAT}=\frac{\tau_{11}}{\tau_{11}+\sigma^2}. \qquad (22)$$

There is no corresponding ICC for the control arm because there is no between-group variance estimated for that arm. Note that the ratio of the variance in the control arm to the variance in the treatment arm is

$$\frac{\sigma^2}{\tau_{11}+\sigma^2}=1-\text{ICC}_{TREAT}. \qquad (23)$$

Thus the ICC in Equation (22) provides a measure of heteroscedasticity across the two study arms. Equation (23) implies that the degree of heteroscedasticity between the treatment and control arms is a direct function of the ICC in the treatment arm.

In comparing Equations (20) and (21) we see that the model implies that this heteroscedasticity takes a special form, that there is added variance in *Y* due to group influences. As such, in contrast to Approaches 1 and 2, which assumed equal variance for *Y* for the treatment and control arms, the model formulated earlier assumes that the variance in the treatment arm exceeds the variance in the control arm (so long as $\tau_{11} \neq 0$). This may not always be the case: Group processes may actually increase the similarity of group members to one another, thereby decreasing within-group differences. Given this possibility, Roberts & Roberts (2005) noted that we can (and in many cases *should*) allow the variance of r_{ij} to differ across arms of the study by modifying the assumption in Equation (18) to instead be

$$r_{ij}|(TREAT=0) \sim N(0,\sigma^2_{Control}) \qquad (24)$$

$$r_{ij}|(TREAT=1) \sim N(0,\sigma^2_{Treatment}). \qquad (25)$$

With this modification, we permit heteroscedasticity between the treatment and control arms but do not constrain the form of heteroscedasticity. Specifically, there need not be added variance within the treatment arm. Further, when $\sigma^2_{Treatment} \neq \sigma^2_{Control}$, the ratio of variances between the two arms will no longer obey the relationship in Equation (23). Not only does this modification make the model potentially more realistic, it also offers the exciting possibility to formally test the hypothesis that participants within a treatment group become similar to one another in their attitudes and behavior, in which case we should find that $\sigma^2_{Treatment} < \sigma^2_{Control}$.

To summarize, of the three approaches, only Approach 3 is fully consistent with a partially nested design. Approach 1 ignores the grouping structure in the treatment arm, potentially leading to inflated Type I errors and spurious treatment effects. Approach 2 assumes a parallel grouping structure in both the control and treatment arms; however, the implied decomposition of variance for the control participants is then nonsensical. Both Approach 1 and Approach 2 assume equal variance in the treatment and control conditions, which may often be unrealistic. In contrast, by taking explicit account of the partially nested study design, Approach 3 offers the appealing benefits that it can account for dependence within treatment groups, capture variability in treatment outcomes across treatment groups, model heteroscedasticity, and reveal whether individuals become more homogeneous in their attitudes and behavior as a function of treatment group membership.

TESTING TREATMENT EFFECTS (AND OTHER FIXED EFFECTS)

Using the model outlined earlier in Approach 3, we are primarily interested in testing the significance of the estimate for γ_{10}, or the average treatment effect. However, we encounter two difficulties in doing so. First, obtaining an unbiased standard error for the effect estimate is not straightforward. Second, the reference distribution for testing the ratio of the estimate to its standard error is unknown. These difficulties are not unique to models for partially nested data, occurring for any mixed-effect (or multilevel) model with a complex covariance structure and/or unbalanced group sizes. We explain each obstacle and then discuss a combined corrective that handles both.

For contrast, let us first consider an ideal situation. If the population values of the variance components of the model (in this case τ_{11} and σ^2, or $\sigma^2_{Treatment}$ and $\sigma^2_{Control}$) were known, then the variance-covariance matrix of the fixed effects estimates, designated $\mathbf{\Sigma}_{\hat{\gamma}}$, could be calculated directly from these values. The square root of the diagonal of this matrix would provide standard errors for the fixed-effect estimates and the ratio of each estimate to its standard error would follow a standard normal distribution, permitting *z* tests of the estimates.

In practice, however, the population values of the variance components are unknown. Accordingly, the sample estimates of these variance components must be used in place of their population values to form an estimate of $\mathbf{\Sigma}_{\hat{\gamma}}$, designated $\hat{\mathbf{\Sigma}}_{\hat{\gamma}}$. The standard errors for the fixed effects are then typically calculated as the square root of the diagonal of $\hat{\mathbf{\Sigma}}_{\hat{\gamma}}$. Unfortunately, these standard errors are negatively biased, presenting our first difficulty for testing the fixed effects of the model (Dempster, Rubin, & Tsutakawa, 1981). One source of bias arises because the variance component estimates are subject to their own sampling variability, and treating them as known fails to account for the imprecision of these estimates. A second source of bias is that, in small samples, $\hat{\mathbf{\Sigma}}_{\hat{\gamma}}$ is a biased estimator of $\hat{\mathbf{\Sigma}}_{\hat{\gamma}}$. Corrections for these two sources of bias were developed in a series of influential papers by Kacker & Harville (1984), Harville and Jeske (1992), and Kenward and Rogers (1997) and have been implemented in some software programs capable of fitting multilevel models.

The second difficulty we face in evaluating the overall treatment effect in partially nested designs is that the reference distribution for the fixed effect estimates is unknown. This difficulty does not arise for designs in which all subjects are independent and there are no grouping effects (for which Approach 1 would be adequate). In that case, only one variance parameter (σ^2)is estimated and the reference distribution for testing the fixed effects is an exact *t* distribution with known degrees of freedom. A few other special cases also exist where exact tests can be obtained (Elston, 1998, p. 1086; Maxwell & Delaney, 2004, p. 479). Unfortunately, for models with unbalanced group sizes and/or complex covariance structures, like the model considered in Approach 3, exact tests for fixed effects cannot be obtained. In practice, it is assumed that the reference distribution can be approximated by a *t* distribution. Kenward and Rogers (1997) suggest estimating the degrees of freedom for the *t* distribution by a method-of-moments approach with origins in the work of Satterthwaite (1941; see also Schaalje, McBride, & Fellingham, 2002, pp. 515–517).

Overall, the Kenward-Rogers (1997) method for testing the fixed effects entails combining the bias correction for the standard errors (for handling the first difficulty) and the Satterthwaite (1941) method for computing degrees of freedom (for handling the second difficulty). In simulations comparing different methods for testing fixed effects in mixed models, Schaalje et al. (2002) found that the Kenward-Rogers method performed better than several competing methods, particularly for complex covariance structures and unbalanced designs. Based on these results, we recommend using the Kenward-Rogers method for testing treatment effects in partially nested study designs.[6] This issue is applicable also to the testing of other fixed effects in more complex models for partially nested data. It is to these extended models that we now turn.

EXTENSIONS OF THE MODEL

In this section we consider how Approach 3 can be extended to some of the more complex situations commonly encountered in evaluation research. In the subsections that follow we discuss how to include pretest measures in the model as covariates, incorporate other individual- and/or group-level covariates into the model, simultaneously evaluate multiple treatment or control arms, and test treatment effects on discrete outcome measures. Of these topics, Roberts and Roberts (2005) provide a short discussion on the use of individual-level covariates in the model (which could include pretest measures). Additionally, Lee and Thompson (2005) discuss binary outcomes and also briefly touch on the issues of multiple treatment arms and group-level covariates, though only in the context of a fully nested design. The other extensions of Approach 3 described here have not, to our knowledge, been presented previously. Each topic is addressed in a separate section, allowing the reader to skim sections of less interest.

Pretest as Covariate

One common approach for modeling treatment effects is to include pretest measures of the outcome as a control covariate in the statistical model, adjusting for preexisting differences among participants. Denoting the pretest measure as *X* and the posttest measure as *Y*, our Level 1 model will then be

$$Y_{ij}=\beta_{0j}+\beta_{1j}TREAT_{ij}+\beta_{2j}X_{ij}+r_{ij}, \tag{26}$$

and our Level 2 model will be

$$\beta_{0j}=\gamma_{00} \tag{27}$$

$$\beta_{1j}=\gamma_{10}+u_{1j} \tag{28}$$

$$\beta_{2j}=\gamma_{20}. \tag{29}$$

Note that the only new feature of this model is the coefficient associated with the pretest measure. Here we have assumed that the relation of the pretest measure to the posttest

[6] SAS code demonstrating this method with the demonstration data is available online at http://www.unc.edu/~dbauer. SPSS code providing Satterthwaite (1941) degrees of freedom, but not corrected standard errors, is also provided. The importance of correcting the standard errors is not known at this time, but failure to do so may result in a higher than nominal rate of Type I errors for tests of fixed effects.

measure is constant over treatment groups and equivalent across arms of the study, assumptions that seem reasonable for most applications. Under these assumptions, the combined model will be

$$Y_{ij}=\gamma_{00}+\gamma_{10}TREAT_{ij}+\gamma_{20}X_{ij}+u_{1j}TREAT_{ij}+r_{ij}. \tag{30}$$

Adjusting for pretest measures, the conditional variance in the posttest for the control arm continues to have only one component (individual-level variance only), whereas the conditional variance for the treatment arm again has two components (one for the individual-level variance and one for the group-level variance). These (now conditional) variances continue to be given by Equations (20) and (21).

It is also worth noting that, for the treatment arm, pretest measures may differ both across and within groups, as shown by decomposing X into a group mean and an individual deviation from the group mean:

$$X_{ij}=\overline{X}_{\cdot j}+\dot{X}_{ij}, \tag{31}$$

where $\bar{X}_{\cdot j}$ is the mean of the pretest values for treatment group j, and $\dot{X}_{ij}$ is the individual deviation from $\bar{X}_{\cdot j}$. Substituting Equation (31) into Equation (30) shows that, when we include pretest scores in the model, we are implicitly adjusting for both preexisting differences among individuals and preexisting differences among treatment groups:

$$Y_{ij}=\gamma_{00}+\gamma_{10}TREAT_{ij}+\gamma_{20}(\overline{X}_{\cdot j}+\dot{X}_{ij})+u_{1j}TREAT_{ij}+r_{ij}, \tag{32}$$

or

$$Y_{ij}=\gamma_{00}+\gamma_{10}TREAT_{ij}+\gamma_{20}\dot{X}_{ij}+\gamma_{20}\overline{X}_{\cdot j}+u_{1j}TREAT_{ij}+r_{ij}. \tag{33}$$

Note that there is an implicit equality constraint for within- and between-group effects of the pretest measure. Although often quite reasonable, it may at times be necessary to relax this assumption. Different within- and between-group effects might be expected if there is a compositional effect of the group above and beyond the individual effect of the covariate. For instance, if X represented a pretest measure of antisocial behavior, we might expect that treatment groups that happen to have higher than average baseline levels of antisocial behavior might be particularly difficult to manage and more resistant to treatment than groups that have lower than average levels of antisocial behavior. Fortunately, the assumption of equal within- and between-group effects can be removed quite easily by including the pretest means for the treatment groups ($\bar{X}_{\cdot j}$) in the model as an additional predictor (see Kreft, DeLeeuw, & Aiken, 1995, for further details on the decomposition of within- and between-group effects). We discuss the inclusion of group-level covariates in the next section.

Adding Individual- and Group-Level Covariates

Other covariates at the individual level can be included in the model in the same fashion as the pretest scores. Specifically, suppose X_1 is a pretest measure of the outcome and X_2 is a background characteristic of the individual (e.g., gender), then the Level 1 model can be modified accordingly to be

$$Y_{ij}=\beta_{0j}+\beta_{1j}TREAT_{ij}+\beta_{2j}X_{1ij}+\beta_{3j}X_{2ij}+r_{ij}. \tag{34}$$

Similarly, group-level covariates, such as treatment fidelity or perceived group cohesiveness, can be incorporated into the model at Level 2 to explain why some groups fare better than others in response to treatment (as called for by Weiss et al., 2005). There is, however, the additional consideration that such group-level variables pertain only to the treatment arm of the study and are, in essence, undefined for participants in the control arm. As an example, suppose that *W* is a variable that reflects some aspect of the group composition. Our Level 2 model for the coefficients in Equation (34) would then be

$$\beta_{0j}=\gamma_{00} \tag{35}$$

$$\beta_{1j}=\gamma_{10}+\gamma_{11}W_j+u_{1j} \tag{36}$$

$$\beta_{2j}=\gamma_{20} \tag{37}$$

$$\beta_{3j}=\gamma_{30}. \tag{38}$$

The combined model is then

$$Y_{ij}=\gamma_{00}+\gamma_{10}TREAT_{ij}+\gamma_{11}(TREAT_{ij}\times W_j)+\gamma_{20}X_{1ij}+\gamma_{30}X_{2ij}+u_{1j}TREAT_{ij}+r_{ij}. \tag{39}$$

The interaction term reflects the fact that the effect of *W* is necessarily conditional on treatment. The omission of a main effect of *W* in the combined model may appear unusual, but this is intentional given that *W* is undefined for control participants. Due to the absence of a main effect for *W*, for control participants the model reduces to the following:

$$Y_{ij}(TREAT=0)=\gamma_{00}+\gamma_{20}X_{1ij}+\gamma_{30}X_{2ij}+r_{ij}, \tag{40}$$

whereas for treated participants the model is

$$Y_{ij}|(TREAT=1)=\gamma_{00}+\gamma_{10}+\gamma_{11}W_j+\gamma_{20}X_{1ij}+\gamma_{30}X_{2ij}+u_{1j}+r_{ij}. \tag{41}$$

Thus, the group-level predictor *W* only affects *Y* for participants in the treatment arm and does not affect *Y* for participants in the control arm, who are ungrouped.

Practically speaking, to incorporate group-level predictors into the model, *W* must be set to an arbitrary nonmissing value for the control participants. If *W* is set to a missing value for the control participants, these participants will be deleted from the analysis, which is nonoptimal for obvious reasons. The value that is chosen for *W* for the control participants (e.g., −999, 2, 127) is, however, irrelevant as this variable will "zero out" of the prediction equation for participants in this arm of the study, as indicated in Equation (40).

Additionally, there may be some individual level predictors that are relevant only for grouped participants. For instance, one might wish to incorporate a measure of treatment exposure or "dosage" to evaluate whether this impacts individual outcomes. Similar to the approach described earlier for group-level covariates, the interaction between this predictor and the treatment indicator would then be included in the Level 1 equation, omitting the main effect. For ungrouped participants, the predictor would again need to be set to an arbitrary value to avoid the deletion of these participants from the analysis.

Multiple Treatment or Control Arms

It is of course also possible to evaluate multiple treatments and/or multiple control groups using the same basic approach indicated here. Specifically, indicator variables would be constructed for each arm of the study, and each of these indicator variables would be included as individual-level predictors. For any study arm that contains groups, the effect of the corresponding indicator variable would be permitted to vary randomly over groups.

Let us first consider a simple case in which one wishes to contrast the efficacy of a group-based treatment (Treatment 1) relative to an individually based treatment (Treatment 2). Additionally, there is a control arm in which participants are not grouped. Extending the simple two-arm model formulated previously in Equations (12) through (14), we can write the Level 1 model for this three-arm study design as follows:

$$Y_{ij}=\beta_{0j}+\beta_{1j}TREAT_{1ij}+\beta_{2j}TREAT_{2ij}+r_{ij}, \tag{42}$$

where $TREAT_1$ and $TREAT_2$ are indicator variables scored 1 if the individual participated in treatment 1 or 2, respectively, and are scored 0 otherwise.

At Level 2, we indicate that only the effect of the group-based intervention has a random component due to groups:

$$\beta_{0j}=\gamma_{00} \tag{43}$$

$$\beta_{1j}=\gamma_{10}+u_{1j} \tag{44}$$

$$\beta_{20}=\gamma_{20}. \tag{45}$$

Note that the individually based treatment does not include a grouping effect. As in the two-arm case, we can allow the variance of the Level 1 residuals (*r*) to vary across study arms to reflect differential effects of treatment on individual variability.

Now suppose that both treatments were group based, but the control arm consisted of ungrouped individuals. The Level 1 equation remains as in Equation (42), but the effect of the grouping structure is then expressed at Level 2 as follows:

$$\beta_{0j}=\gamma_{00} \tag{46}$$

$$\beta_{1j}=\gamma_{10}+u_{1j} \tag{47}$$

$$\beta_{2j}=\gamma_{20}+u_{2j}. \tag{48}$$

Note that both group-based treatments have effects that depend partly on the individuals' treatment group. Assuming that individuals in the two treatment arms are independent (e.g., randomly assigned to treatment groups), we can reasonably assume that the u_{1j} and u_{2j} terms are also independent. Following Lee and Thompson (2005), we thus express their distribution as

$$\begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\begin{pmatrix} \tau_{11} & 0 \\ 0 & \tau_{22} \end{pmatrix}\right]. \tag{49}$$

In practice, it is worth noting that many multilevel modeling software programs include a covariance between u_1 and u_2 by default, in which case this covariance must be manually set to zero.

Discrete Outcomes

For simplicity, we have thus far described each of the aforementioned models for a continuous outcome variable Y that could reasonably be assumed to have a conditional normal distribution. However, dichotomous, ordinal, and count outcomes are also quite common in intervention and prevention research. Fortunately, great strides have been made in the past decade in the estimation of multilevel generalized linear models, permitting the extension of all of the models presented earlier to discrete outcome variables. As a basic example, we show how the model in Equations (12) through (14) could be reformulated as a generalized linear model. Although the Level 2 equations remain unchanged, we must rewrite the Level 1 equation as

$$\eta_{ij}=\beta_{0j}+\beta_{1j}TREAT_{ij}, \tag{50}$$

where η_{ij} is referred to as the *linear predictor* and is related to the expected value (or conditional mean) of Y though the equation $\eta_{ij} = f(\mu_{ij})$, where f is known as the *link function.* The purpose of the link function is to map the possibly continuous values of the linear predictor onto the logical range of the expected value for the outcome. For continuous outcomes that have a broad range, an identity link is typically used: $\eta_{ij} = \mu_{ij}$. In contrast, for count outcomes, the expected value has a lower bound of zero, but no upper bound, and hence the log link is often used: $\eta_{ij} = \log(\mu_{ij})$. For dichotomous outcomes the value of the linear predictor must be mapped onto the [0,1] interval and hence the logit link is a natural choice: $\eta_{ij} = \log(\mu_{ij}/(1 - \mu_{ij}))$. Other common choices for dichotomous outcomes are the probit and complementary log-log link functions.

Another important feature of Equation (50) is that there is no Level 1 residual. This is because this equation only provides information on the *expected value* of Y and not the variation around this expected value. To account for this variation, we must also specify a conditional distribution for Y. For continuous outcomes, the conditional distribution is often specified as normal, such that $Y_{ij}|\mu_{ij} \sim N(\mu_{ij,}\ \sigma^2)$ In combination with the identity link, this specification gives rise to the models previously presented. For count outcomes, the conditional distribution might instead be specified as Poisson, or $Y_{ij}|\mu_{ij} \sim$ Poisson (μ_{ij}). Alternatively, for dichotomous outcomes, the conditional distribution would be specified as Bernoulli, or $Y_{ij}\ |\mu_{ij} \sim$ Bernoulli(μ_{ij}).

As for this most basic of models, appropriate choices for the link function and conditional distribution could be made for any of the models discussed previously for the case where Y is discretely distributed.

To summarize, the analysis of partially nested data can be extended beyond the relatively simple expository examples provided by Lee and Thompson (2005) and Roberts and Roberts (2005) to the more complex situations that are encountered in much intervention and prevention research. Although the model extensions delineated earlier are far from exhaustive, we believe they cover many of the possible analysis scenarios that investigators encounter when using partially nested designs. We now turn to a demonstrative application of these models.

AN EMPIRICAL DEMONSTRATION OF THE ANALYSIS OF PARTIALLY NESTED DATA

For our demonstration, we evaluate an effectiveness trial of the Reconnecting Youth (RY) preventive intervention program that employed a partially nested design. This RY trial involved a large sample of adolescents from five high schools in an urban school district in the Southwest and four high schools in an urban school district on the Pacific coast. The goals, methods, and recruitment procedures for the RY program have been described in Eggert, Thompson, Herting, Nicholas, and Dicker (1994), and prior analyses of the current sample may be found in Hallfors et al. (2006), Cho, Hallfors, and Sanchez (2005), and Sanchez et al. (2007). Briefly, an initial screened sample was stratified on risk status (based on criteria including highest 25% for truancy and bottom 50% for grade point average (GPA), or referred for treatment by a schoolteacher or counselor). High-risk children were oversampled (N = 1370) and low-risk children were randomly sampled (N = 598). High-risk children were *individually* randomly assigned to either the intervention arm (N = 695) or to the control arm (N = 675). Of those assigned to the intervention, 47% did not participate, usually because of scheduling issues or other external constraints (remaining N = 370). The participation rate varied across schools and participation was negatively related to age, unrelated to gender or ethnicity, and inconsistently related to baseline measures of deviancy.[7]

[7] Unlike many treatment studies, in this study noncompliance was typically dictated by third parties (e.g., guidance counselors) largely as a function of external constraints (e.g., classes needed for graduation) rather than being an active choice of the participant assigned to treatment. Compliance is thus unlikely to be related to differential motivation, openness to treatment, or treatment effectiveness. For this reason, and to simplify the presentation of the example, we chose to exclude noncompliers from the present analyses (i.e., using an “on treatment” approach). Parallel analyses using an “intent to treat” approach produced broadly similar results (not shown here).

High-risk participants in the intervention arm received RY treatment *administered in groups* (i.e., RY classes), whereas high-risk participants assigned to the control arm were left *ungrouped*. The number of RY classes in the intervention arm totaled 41, with 2 to 7 classes per school, 5 to 15 students per class, and an average class size of 9 students. Additionally, the low-risk children constituted a second *ungrouped* control arm—representing "typical" adolescents from their respective schools. Thus, this is a multiple-control-arm, partially nested design—features that we will now address using the analytic strategy advocated in the prior sections.

The RY trial is particularly suited for this demonstration because modeling RY's partial nesting is shown to shed important light on some earlier results. Specifically, in previous analyses of the data, Hallfors et al. (2006) found that the intervention program did *not* have the anticipated positive effects. Instead, it appeared to *exacerbate* some problem behaviors, particularly when measured one semester posttreatment. Here, we reevaluate the negative effect of treatment on one outcome variable, deviant peer bonding (DPB). DPB was measured as the average of eight 5-point items (e.g., How many of your close friends skip school, drink alcohol, have gotten into physical fights with other kids, etc.). With the methods proposed in this article, we are able to estimate a series of models that systematically examine whether, concomitant with the negative mean shift in DPB for RY participants, there is (a) within-class homogenization of DPB suggestive of an iatrogenic "contagion" effect, (b) there are differences across RY classes in treatment outcomes suggestive of iatrogenic compositional or provider effects, and (c) whether these effects persist when controlling for preexisting differences in DPB and demographic characteristics. Without the methods proposed in this article, these hypotheses could not be precisely examined.

In total, we estimated a sequence of three models. Our goals in fitting the first model were (a) to evaluate whether mean DPB differed across the three conditions at one semester posttreatment, with particular attention to the effect of RY relative to control; (b) to ascertain whether there was variability across RY classes in the magnitude of the treatment effect; and (c) to determine the intraclass correlation among RY students attending the same class. To satisfy these goals, we also needed to control for the effects of school-based clustering. Because only nine schools were included in the study and these schools were nonrandomly selected based on urbanicity and demographic composition, we chose to model school as a fixed factor. Although this is consistent with the study design and controls for school-based clustering, we must limit our inference space to these schools (i.e., we cannot speculate as to what the effects might be in other schools). With a larger, random sample of schools it would be possible to move to a three-level model with a random effect of school, and this would permit inferences to the larger population of schools from which the sample was drawn.[8]

For our first model, we thus included the treatment condition as the primary predictor of DPB and school as a control factor. Because only the RY condition was grouped, the Level 1 and 2 equations are similar to those shown in Equations (42) to (45), with the addition of eight dummy variables to represent the nine-level school factor. The Level 1 Equation is thus

[8] In response to reviewer concerns, we also ran these models with a random effect of school rather than including school as a fixed factor. The results were highly similar to those presented here.

$$DPB_{ij}=\beta_{0j}+\beta_{1j}RY_{ij}+\beta_{2j}Typical_{ij}+\sum_{c=1}^{8}\beta_{(2+c)j}School(c)_{ij}+r_{ij}, \tag{51}$$

where $School(c)_{ij}$ is a dummy variable indicating whether the participant i is a student in school c (coded 1) or not (coded 0). In turn, the Level 2 Equations are

$$\begin{gathered}\beta_{0j}=\gamma_{00}\\ \beta_{1j}=\gamma_{10}+u_{1j}\\ \beta_{2j}=\gamma_{20}\\ \beta_{(2+c)j}=\gamma_{(2+c)0}\end{gathered} \tag{52}$$

The combined model equation is then

$$DPB_{ij}=\gamma_{00}+\gamma_{10}RY_{ij}+\gamma_{20}Typical_{ij}+\sum_{c=1}^{8}\gamma_{(2+c)0}School(c)_{ij}+u_{1j}RY_{ij}+r_{ij}. \tag{53}$$

We begin by assuming homoscedasticity for the residual variance r, then go on to test this assumption by allowing the residual variance to differ between the RY study arm and the other two arms. This model and all of those to follow were estimated using the MIXED procedure in SAS with the REML estimator and the Kenward-Rogers (1997) method of testing the fixed effects (both SAS and SPSS code and annotated output for this analysis are provided in online supplementary material posted at http://www.unc.edu/~dbauer).

The results from the homoscedastic model, reported in the first column of Table 1, indicated that the three conditions *did* differ significantly from one another in levels of DPB, with RY students displaying higher DPB than controls and controls displaying higher DPB than typical students. A significant school effect was also detected, indicating that DPB levels vary across schools. Planned contrasts revealed that the school effect was due almost entirely to site differences. Controlling for the school effects, the variance component for the RY classes, .053, was relatively small in magnitude. In relation to the residual variance, .79, this value yields an intraclass correlation of .06, indicating that the DPB scores of students attending the same RY class were positively correlated (as expected) but not highly. When the residual variance was permitted to differ between the RY arm and the other two study arms, the obtained estimates were .81 and .78, respectively. The direction of this difference is counter to the hypothesis of within-group homogenization due to iatrogenic effects in the RY arm, and the difference itself was not statistically significant by the likelihood ratio test ($\chi^2(1)= .20$, $p = .65$). In total, there is little evidence of “contagion” effects for DPB and only modest evidence of compositional or provider effects.

The second model that we fit extended the model in Equation (53) to include several additional Level 1 covariates measured in all three groups. Specifically, we controlled for baseline DPB (measured pretreatment), age, sex, and ethnicity. As shown in the second column of Table 1, the differences among the three experimental conditions were diminished but still statistically significant following the inclusion of the control variables. Of the control variables, baseline DPB and ethnicity were statistically significant. The ethnicity effect was largely due to Asian students having lower DPB scores. Notably, after the inclusion of these covariates, the variance component for the RY classes dropped to .01

and the residual intraclass correlation dropped to .02. In other words, much of the variability in the effects of RY on DPB could be explained on the basis of preexisting individual factors that were unevenly distributed across RY classes.

Finally, in Model 3, we extended Model 2 by including several covariates relevant *only* to the RY condition (as discussed in the “adding individual- and group-level covariates” section). This allowed us to evaluate whether characteristics of the individual or classroom *moderated* the effect of treatment on participants. First, we considered whether students attending more RY classes (absent less) showed greater negative effects (i.e., a dosage effect). Second, we assessed whether two aspects of class composition moderated treatment effects: the average age of the students within the class and the percentage of female students in the class. Because these three variables were only definable for students assigned to the RY condition, they were set to –999 for students in the control and typical conditions (i.e., following the strategy outlined previously). To test the effects of these predictors, three interaction terms were then added to the model, RY × Absences, RY × Mean Age, and RY × Percentage Female. The results, shown in the third column of Table 1, indicated that students who were absent from more classes had slightly higher levels of DPB, as did students attending classes composed of mostly younger, male students; however, none of these effects was statistically significant. These null findings were consistent with the results of Model 2, which indicated that there were few differences among RY classes left to be explained following the inclusion of the control covariates.

To summarize, these analyses provide one demonstration that data obtained from a partially nested design can be appropriately analyzed using specifically tailored multilevel models. To match the complexity of real evaluation research, we have shown how pretest measures, “common” covariates, and covariates relevant only for the grouped condition can be included in the model. Many other extensions of the analytic model could also be contemplated, such as the inclusion of multiple pre- or posttest measures. At this point, however, we turn to a more fundamental epistemological question that has, to date, gone unaddressed: To what extent can valid inferences about treatment effects be made using data obtained from a partially nested study design?

ASSESSING TREATMENT EFFECTS WITH PARTIALLY NESTED DATA

A fundamental difficulty with partially nested study designs is that the structure of the data is not parallel between the treatment and control arms, and thus treatment effects may be conflated with grouping effects. Specifically, the act of placing participants into groups may have either positive or negative effects, and if only treated participants are grouped then treatment effects cannot be disentangled from these grouping effects. At its core, this is an issue of internal validity: Does a partially nested study design allow for strong tests of the theoretical model guiding treatment? Unfortunately, the answer must be “No,” given that one cannot ascribe effects, either positive or negative, uniquely to the theoretically motivated aspects of treatment. That is, these effects may be a consequence simply of grouping participants together, regardless of treatment. Thus, we *cannot* say with certainty that the negative effects observed in the RY effectiveness trial were due to an inadequacy of the theoretical model that motivated the intervention. It may simply be that the intended positive effects of the manipulation were overwhelmed by the negative effects of grouping high-risk youth together. Put simply, with partially nested designs, it is not possible to make definitive statements regarding the causal effects of the intervention.

This is not to say that partially nested study designs are without merit. Although the internal validity of the partially nested study design is not as strong as would be afforded by a design with parallel structure for treatment and control arms, in many instances the external validity is stronger. For instance, to assess the public health benefits of the RY program, the ungrouped controls provide the most appropriate contrast condition, as they reflect the true condition of high-risk students when no intervention is implemented. Constructing untreated groups of high-risk adolescents for comparison to the treatment groups might be more internally valid for theory evaluation but would not provide a good indication of what could be gained (or lost) by implementing this intervention in like schools. Additionally, for many studies, there would be ethical limitations to the construction of groups within the control arm. For instance, it would probably be unwise to intentionally congregate high-risk youth into groups without providing some sort of directed intervention or treatment. Thus, partially nested study designs have a clear place in prevention and intervention research and in the investigation of treatment effects more broadly. The limitations of these designs for evaluating theoretical models of treatment effects have, however, gone largely unappreciated.

CONCLUSIONS

Partially nested study designs are a common and necessary presence in prevention and intervention research. Until recently, analytic methods for properly evaluating treatment effects with these kinds of designs have been unavailable. By extending three recent publications on this topic (Hoover, 2002; Lee & Thompson, 2005; Roberts & Roberts, 2005), this article explicated a general approach to the analysis of data with a partially nested structure using multilevel (or mixed-effects) models. This approach not only provides estimates with appealing interpretations, it is also amenable to the inclusion of covariates and predictors at both the individual and group levels. We demonstrated the features analytically and then empirically with RY trial data. Although this approach solves many of the data-analytic challenges associated with partially nested designs, it does not resolve the basic question of whether and when valid inferences about treatment effects can be made with these designs. In our view, these designs often increase external validity at the expense of internal validity, due to the conflation of treatment and grouping effects. Investigators should be aware of this trade-off when selecting a partially nested design for their research and take appropriate precautions in interpreting their findings.

Acknowledgments

This research was supported by the National Institutes on Drug Abuse grant R01 DA13666, awarded to Denise Hallfors.

References

Baldwin SA, Murray DM, Shadish WR. Empirically supported treatments or Type I errors? Problems with the analysis of data from group-administered treatments. Journal of Consulting and Clinical Psychology 2005;73:924–935. [PubMed: 16287392]

Bland JM. Cluster randomized trials in the medical literature: Two bibliometric surveys. BMC Medical Research Methodology 2004;4:1–6.

Carey M, Cary K, Maisto S, Gordon C, Schroder K, Vanable P. Reducing HIV-risk behavior among adults receiving outpatient psychiatric treatment: Results from a randomized controlled trial. Journal of Consulting and Clinical Psychology 2004;72:252–268. [PubMed: 15065959]

Cho H, Hallfors DD, Sanchez V. Evaluation of a high school peer group intervention for at-risk youth. Journal of Abnormal Child Psychology 2005;33(3):363–374. [PubMed: 15957563]

Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. Journal of Consulting and Clinical Psychology 2001;59:20–26. [PubMed: 2002139]

Crits-Christoph P, Tu X, Gallop R. Therapists as fixed versus random effects—Some statistical and conceptual issues: A comment on Siemer and Joormann (2003). Psychological Methods 2003;8:518–523. [PubMed: 14664686]

Dempster AP, Rubin DB, Tsutakawa RK. Estimation in covariance component models. Journal of the American Statistical Association 1981;76:341–353.

Eggert LL, Thompson EA, Herting JR, Nicholas LJ, Dicker BG. Preventing adolescent drug abuse and high school dropout through an intensive school-based social network development program. American Journal of Health Promotion 1994;8(3):202–215. [PubMed: 10172017]

Elston DA. Estimation of denominator degrees of freedom of F-distributions for assessing wald statistics for fixed-effect factors in unbalanced mixed models. Biometrics 1998;54:1085–1096.

Fromme K, Corbin W. Prevention of heavy drinking and associated negative consequences among mandated and voluntary college students. Journal of Consulting and Clinical Psychology 2004;72:1038–1049. [PubMed: 15612850]

Hallfors D, Cho H, Sanchez V, Khatapoush S, Kim H, Bauer D. Efficacy vs effectiveness trial results of an indicated “model” substance abuse program: Implications for public health. American Journal of Public Health 2006;96:2254–2259. [PubMed: 16809591]

Harville DA, Jeske DR. Mean squared error of estimation or prediction under a general linear model. Journal of the American Statistical Association 1992;87:724–731.

Hoover DR. Clinical trials of behavioural interventions with heterogeneous teaching subgroup effects. Statistics in Medicine 2002;21:1351–1364. [PubMed: 12185889]

Janega JB, Murray DM, Varnell SP, Blitstein JL, Birnbaum AS, Lylte LA. Assessing the most powerful analysis method for school-based intervention studies with alcohol, tobacco, and other drug outcomes. Addictive Behaviors 2004;29:595–606. [PubMed: 15050677]

Kacker RN, Harville DA. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. Journal of the American Statistical Association 1984;79:853–862.

Kenny DA, Mannetti L, Pierro A, Livi S, Kashy DA. The statistical analysis of data from small groups. Journal of Personality and Social Psychology 2002;83:126–137. [PubMed: 12088122]

Kenward MG, Rogers JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 1997;53:983–997. [PubMed: 9333350]

Kreft IGG, DeLeeuw J, Aiken LS. Variable centering in hierarchical linear models: Model parameterization, estimation, and interpretation. Multivariate Behavioral Research 1995;30:1–21.

Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. Clinical Trials 2005;2:163–173. [PubMed: 16279138]

Livert D, Rindskopf D, Saxe L, Stirratt M. Using multilevel modeling in the evaluation of community-based treatment programs. Multivariate Behavior Research 2001;36:155–183.

Maxwell, SE.; Delaney, HD. Designing experiments and analyzing data: A model comparison perspective. 2. Mahwah, NJ: Erlbaum; 2004.

Murray, DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.

Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Evaluation Review 2003;27:79–103. [PubMed: 12568061]

Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: A review of recent methodological developments. American Journal of Public Health 2004;94:423–432. [PubMed: 14998806]

Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. Applied linear statistical models. 4. Chicago: Irwin Press; 1996.

Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods 1997;2:173–185.

Raudenbush, S.; Bryk, A. Hierarchial linear models: Applications and data analysis methods. Newbury Park, CA: Sage; 2002.

Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials 2005;2:152–162. [PubMed: 16279137]

Sanchez V, Steckler A, Nitirat P, Hallfors D, Cho H, Brodish P. Fidelity of implementation in a treatment effectiveness trial of Reconnecting Youth. Health Education and Research 2007;22:95–107.

SAS Institute Inc.. SAS 9.1.3 Help and documentation. Cary, NC: Author; 2004.

Satterthwaite FE. Synthesis of variance. Psychometrika 1941;6:309–316.

Schaalje GB, McBride JB, Fellingham GW. Adequacy of approximations to distributions of test statistics in complex linear models. Journal of Agricultural, Biological, and Environmental Statistics 2002;7:512–524.

Serlin RC, Wampold BE, Levin JR. Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: A comment on Siemer and Joormann (2003). Psychological Methods 2003;8:524–534. [PubMed: 14664687]

Siemer M, Joormann J. Power and measures of effect size in analysis of variance with fixed versus random nested factors. Psychological Methods 2003a;8:497–517. [PubMed: 14664685]

Siemer M, Joormann J. Assumptions and consequences of treating providers in therapy studies as fixed versus random effects: Reply to Crits-Christoph, Tu, and Gallop (2003) and Serlin, Wampold, and Levin (2003). Psychological Methods 2003b;8:535–544. [PubMed: 14664688]

Wampold BE, Brown GS. Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. Journal of Consulting and Clinical Psychology 2005;73:914–923. [PubMed: 16287391]

Wampold BE, Serlin RC. The consequences of ignoring a nested factor on measures of effect size in analysis of variance. Psychological Methods 2000;5:425–433. [PubMed: 11194206]

Weiss B, Caron A, Ball S, Tapp J, Johnson M, Weisz JR. Iatrogenic effects of group treatment for antisocial youth. Journal of Consulting and Clinical Psychology 2005;73:1036–1044. [PubMed: 16392977]

TABLE 1

Estimates From Analysis of Iatrogenic Effects in Reconnecting Youth (RY) Effectiveness Trial on Posttest Measures of Deviant Peer Bonding

	Model 1		Model 2		Model 3	
Predictor	**Estimate**	**Test**	**Estimate**	**Test**	**Estimate**	**Test**
			Fixed Effects			
Condition		*F* (2, 132) = 39.48***		*F* (2, 149) = 8.48***		*F* (2, 123) = 8.39***
RY	.19	*t*(68.3) = 2.63*	.14	*t*(79.4) = 2.43*	.14	*t*(64.7) = 2.42*
Typical	−.37	*t*(1432) = −6.99***	−.11	*t*(1428) = −2.19*	−.11	*t*(1405) = −2.20*
School		*F* (8, 891) = 15.95***		*F* (8, 781) = 3.03**		*F* (8, 778) = 3.06**
Site		*F* (1, 936) = 80.75***		*F* (1, 1133) = 12.49***		*F* (1, 1113) = 12.62***
Baseline deviant peer bonding			.50	*t*(1444) = 21.31***	.49	*t*(1413) = 20.97***
Age			−.03	*t*(1450) = −1.40	−.03	*t*(1419) = −1.07
Male			.06	*t*(1455) = 1.57	.06	*t*(1409) = 1.35
Ethnicity				*F* (4, 1452) = 4.07*		*F* (4, 1425) = 3.88**
RY × Absences					.002	*t*(662) = .51
RY × Mean Age					−.14	*t*(33.4) = −1.02
RY × Percentage Female					−.001	*t*(34.7) = −.48
			Variance Components			
Level 1 (σ^2	.789	$z = 26.73^{***}$	.593	$z = 26.63^{***}$	.591	$z = 26.41^{***}$
Level 2, RY(τ_{11})	.053	$z = 1.51$	.014	$z = .71$	.012	$z = .56$

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.