



# Psychological Methods

**Manuscript version of**

## Multilevel Factorial Designs With Experiment-Induced Clustering

Inbal Nahum-Shani, John J. Dziak, Linda M. Collins

Funded by:

- National Institutes of Health

© 2017, American Psychological Association. This manuscript is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final version of record is available via its DOI: <https://dx.doi.org/10.1037/met0000128>

This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.



## Multilevel Factorial Designs with Experiment-Induced Clustering

### Abstract

Factorial experimental designs have many applications in the behavioral sciences. In the context of intervention development, factorial experiments play a critical role in building and optimizing high-quality, multi-component behavioral interventions. One challenge in implementing factorial experiments in the behavioral sciences is that individuals are often clustered in social or administrative units and may be more similar to each other than to individuals in other clusters. This means that data are dependent within clusters. Power planning resources are available for factorial experiments in which the multilevel structure of the data is due to individuals' membership in groups that existed before experimentation. However, in many cases clusters are generated in the course of the study itself. Such experiment-induced clustering (EIC) requires different data analysis models and power planning resources from those available for multilevel experimental designs in which clusters exist prior to experimentation. Despite the common occurrence of both experimental designs with EIC and factorial designs, a bridge has yet to be built between EIC and factorial designs. Therefore, resources are limited or nonexistent for planning factorial experiments that involve EIC. This article seeks to bridge this gap by extending prior models for EIC, developed for single-factor experiments, to factorial experiments involving various types of EIC. We also offer power formulas to help investigators decide whether a particular experimental design involving EIC is feasible. We demonstrate that factorial experiments can be powerful and feasible even with EIC. We discuss design considerations and directions for future research.

## Multilevel Factorial Designs with Experiment-Induced Clustering

### Introduction

Factorial experimental designs have many potential advantages for behavioral scientists. In the context of intervention development, factorial designs play a critical role in building and optimizing multi-component interventions based on empirical evidence. Multi-component interventions are interventions that include several aspects (i.e., components), pertaining to the intervention's content, type, methods of delivery, and/or implementation strategies (see Collins, Kugler, & Gwadz, 2016). Factorial experiments help investigators screen several candidate intervention components *simultaneously* and decide which are likely to offer a detectable benefit. With each factor corresponding to one intervention component, the purpose of screening experiments is to identify which intervention components have a substantial influence on the response variable and are, therefore, candidates for selection into a high-quality intervention package, which will be evaluated in a subsequent randomized controlled trial (RCT; Myers, Montgomery, & Anderson-Cook, 2016; Wu & Hamada, 2011). These screening experiments are an important step *before* confirming the efficacy/effectiveness of the intervention package as a whole via a RCT (Collins, Nahum-Shani, & Almirall, 2014; Collins et al., 2016).

One challenge in implementing factorial experiments in the behavioral sciences is that individuals are often clustered rather than independent: students are clustered in schools, employees are clustered in organizations, patients are clustered in clinics, and so on. These individuals may be more similar to each other than to individuals in other clusters, on average. Statistically, this means that their data may be dependent within clusters. When the experiment involves randomizing these pre-existing clusters to experimental conditions, statistical power is

affected not only by the standard considerations that are relevant in any experiment (i.e., effect size, chosen Type I error rate, and sample size ( $N$ )), but also by the number of clusters ( $J$ ), the number of individuals within each cluster ( $n$ ), and the intraclass correlation (ICC). The ICC reflects the degree of dependence in the response (i.e., outcome) among individuals within clusters. To the extent that ICC is large or  $J$  is small, statistical power for an experiment with clustered individuals is expected to be lower compared to an experiment with unclustered individuals. Thus, sample size planning becomes somewhat more complex when an investigator is considering a multilevel experiment.

Dziak, Nahum-Shani, and Collins (2012) examined the feasibility of conducting factorial experiments in a multilevel setting. They showed that in scenarios with a reasonable number of clusters, number of individuals within clusters, and ICC, it is often possible to conduct a factorial experiment with adequate power for addressing scientific questions of primary interest, even when the target population is multilevel. However, their work focused on multifactor experiments in which the multilevel structure of the data is due to clusters that exist prior to experimentation (e.g., schools, clinics, organizations). In such settings, data may be dependent within clusters both at pretest and at posttest assessments.

By contrast, in many cases the multilevel structure of the data is due to clusters that are generated in the course of the study itself, such that while individuals are independent at pretest, their data may be dependent within clusters at posttest. This can occur for reasons that are practical, scientific, or both. Practical reasons for inducing clusters typically include the availability of resources and/or the feasibility of intervention delivery. For example, intervention science experiments commonly include a staff of therapists, each of whom delivers the intervention to a subset of individuals (e.g., Cloitre, Koenen, Cohen, & Han, 2002). Hence,

individuals' outcomes may be correlated due to shared provider effects. At other times, there are scientific or therapeutic reasons to induce clusters. Interventions are sometimes designed to be delivered in group settings, in order to facilitate therapeutic group processes and capitalize on social reinforcers such as social support, sense of belonging, cohesiveness, and social accountability (Schulz, Cowan, & Cowan, 2006). In these cases, the outcome for treated individuals may be correlated due to common experiences, informal processes of socialization, and group dynamics.

Experimental designs in which clusters are generated as part of the study are known as "individually randomized group treatment (IRGT) trials" (Pals, Murray, Alfano, Shadish, & Hannan, 2008; Candel & Van Breukelen, 2009), or "clinical trials with clustering effects due to treatment" (Roberts & Roberts, 2005). In this article we use the broader term experiment-induced clustering (EIC) to refer generally to designs in which one or more of the experimental conditions involve generating dependence between the units of analysis (i.e., individuals).

Many experiments in the behavioral sciences involve EIC. In applied and social psychology, these experiments are often used to study group dynamics and intergroup relations; among these, factorial designs are highly prevalent, although usually with only two (e.g., Kramer, Fleming, & Mannis, 2001; Nye, 2002; Valacich, Wheeler, Mennecke, & Wachter, 1995) or three factors (e.g., Erez & Arad, 1986; Derlega, Winstead, Wong, & Hunter, 1985; Karakowsky & McBey, 2001). In the context of intervention development, many experiments involve generating dependence between individuals, via the assignment of individuals to receive treatment in groups or from providers. Although only a few of these experiments have involved multiple factors (e.g., Charlesworth et al., 2011; Kasari, Rotheram-Fuller, Locke, & Gulsrud, 2012; Nackers et al., 2015; Wilson et al., 2015), interest in factorial designs with EIC is growing

in the context of behavioral intervention development. This trend is facilitated by the increased use of factorial designs to screen multiple intervention components (e.g., Cook et al., 2016; Pellegrini, Hoffman, Collins, & Spring, 2014, 2015) and by technological advances that not only make the integration of group-based support (e.g., online group therapy; Chebli, Blaszczynski, & Gainsbury, 2016; Gainsbury & Blaszczynski, 2011) more feasible, but also enhance the experimenter's control over the combination of intervention components delivered (Crespi, 2016; Dallery, Riley & Nahum-Shani, 2015; Peters, de Bruin, & Crutzen, 2015).

Experimental designs involving EIC require data analysis models and power planning resources that differ from those available for multilevel experimental designs in which clusters exist prior to experimentation. Existing resources for planning experimental designs that involve EIC are limited to the standard two-arm (i.e., a single-factor case) RCTs (Candel & Van Breukelen, 2009; Moerbeek & Wong, 2008; Tokola, Larocque, Nevalainen, & Oja, 2011; Roberts & Roberts, 2005). Factorial designs with EIC have received little methodological attention, and their power properties have not specifically been explored. Given the common occurrence of experimental designs with EIC, as well as the growing interest in (e.g., Baker, Gustafson, & Shah, 2014; Czajkowski et al., 2015; Jacobs & Graham, 2016) and use of (e.g., Cook et al., 2016; Howard & Jacobs, 2016; Pellegrini et al., 2014; 2015) factorial designs to inform the construction of high-quality multi-component behavioral interventions, it is critical to close this gap and build a bridge between these two experimental design features.

### The Present Article

The present article has three purposes. The first is to extend prior models for EIC, developed for single-factor experiments, to factorial experiments involving various types of EIC. The second is to demonstrate that factorial experiments with EIC can be powerful and feasible.

The third is to offer power formulas and Monte Carlo simulation results to help investigators decide whether a particular experimental design involving EIC is feasible to implement in their situation. The present article considers only models with normally distributed responses. However, this work can serve as the basis for future extension to generalized models with other response types (e.g., binary, count).

We begin by briefly reviewing complete and fractional factorial designs. We then discuss two common situations in which there is EIC: full EIC, in which all experimental conditions involve the creation of clusters, and partial EIC, in which only a subset of the experimental conditions involves the creation of clusters. We discuss modeling for a single-factor experiment and then for factorial experiments. To simplify the discussion, we assume that all factors in an experiment are dichotomous, and that the outcome is continuous and measured at the individual level. Finally, we discuss power-planning resources for factorial designs with EIC, using simulation studies to evaluate these resources and explore design elements that are likely to affect the power of factorial designs with EIC.

### **Factorial and Fractional Factorial Designs**

Consider the following hypothetical example. Suppose an investigator wishes to develop an intervention program to promote weight loss among overweight individuals. There are three factors of theoretical interest to the investigator, each with two levels, which could be labeled *On* (experimental) and *Off* (control) for convenience (the levels could instead be “low” and “high” or any other dichotomy). The factors are whether or not the individual is offered (1) weekly videos that provide instructional and motivational training (Video), (2) encouraging text messages (Texts), and (3) meal replacement (Meals). We assume the investigator is interested in screening these intervention components in order to construct an efficient and high-quality

intervention package. In particular, the investigator would like to address three scientific questions concerning the selection of these intervention components: (1) would the targeted outcome be improved by including weekly videos in the intervention? (2) would it be improved by including text messages? and (3) would it be improved by including meal replacement?

To simultaneously address all three scientific questions, the investigator can use a complete factorial design. With  $K$  dichotomous factors, a complete factorial design requires  $2^K$  conditions. In the current example, this would be a  $2 \times 2 \times 2$  (or  $2^3$ ) factorial design, which involves eight experimental conditions. Table 1 shows the conditions of this hypothetical factorial design. This factorial design is balanced at the condition level, meaning that for each factor, four of the eight conditions have the *Off* level, and four have the *On* level. This property makes the experiment more efficient, in other words, provides greater power for testing the effects of interest, relative to other design alternatives (see Collins, Dziak, & Li, 2009 for more details). Here, we use the term *level* when referring to the value of one of the independent factors (e.g., Video = *On*), and the term *condition* when referring to a combination of levels of all the factors (e.g., Video = *Off*, Texts = *On*, Meals = *Off*). We use the term *main effect* when referring to the difference between levels of a particular factor, averaging across all experimental conditions (e.g., the main effect of Video is the average difference in response between the four conditions with Video = *On* and the four with Video = *Off*; see Myers & Well, 2003).

Table 2 shows the effect-coded design matrix, coding *On* = +1, *Off* = -1. The use of effect coding (-1 and +1) instead of dummy coding (0 and 1) is highly recommended when analyzing data from factorial screening experiments (Collins, Dziak, Kugler & Trail, 2014). Effect coding conveniently enables the significance tests of the regression coefficients to be

directly interpreted as significance tests of main effects and interactions in an ANOVA framework (see Chakraborty, Collins, Strecher & Murphy, 2009; Myers & Well, 2003).

As the number of factors in a factorial experiment increases, the number of experimental conditions increases rapidly, although the total sample size needed to maintain power may not increase appreciably. With a large number of factors a complete factorial experiment may not be feasible, due to the expense and complexity of implementing so many experimental conditions. In these cases, the investigator might consider a *fractional* factorial design as an alternative. Fractional factorial designs offer many of the advantages of a complete factorial design, while requiring considerably fewer experimental conditions (Kirk, 2003; Wu & Hamada, 2011). These designs are a variation upon factorial designs, involving the use of a subset of the experimental conditions of a complete factorial design, *carefully chosen* to preserve key statistical properties. Consider our weight loss example. One possible fractional factorial design would consist of only half of the conditions in the complete factorial design, represented by rows 2, 3, 5, and 8 from Table 2. This subset preserves the property that all effects are represented by a balanced number of conditions (e.g., each factor is  $-1$  for half of the rows and  $+1$  for the other half). Hence, the main effects can be efficiently tested without implementing all eight conditions. The 2, 3, 5, 8 design would be described as a  $2^{3-1}$  fractional factorial, indicating that this particular fractional factorial design is  $2^{-1} = 1/2$  fraction of the complete  $2^3$  factorial; for this reason it is also called a half-fraction factorial. The advantages of complete and fractional factorial designs for studying multi-component behavioral interventions are detailed in Collins et al. (2009).

### Full and Partial EIC

The factorial experiment described above for the weight loss intervention example is a straightforward randomized factorial design in which there is no EIC. When a factorial design

involves EIC, dependence in data is generated within clusters of individuals. In this context we distinguish between factorial designs with full EIC and factorial designs with partial EIC.

### **Example: Factorial Designs with Full EIC**

Consider the hypothetical factorial experiment described above. Suppose the two levels of the Video factor both involve a treatment that is delivered in groups of about five individuals each. These are experiment-induced clusters. Each group meets weekly and involves discussions and social support facilitated by a trained practitioner. The videos are offered only to individuals randomized to the *On* level of Video, who view and discuss the videos at their weekly group meetings. Groups receiving the *Off* level still meet and discuss their experiences, but do not view the videos. In this design, individuals will be randomly assigned not only to a level of Video, but also to a group within their level of Video. To avoid contamination of experimental factors and/or perceptions of inequalities within the group, the other two factors (i.e., Texts and Meals) must also be held constant within each group. This can be accomplished by first randomly assigning individuals to groups, and then randomly assigning each group to the levels of each of the three factors. Thus, each individual is nested within a cluster (group), and each cluster will belong to one of the eight conditions in Table 3. In this scenario, the individual-level outcomes (e.g., weight measurements) are independent at pretest, but no longer independent at posttest, because group members potentially influence each other, and all will be influenced by the shared practitioner. For simplicity, we assume that it is not necessary to model a practitioner effect in addition to a group effect; we return to this in the Discussion section.

This situation differs from a typical cluster-randomized factorial experiment (as described by Dziak et al. 2012), in which the clusters are units that exist prior to experimentation, such as schools, clinics, or workplaces. In a cluster-randomized factorial experiment, the response is

expected to have a positive intraclass correlation (ICC) both at pretest (i.e., prior to the intervention) and at posttest (following the intervention). However, in the current example, the clusters are created as part of the study by random assignment. Thus, the expected pretest ICC is zero because individuals have no shared experience prior to the intervention, whereas the expected posttest ICC is positive because individuals from the same group are likely to have shared experiences during the study. Here, we use *factorial designs with full EIC* to label factorial experiments in which clusters are generated for all individuals in the course of the study. The models and power formulas from Dziak et al. (2012) do not apply to these studies, because Dziak et al. assumed that a positive ICC exists at both pretest and posttest.

### **Example: Factorial Designs with Partial EIC**

Consider yet another variation of the weight loss intervention example, in which all individuals receive the weekly videos, so that Video is no longer a factor in the design. Instead, the investigator considers a different factor, Support, aiming to assess the efficacy of group support (i.e., Support = *On*) vs. no group support (i.e., Support = *Off*). Suppose that only individuals randomized to the *On* level of Support are assigned to groups of about five individuals each, who meet weekly to watch and discuss the videos. Those randomly assigned to the *Off* level of Support are simply given the videos and asked to view them at their own convenience. The other two factors (Texts and Meals) remain the same. As before, each individual in a given support group is given the same levels of all of the assigned experimental factors (i.e., the same levels of Text and Meals) as his/her fellow group members. In this scenario, the individuals in the *On* level of Support are clustered, whereas those in the *Off* level remain independent. The multilevel data generated from such studies is known as partially nested (Bauer, Sterba, & Hallfors, 2008). We use *factorial designs with partial EIC* to label factorial

experiments in which clusters are generated by experimentation for only a subset of the individuals. Once again, these studies require new regression models and power formulas.

### **Summary of Examples**

In both full and partial EIC it is necessary to take cluster-level variation into account when planning the sample size for powering the experiment and when analyzing the resulting data. However, there is an important difference. In full EIC the multilevel structure of the data is the same for all experimental conditions; in partial EIC it may not be (Baldwin, Bauer, Stice, & Rohde, 2011; Moerbeek & Wong, 2008). Because factorials with full EIC and factorials with partial EIC may induce different structures of variation in the outcome, power planning and data analysis should be conducted in a way that appropriately reflects each approach.

In the following section we discuss modeling issues separately for factorial designs with full EIC and factorial designs with partial EIC. For clarity, we begin with the simple case of a single-factor experiment (i.e., a RCT), and continue with factorial experiments. Throughout, for consistency, we use the multilevel modeling approach, rather than the ANOVA framework for analyzing multilevel data. The former offers more flexibility in that it does not require each cluster to contain exactly the same number of individuals. Readers more familiar with the latter can rely on the extant literature, which explains the link between ANOVA and multilevel models in detail (e.g., Kenny, Bolger & Kashy, 2002; Hox & Kreft, 1994).

### **Modeling for a Single-Factor EIC**

#### **A Single-Factor Experiment with Full EIC**

As noted earlier, when clusters are generated for all individuals, cluster-level variance might be non-zero for all individuals, but only at posttest. Therefore, if there is *no pretest*, it is reasonable to simply treat this design as a between-clusters design, as in Model 2 of Dziak et al.

(2012) or standard references on cluster-randomized experiments (e.g., Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). Specifically, let  $X$  denote the treatment variable, where 0 represents the control condition and 1 represents the experimental condition, and let  $Y$  denote the outcome of interest at posttest, which might correspond to the individual's weight at a 6-month follow-up. With Level 1 representing the individual and Level 2 representing the cluster, the response  $Y_{ij}$  for individual  $i$  in cluster  $j$  can be modeled as

$$\begin{aligned} \text{Level 1: } Y_{ij} &= \beta_{0j} + e_{ij} \\ \text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01}X_j + u_j \\ \text{Combined: } Y_{ij} &= \gamma_{00} + \gamma_{01}X_j + u_j + e_{ij}, \end{aligned} \tag{1}$$

where  $e_{ij} \sim N(0, \sigma_e^2)$  and  $u_j \sim N(0, \tau_u^2)$ , where  $N()$  represents the normal distribution with the mean and variance specified. Because cluster-randomized designs have been described in detail in the references mentioned above, we do not elaborate on this scenario.

Model 1 can be extended to include individual-level pretest response  $P_{ij}$  as a covariate, where  $P_{ij}$  might correspond to the individual's weight at baseline. Other covariates can be included in a similar manner:

$$\begin{aligned} \text{Level 1: } Y_{ij} &= \beta_{0j} + \beta_{1j}P_{ij} + e_{ij} \\ \text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01}X_j + u_j \\ &\quad \beta_{1j} = \gamma_{10} \\ \text{Combined: } Y_{ij} &= \gamma_{00} + \gamma_{01}X_j + \gamma_{10}P_{ij} + u_j + e_{ij}, \end{aligned} \tag{2}$$

where, as before,  $e_{ij} \sim N(0, \sigma_e^2)$ , and  $u_j \sim N(0, \tau_u^2)$ . Here,  $\sigma_e^2$  describes the random error after adjusting for the pretest and cluster membership, and  $u_j$  represents the between-cluster variation. Throughout, for simplicity, we work with the assumptions that the effect of pretest does not vary

between clusters and that the pretest does not interact with treatment (e.g., Model 30 in Bauer et al., 2008). An interesting feature of Model 2 is that it does not require assuming zero ICC at pretest or no systematic differences between the control and the experimental conditions in pretest responses. Hence, it remains appropriate even when individuals are not randomly assigned to clusters and conditions. An alternative to Model 2 might be a repeated-measures model for changes in the response between pretest and posttest. Such a model will include three levels for observations time (Level 1), individual (Level-2) and cluster (Level-3). Throughout, we focus solely on covariate-adjusted models because they require fewer levels of nesting and hence are computationally simpler than the corresponding repeated-measures models. However, either approach might be deemed by investigators to be more relevant to the specific process they are studying. Therefore, as a supplement we provide a technical report (Dziak & Nahum-Shani, 2016) containing models and power formulas for the repeated-measures approach as well.

### A Single-Factor Experiment with Partial EIC

Where there is partial EIC, it is necessary to take cluster-level variation into account for individuals in the experimental condition, but not for those in the control condition, because the latter are not clustered. However, following Roberts and Roberts (2005) and Bauer et al. (2008), and in order to conveniently use multilevel notation for the entire sample, we treat individuals in both conditions as clustered, with control participants comprising clusters of size one. This requires setting up the model in a special way. We continue to dummy-code the levels of  $X$  as 0 (for control condition) and 1 (for experimental condition), and assume for the moment that there are no pretest measurements. With Levels 1 and 2 representing the individual and the cluster, respectively, the response  $Y_{ij}$  for individual  $i$  in cluster  $j$  can be modeled as

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

$$\begin{aligned} \text{Level 2: } & \beta_{0j} = \gamma_{00} + (\gamma_{01} + u_j)X_j \\ \text{Combined: } & Y_{ij} = \gamma_{00} + (\gamma_{01} + u_j)X_j + e_{ij}, \end{aligned} \quad (3)$$

where  $e_{ij} \sim N(0, \sigma_e^2)$  and  $u_j \sim N(0, \tau_u^2)$ . The control-condition mean response is  $\gamma_{00}$  and the experimental-condition mean response is  $\gamma_{00} + \gamma_{01}$ ; thus  $\gamma_{01}$  is the overall treatment effect. An interesting feature of Model 3 is that the Level 2 slope is allowed to vary across clusters, while the Level 2 intercept is fixed. This feature allows a cluster-level variation in mean response, but only for those individuals who were randomized to the experimental condition (i.e., who have  $X_j = 1$ , not  $X_j = 0$ ). If a random  $\beta_{0j}$  had been present for individuals with  $X_j = 0$  (control condition) also, then they too would have a cluster-level variation in mean response. However, because individuals in the control condition are, in fact, not clustered (except trivially in clusters of size one), such cluster-level variation cannot be properly estimated or interpreted in their case. To avoid this difficulty, Model 3 is specified in a way that eliminates the cluster-level random component for individuals in the control condition. This allows the response to be treated as cluster-correlated for individuals in the experimental condition ( $Y_{ij} = \gamma_{00} + \gamma_{01} + u_j + e_{ij}$ ) and as independent for individuals in the control condition ( $Y_{ij} = \gamma_{00} + e_{ij}$ ).

Model 3 can be extended further to allow the variance of the Level 1 residuals  $\sigma_e^2$  to differ between the unclustered control and the clustered experimental conditions (see Roberts & Roberts, 2005; and Bauer et al., 2008 for more details). It would be somewhat unreasonable to expect the variance of the Level 1 residuals to be the same for clustered and unclustered individuals (see Bauer et al., 2008), and therefore, for the remainder of this paper, we allow them to differ whenever partial EIC is being used. We denote the Level 1 error variance for unclustered individuals as  $\sigma_{e0}^2$  and for clustered individuals as  $\sigma_{e1}^2$ .

As with Model 2, individual-level pretest  $P_{ij}$  (as well as other covariates) can be added to Model 3 as follows:

$$Y_{ij} = \gamma_{00} + (\gamma_{01} + u_j)X_j + \gamma_{10}P_{ij} + e_{ij} \quad (4)$$

where  $e_{ij} \sim N(0, \sigma_e^2)$  and  $u_j \sim N(0, \tau_u^2)$ .

### Modeling for a Multiple-Factor Case

The models discussed so far have assumed only a single randomized independent variable, but they can all be extended to allow multiple factors in a factorial experiment. The models developed in this section can accommodate either factorial designs with full EIC or factorial designs with partial EIC. However, in order to simplify the description, additional assumptions are made about how the clustering is determined in each case. Specifically, in the case of full EIC, we assume that each individual is assigned to only one experiment-induced cluster, and that to minimize the risk of contamination (see Slymen & Howell, 1997) all factors  $X_1, \dots, X_K$ , are assigned in such a way that all members of a cluster will share the same levels. In the case of partial EIC, we assume that only one of the factors, denoted  $X_1$ , induces clustering, in the sense that individuals assigned to the +1 level of  $X_1$  are clustered and those assigned to the -1 level of  $X_1$  are not. The other factors  $X_2, \dots, X_K$  are assumed to have levels that can be delivered both to clustered and unclustered individuals. Hence, in this setting, each individual might belong (or not) to only one experiment induced cluster. As before, to minimize the risk of contamination, we assume that the other factors  $X_2, \dots, X_K$  are assigned in such a way that all members of a cluster will share the same levels. Factorial designs in which an individual may belong to several different clusters are beyond the scope of this article.

Earlier we explained that in the case of partial EIC it is convenient to dummy-code the  $X$  variable for group treatment (i.e., to represent *Off* and *On* levels as 0 and 1 respectively) so that

the random cluster effect will automatically become zero for individuals in the unclustered condition. However, the dummy-coding approach is disadvantageous when multiple factors are under investigation, because the resulting coefficients in a linear model would not be independent in their distributions or interpretations and would not correspond to main effects or interactions in the usual ANOVA sense (see Chakraborty et al., 2009). In this case, the test of main effects and interactions has to be done by using linear combinations of model coefficients, which can be a source of confusion and inconvenience. Alternatively, as noted by Dziak et al. (2012), effect coding (-1 and +1) can be used to make the interpretation of the regression coefficients in this setting more convenient. As we explain below and in Appendix A, when effect coding is used, scientists can directly interpret each coefficient in the linear model for multiple factors as an independent test of a main effect or an interaction, without having to test more complicated linear combinations of model parameters. Hence, to be able to both conveniently model the random cluster effect only for individuals in the clustered condition and interpret the effects of multiple factors in a straightforward manner, we employ an approach that combines *both* a dummy-coded and an effect-coded clustering factor. The justification for this seemingly unusual approach is given below and elaborated upon in Appendix A.

Continuing with our three-factor example for simplicity, let  $X_1$ ,  $X_2$  and  $X_3$  be the effect-coded representations (-1 for *Off* and +1 for *On*) of the first, second, and third factors. Also, let  $C$  be an indicator of whether the individual has been assigned to a nontrivial cluster ( $C = 1$ ) or to a trivial cluster of size 1 ( $C = 0$ ). In the case of partial EIC,  $C$  is a dummy-coded version of the assumed cluster-generating factor  $X_1$ ; that is,  $C = 1$  if  $X_1 = +1$ , and  $C = 0$  if  $X_1 = -1$ . In the case of full EIC,  $C = 1$  for everyone, since all individuals are assigned to nontrivial clusters. Thus, we integrate Models 1 and 3 and extend them to multiple factors as follows:

$$\begin{aligned}
\text{Level 1:} \quad & Y_{ij} = \beta_{0j} + e_{ij} \\
\text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01}X_{1j} + \gamma_{02}X_{2j} + \gamma_{03}X_{3j} + \\
& \gamma_{04}X_{1j}X_{2j} + \gamma_{05}X_{1j}X_{3j} + \gamma_{06}X_{2j}X_{3j} + \gamma_{07}X_{1j}X_{2j}X_{3j} + u_jC_j.
\end{aligned}$$

The corresponding mixed model is

$$\begin{aligned}
Y_{ij} = & \gamma_{00} + \gamma_{01}X_{1j} + \gamma_{02}X_{2j} + \gamma_{03}X_{3j} + \\
& \gamma_{04}X_{1j}X_{2j} + \gamma_{05}X_{1j}X_{3j} + \gamma_{06}X_{2j}X_{3j} + \gamma_{07}X_{1j}X_{2j}X_{3j} + u_jC_j + e_{ij},
\end{aligned} \tag{5}$$

where  $e_{ij} \sim N(0, \sigma_e^2)$ , and  $u_j \sim N(0, \tau_u^2)$ .

As noted earlier, Model 5 can be used to describe factorial experiments with either full or partial EIC. If  $C_j = 1$  for all individuals, it represents a factorial design with full EIC, and  $u_j$  becomes a familiar additive random effect (random intercept term) representing cluster-level variability. If  $C_j = 1$  for only those having  $X_1 = +1$  and  $C_j = 0$  otherwise, then there is partial EIC, and the cluster-level random effect  $u_j$  becomes relevant only for a subset of the individuals.

Even though the variables  $C_j$  and  $X_{1j}$  represent the same factor, they are not confounded in Model 5. This is because  $X_{1j}$  is used to model the fixed effect of that factor, and  $C_j$  is used to model the cluster-specific random effects of that factor. One way to clarify this distinction is to interpret  $\gamma_{01}X_{1j}$  as the overall average effect of receiving group treatment, and to interpret  $u_jC_j - \gamma_{01}X_{1j}$  as the cluster-specific random deviation from that average (i.e., how well or poorly a particular group functioned). Hence, inferences for  $\gamma_{01}$  and  $\tau_u^2$  are not confounded even though the values of  $C_j$  and  $X_{1j}$  are collinear. This approach is similar to the one employed in Models 3 and 4, as well as the models employed by Roberts and Roberts (2005) and Bauer et al. (2008), where  $X_{1j}$  is effectively used twice, once to model the fixed effect of group treatment ( $\gamma_{01}X_{1j}$ ) and another time to model the random effect of group treatment ( $u_jX_{1j}$ ).

As noted earlier, because effect coding is used for the fixed effects part of the model, model coefficients can be interpreted as tests of main effects in the usual ANOVA sense. For any given cell of the  $2 \times 2 \times 2$  design,  $C_j$  is a constant and  $u_j$  is a random variable with expectation zero. Hence, the expectation of the product  $C_j u_j$  is zero and the mean of any cell in the design is given by  $E(Y_{ij}|X_{1j}, X_{2j}, X_{3j}) = \gamma_{00} + \gamma_{01}X_{1j} + \gamma_{02}X_{2j} + \gamma_{03}X_{3j} + \gamma_{04}X_{1j}X_{2j} + \gamma_{05}X_{1j}X_{3j} + \gamma_{06}X_{2j}X_{3j} + \gamma_{07}X_{1j}X_{2j}X_{3j} + 0 + 0$ . In other words, the structure of the clustering does not affect the interpretation of the linear regression coefficients.

For example, based on Model 5 the main effect of  $X_1$  is

$E(Y_{ij}|X_{1j} = +1) - E(Y_{ij}|X_{1j} = -1) = (+\gamma_{01}) - (-\gamma_{01}) = 2\gamma_{01}$ . This is because the other effects, including main effects and interactions, average out when contrasting the average of the four means of the cells with  $X_1 = 1$  with the four means of the cells with  $X_1 = -1$ . More generally, a test of the main effect of any factor  $k$  is a test of the difference in average expected responses between the two levels of  $X_k$  (defining each level's average by averaging the means of the cells composing the level), namely by  $(+\gamma_k) - (-\gamma_k) = 2\gamma_k$ . Because effect coding is used, all coefficients other than  $\gamma_k$  cancel out when calculating the main effect of factor  $k$ .

Similarly, the interaction of  $X_1$  and  $X_2$ , which represents the extent to which the difference in the response between the two levels of  $X_1$  varies across the two levels of  $X_2$ , averaging across  $X_3$ , is  $[E(Y_{ij}|X_{1j} = +1, X_{2j} = +1) - E(Y_{ij}|X_{1j} = -1, X_{2j} = +1)] - [E(Y_{ij}|X_{1j} = +1, X_{2j} = -1) - E(Y_{ij}|X_{1j} = -1, X_{2j} = -1)] = (\gamma_{01} + \gamma_{02} + \gamma_{04}) - (-\gamma_{01} + \gamma_{02} - \gamma_{04}) - (\gamma_{01} - \gamma_{02} - \gamma_{04}) + (-\gamma_{01} - \gamma_{02} + \gamma_{04}) = 4\gamma_{04}$ . More generally, a test of the interaction between any two factors  $X_a$  and  $X_b$  is a test of the difference in averaged simple effects of  $X_a$  between levels of  $X_b$  or vice versa, namely  $4\gamma_{a,b}$ . Each simple effect is defined by averaging across means of cells composing a specific combination of levels of  $X_a$  and

$X_b$ , and calculating the difference in averages between the *On* and *Off* levels of one factor ( $X_a$ ) for each specific level of the other factor ( $X_b$ ). Researchers sometimes define the interaction effect as half this quantity (i.e.,  $2\gamma_{a,b}$ ) to make its scale comparable to the main effects.

Regardless, a test of whether  $\gamma_{a,b} = 0$  is a test of whether factors  $X_a$  and  $X_b$  interact in the ANOVA sense, averaging over levels of the other factors under investigation.

Note that Model 5 includes all interactions, but it could alternatively include only lower-order ones (e.g., two-way but not three-way, by constraining  $\gamma_{07} = 0$ ). Because of the effect coding, the lower-order coefficients would have roughly the same interpretation, regardless of whether or not the higher-order interactions were included (see more details in Myers & Well 2003, and Dziak et al., 2012). This is because we assume either that cell sizes are balanced, or that a weighted average across cells is used in defining the main effects and interactions.

In summary, despite the unusual form of the error structure, the fixed effects coefficients have the same interpretations as in the linear model representation of classic ANOVA without clustering (see Myers & Well, 2003) and as in the clustered factorial designs previously described in the literature (e.g., Dziak et al., 2012). This is explained further in Appendix A.

Model 5 can also be extended to include a pretest or other covariates as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{10}P_{ij} + \gamma_{01}X_{1j} + \gamma_{02}X_{2j} + \gamma_{03}X_{3j} + \gamma_{04}X_{1j}X_{2j} + \gamma_{05}X_{1j}X_{3j} + \gamma_{06}X_{2j}X_{3j} + \gamma_{07}X_{1j}X_{2j}X_{3j} + u_jC_j + e_{ij}, \quad (6)$$

where  $e_{ij} \sim N(0, \sigma_e^2)$  and  $u_j \sim N(0, \tau_u^2)$ . As noted earlier, the error variance in Model 6 (as well as in Models 3-5) can be allowed to differ between unclustered ( $\sigma_{e0}^2$ ) and clustered ( $\sigma_{e1}^2$ ) individuals. Further, one could extend Model 6 (as well as Models 2 and 4) to allow the effect of pretest to vary between clusters or allow the pretest to interact with treatment. Note that if  $C_j$  is a

dummy-coded version of  $X_{1j}$  (i.e., the partial EIC setting), then Model 6 becomes a multiple-factor generalization of Model 4; if  $C_j$  is 1 for all individuals (i.e., the full EIC setting), then Model 6 becomes a multiple-factor generalization of Model 2. Thus, Models 5 and 6 can be used with factorial experiments having either full or partial EIC. In the following section we propose formulas for calculating power for these models.

### Estimating Power

Because the test for a main effect or an interaction in any of the designs considered in this paper can be viewed as a significance test for a coefficient in a linear mixed model, it is reasonable to estimate the power for this test using the noncentral  $F$  distribution, as in Dziak et al. (2012). That is, we assume that the power can be approximated by the probability that a noncentral  $F_{1,v}$  variate, having noncentrality parameter  $\lambda$ , exceeds the critical value  $\kappa$  of the test to be performed. Here,  $\kappa$  is the value such that a central  $F_{1,v}$  variate has a probability  $\alpha$  of exceeding  $\kappa$  under  $H_0$ . The  $\lambda$  parameter represents the amount of evidence against  $H_0$  that the sample is expected to provide; it is calculated as

$$\lambda = \frac{\gamma^2}{\text{Var}(\hat{\gamma})} \quad (7)$$

where  $\gamma$  is the regression parameter in question. The numerator degrees of freedom of the test of a given main effect or interaction, assuming dichotomous factors, is 1, because a single coefficient is being set to zero under  $H_0$ . A good estimate for the denominator degrees of freedom  $v$  in the full-EIC design is the number of clusters minus the number of regression coefficients to be estimated. This is also a good conservative initial estimate of the degrees of freedom  $v$  when planning power for partial EIC designs. However, for actual data analysis in

partial EIC designs  $\nu$  should be empirically estimated using Satterthwaite's approximation (Roberts & Roberts, 2005).

The formula for  $\text{Var}(\hat{\gamma})$  depends on whether a pretest is present and whether full or partial EIC is being used. The formulas for full EIC with or without a pretest are presented in Table 4, and the formulas for partial EIC with or without a pretest are presented in Table 5. These formulas apply to both main effects and interactions. The derivations for these formulas are described in Appendix B. The variance formulas in Tables 4 and 5 are presented in two forms: one expressed directly in terms of the different variance components and one re-expressed in terms of the posttest variance, the pretest-posttest correlation, and the posttest ICC. The latter form may be easier to use in sample size planning because the posttest variance cancels out in practice if standardized effect sizes are being used, and plausible values for the correlations can be found in the literature.

The size of the coefficient  $\gamma$  in Expression 7 expresses the magnitude of its corresponding main effect or interaction. As noted earlier, in Models 5 and 6, the main effect of any factor  $X_k$  is quantified by  $ME = 2\gamma_k$ . Thus, if for power planning purposes the minimum detectable main effect of  $X_k$  is desired to be a quantity  $ME$ , then set  $\gamma_k = ME/2$  in the numerator of Expression 7. The expression can also be restated in terms of Cohen's standardized difference  $d = ME/\sigma_Y$ , where  $\sigma_Y$  is the posttest standard deviation, adjusting for any cluster and treatment effects that may exist but not adjusting for pretest. Specifically,

$$\lambda = \frac{\sigma_Y^2 d^2}{4\text{Var}(\hat{\gamma})}. \quad (8)$$

Similarly, an interaction can be quantified by  $4\gamma_{a,b}$ , with  $\gamma_{a,b}$  representing the coefficient for the interaction between  $X_a$  and  $X_b$  (e.g.,  $4\gamma_{04}$  in the case of the interaction between  $X_1$  and  $X_2$  in Models 5 and 6). Hence, if for power planning purposes the minimum detectable interaction is

desired to be some quantity denoted  $q$ , then set  $\gamma_{a,b} = q/4$  in the numerator of Expression 7 (see Dziak et al., 2012, Appendix A). Alternatively, if the interaction is defined as half the difference in simple effects, it is represented by  $2\gamma_{a,b}$ , and one would set  $\gamma_{a,b} = q/2$ . Recall that these equalities hold regardless of the random effects structure because the population mean for each of the random effects is always zero for each cell in the design.

Finally, under the usual assumption of asymptotic normality of the maximum likelihood estimate,  $\text{Var}(\hat{\gamma})$  can be used to estimate the minimum detectable effect  $\gamma_{MD}$ . Supposing without loss of generality that  $\gamma > 0$ , assuming for simplicity that  $\text{Var}(\hat{\gamma})$  is known instead of estimated (as is usual in power calculations), and considering  $H_0$  rejections only in the correct direction,  $\hat{\gamma}$  would be judged statistically significant if  $\hat{\gamma}/\sqrt{\text{Var}(\hat{\gamma})} > z_{1-\frac{\alpha}{2}}$ , where  $z_{1-\frac{\alpha}{2}} \approx 1.96$  for an  $\alpha = .05$  two-sided test. Then  $\gamma_{MD}$  for a desired power  $1 - \beta$  (e.g., .80) is the value of  $\gamma$  such that

$$P\left(\frac{(\hat{\gamma} - \gamma)}{\sqrt{\text{Var}(\hat{\gamma})}} > z_{1-\frac{\alpha}{2}} - \frac{\gamma}{\sqrt{\text{Var}(\hat{\gamma})}}\right) = 1 - \beta,$$

where  $(\hat{\gamma} - \gamma)/\sqrt{\text{Var}(\hat{\gamma})}$  has a standard normal distribution under  $H_1$ . So  $z_{1-\frac{\alpha}{2}} - \frac{\gamma_{MD}}{\sqrt{\text{Var}(\hat{\gamma})}} = z_{1-\beta}$ ,

where  $z_{1-\beta} \approx .8414$  for  $1 - \beta = .80$ . Then

$$\gamma_{MD} = \sqrt{\text{Var}(\hat{\gamma})} (z_{1-\beta} - z_{1-\beta}). \quad (9)$$

The minimum detectable difference and minimum detectable scaled difference (e.g., Oakes and Feldman, 2001) between levels of the factor of interest would then be  $2\gamma_{MD}$  and  $2\gamma_{MD}/\sigma_Y$ , respectively. That is,  $2\gamma_{MD}/\sigma_Y$  would be the minimum detectable Cohen's  $d$ .

In the formulas for factorial designs with full EIC (Table 4), it is assumed that all experiment-induced clusters have the same size  $n$ . Additionally, balance is assumed on all of the factors. This means that each cell defined by the combinations of  $X_1, X_2, \dots, X_K$  has the same number of clusters, each containing the same number of individuals. In the formulas for factorial

designs with partial EIC (Table 5), it is assumed that all experiment-induced clusters have the same size  $n$ . Additionally, balance is assumed on all of the factors, *except* for  $X_1$ . Thus, it is assumed that each cell defined by the combinations of  $X_2, \dots, X_K$  has the same number of clusters and the same number of individuals. Of course, such balance will not hold exactly in practice, but in the simulation experiments presented later we found that the power formulas still perform very well under small imbalances.

The formulas for partial EIC (Table 5) do *not* assume that the same number of individuals have  $X_1 = +1$  as  $X_1 = -1$  (i.e., that the same number of individuals are clustered as are unclustered), because that might not be feasible or even desirable. Cells with clustering are subject to cluster-level variance in addition to their individual-level variance, so the means of cells with  $X_1 = +1$  might be estimated with higher design effects. Note that the design effect is defined as the ratio of the sampling variance in the clustered population to the corresponding sampling variance obtained if individuals were independent; the larger the design effect, the larger the sample size required for achieving adequate power (see Dziak et al., 2012 for a detailed discussion). Therefore, perhaps more total individuals should be assigned to the  $X_1 = +1$  cells to compensate for the unequal amount of estimation error. Of course, it is also not assumed that the number of clusters in the  $X_1 = +1$  conditions is the same as the number of individuals in the  $X_1 = -1$  conditions. A reasonable conjecture, which we test in Simulation Study 2, is that the optimal allocation would be somewhere between these two extremes.

The formulas presented here are informally derived in Appendix B. Further evidence that they are valid can be obtained via simulation. In the following section we report the results of simulation studies illustrating the performance of the power formulas in factorial designs with full EIC (Simulation Experiment 1) or partial EIC (Simulation Experiment 2), respectively.

### Monte Carlo Simulation Studies

We conducted two simulation studies to address three primary questions about the feasibility of factorial designs with EIC. First, are the null hypothesis tests for the main effects and interactions valid (i.e., is the Type I error rate no higher than  $\alpha$  when  $H_0$  is true)? Second, is there acceptable statistical power for main effects and interactions in the context of a screening experiment with a realistic number of individuals? Third, do the proposed power formulas give reasonably accurate estimates of the power over a range of situations (i.e., for either main effects or interactions, and in complete or fractional factorial designs with EIC, and for a range of sample sizes, cluster sizes, and ICCs)? We address these questions separately for factorial designs with full EIC (Study 1) and partial EIC (Study 2). In the case of partial EIC, we investigate two additional questions—whether individuals should be assigned in a balanced or intentionally unbalanced way on the clustering factor  $X_1$ , and whether the answers to the above questions change appreciably depending on whether error variances are assumed equal between clustered and unclustered individuals. The data were simulated and analyzed using SAS.

#### **Simulation Study 1: Factorial Designs with Full EIC**

##### **Methods**

**Data-generating model.** Each simulated dataset was based on a simulated randomized experiment with five effect-coded dichotomous factors  $X_{1j}, X_{2j}, \dots, X_{5j}$ , using the following ANCOVA model:

$$Y_{ij} = \gamma_P P_{ij} + u_j + \gamma_1 X_{1j} + \gamma_3 X_{3j} + \gamma_{1,3} X_{1j}X_{3j} + e_{ij}. \quad (10)$$

Here,  $e_{ij} \sim N(0, \sigma_e^2)$  is the individual-level random error, and  $u_j \sim N(0, \tau_u^2)$  is the random effect of cluster. This model is essentially the full-EIC ( $C_j = 1$  for all  $j$ ) version of Model 6, but with five factors instead of three, and some coefficients set to zero. We use slightly simplified

notation here instead of the more formal multilevel subscripts in the earlier models; for example we use  $\gamma_P$  (rather than  $\gamma_{10}$  as in Model 6) to denote the pretest effect on posttest. Also, for simplicity we set the intercept to zero because the intercept cancels out in the contrasts of interest and therefore does not matter for power of the tests of interest.

For convenience we assume that the pretest variance  $\sigma_P^2$  is 1 and that  $\sigma_Y^2$ , which denotes the posttest variance after controlling for treatment and cluster effects, is 1. As in the simulations of Dziak et al. (2012), we assume that  $\rho_{\text{pre},\text{post}}$ , which denotes the pretest-posttest correlation after controlling for treatment and cluster, is 0.65. Because the model implies that  $\sigma_Y^2 = \text{Var}(Y_{ij}|u_j) = \gamma_P^2\sigma_P^2 + \sigma_e^2$  and because  $\rho_{\text{pre},\text{post}} = \gamma_P/\sigma_Y$ , some algebra gives  $\gamma_P = 0.65$ , and  $\sigma_e^2 = 1 - 0.65^2 \approx 0.5775$ . We assume that the posttest ICC  $\rho_Y$  is either 0.10 or 0.20. Because the model implies that  $\rho_Y = \tau_u^2/(\tau_u^2 + \gamma_P^2\sigma_P^2 + \sigma_e^2)$ , some algebra gives  $\tau_u^2 = \rho_Y/(1 - \rho_Y)$ . This means we set  $\tau_u^2$  to either  $0.1/(1 - 0.1) \approx .1111$  or  $0.2/(1 - 0.2) = .25$ .

We assume for simplicity that  $\gamma_1 = \gamma_3 = \gamma_{1,3}$  and set this value to 0, 0.1, 0.15, or .25. The conditions in which the coefficients are set to 0 enable estimation of the Type I error rate, and the conditions in which the coefficient is nonzero allow estimation of power for different effect sizes. Because  $\sigma_Y = 1$ , setting  $\gamma = 0.1$  for a main effect corresponds to  $d = 2\gamma/\sigma_Y = 0.2$  ("small" in Cohen, 1988),  $\gamma = 0.15$  corresponds to  $d = 0.3$  ("small to moderate"), and  $\gamma = 0.25$  corresponds to  $d = 0.5$  ("moderate").

**Design.** We assumed two scenarios regarding the overall design of the study. The first was a *complete factorial*. In this case, each of the  $2 \times 2 \times 2 \times 2 \times 2 = 32$  possible conditions defined by the five factors were used. The second was a *fractional factorial*, specifically a half fraction, in which only 16 of these conditions were used. This half factorial design was the same as the

one shown in Table 4 of Dziak et al. (2012). In this design, each main effect is aliased with a four-way interaction, and each two-way interaction is aliased with a three-way interaction.

**Simulated sample size.** Depending on the simulation scenario, we modeled an experiment with 300, 400, 500, or 600 total individuals available, to be assigned to clusters of size 5 or 10. Note that the upper end of this range of sample sizes is not uncommon in health behavior intervention research; for example, in a meta-analysis of randomized trials (or quasi-experimental designs) in the area of tailored health behavior change interventions, Noar, Benac, and Harris (2007) found that over 50% of the reviewed studies had a sample size greater than 500. We simulated the data such that each simulated individual drops out of the study with independent random probability 0.20, which sometimes causes the cluster sizes to be unequal. When predicting power using the power formula, the 20% dropout was taken into account by treating the cluster size as 4 or 8 instead of 5 or 10, but the inequality of the cluster sizes was not further taken into account. This provides an opportunity to use simulation results to check for the robustness of the formula to somewhat unbalanced cluster sizes.

**Assumed model for performing tests.** In real life, a researcher analyzing data from a factorial study does not know for certain which interactions are negligible and which are not. That is, the investigator does not know in advance that many of the possible interactions have coefficients of zero. Therefore, we assume that the investigator fits the following model.

$$\begin{aligned}
 Y_{ij} = & \gamma_0 + \gamma_P P_{ij} + u_j + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \gamma_{1,2} X_{1j}X_{2j} + \gamma_{1,3} X_{1j}X_{3j} \\
 & + \gamma_{1,4} X_{1j}X_{4j} + \gamma_{1,5} X_{1j}X_{5j} + \gamma_{2,3} X_{2j}X_{3ij} + \gamma_{2,4} X_{2j}X_{4j} \\
 & + \gamma_{2,5} X_{2j}X_{5j} + \gamma_{3,4} X_{3j}X_{4j} + \gamma_{3,5} X_{3j}X_{5j} + \gamma_{4,5} X_{4j}X_{5j} + e_{ij}
 \end{aligned} \tag{11}$$

Model 11 does not include three- or four-way interactions, for three reasons. First, they are zero in the true data-generating model, although of course this would not be known to a real-

world investigator. Second, researchers in practice may choose to fit parsimonious models lacking third- and higher-order interactions because such complex interaction effects are typically difficult to detect and interpret. Third, in the half factorial scenario it would be impossible to fit a model that includes these interactions, because each would be aliased with an effect that is already in Model 10 (see Collins et al., 2009). The SAS code for fitting this model is provided in Appendix C.

**Other technical details.** The significance of each effect was decided by a marginal ("Type III" in SAS) significance test on the appropriate coefficient. For the purpose of predicting power, the assumed denominator degrees of freedom was counted as the number of clusters minus the number of regression parameters (the latter was counted as 17, for one intercept, one coefficient for pretest, five main effects and  $\binom{5}{2} = 10$  interactions). When actually performing the test, a Satterthwaite approximation was used to provide slightly more power.

**Summary.** Table 6 summarizes all the simulation study scenarios defined by the characteristics described above (128 scenarios). For each scenario, 5000 datasets were generated.

## Results

For simplicity, we focus only on  $\gamma_1$  when reporting the Type I error rate and power, although we could equivalently have examined a different parameter. Results for the main effect  $\gamma_3$  or the interaction  $\gamma_{1,3}$  are very similar to those for  $\gamma_1$ , in each of the 128 scenarios. The absolute difference in Type I error or in power between these parameters was never more than .026 in any scenario, and was often less than .005. Moreover, as expected, there was little systematic difference, either in Type I error or in power, between complete and fractional factorial experiments for comparable conditions. When comparing these designs, the absolute difference in Type I error was always less than .015, and the absolute difference in power was

always less than .04 and usually less than .02. Thus, we consider the two types of design together when reporting the results. Below, we summarize the results by considering each of the motivating questions in turn.

First, the tests of the main effects and interactions were valid, in that for conditions with true effect size of zero, the Type I error rate for  $\gamma_1$  was very close to nominal (between .038 and .059 for a nominal .05 test). Similarly, Type I error for the effects omitted from Model 9 (e.g.,  $\gamma_2, \gamma_4, \gamma_5, \gamma_{1,2}, \gamma_{1,4}$ ) was always between .039 and .060, regardless of scenario.

Second, the results (in Table 7) indicated that acceptable statistical power can be obtained with feasible sample sizes. For example, suppose that 500 individuals are assigned to clusters of size 5. If ICC is 0.1, then we observe a power of 0.82 for detecting a small to moderate ( $d = .3$ ) effect size. As expected, the same scenario with a larger ICC (0.2) yielded lower power (0.65) for detecting a small to moderate ( $d = .3$ ) effect size, yet acceptable power for detecting a moderate ( $d = .5$ ) effect size.

Third, the simulations indicated that the power formula provides a very good estimate of power, except in cases with only 30 clusters. In these cases, power was limited because of the extremely limited degrees of freedom (17 parameters to estimate but only 30 independent experimental units). In most cases, the power formula over-predicted power by about 1% or 2%, perhaps due to failure to correct for the effects of unequal numbers of members per cluster or unequal numbers of clusters per condition, but this is a very small difference considering the large amount of uncertainty inherent in power analysis.

### **Simulation Study 2: Factorial Designs with Partial EIC**

In addition to the three primary questions noted earlier, this simulation study was also designed to address two questions concerning the effects of the allocation proportion and the

equality of error variances, two topics that are of special interest in partial EIC settings.

Specifically, the fourth question is as follows: for maximum power, should individuals be assigned in a balanced way on  $X_1$ ? Ordinarily, it is desirable to have balanced assignment on factors. However, as discussed earlier, individuals with  $X_1 = +1$  will be subject to cluster-level variance in addition to their individual-level variance, so the means of cells with  $X_1 = +1$  essentially have larger sampling variance. Thus, we hypothesized that more individuals should be assigned to the +1 level to counteract this unequal amount of estimation error, such that the optimal allocation proportion (of individuals to  $X_1 = +1$ ) would be more than 50%. The fifth question was as follows: does the proposed power formula continue to accurately predict power in cases where the error variances for unclustered ( $\sigma_{e0}^2$ ) and clustered ( $\sigma_{e1}^2$ ) individuals are not equal? Specifically, we considered the possibility that  $\sigma_{e0}^2 > \sigma_{e1}^2$ , so that being in a cluster makes individuals more similar, above and beyond the shared cluster intercept.

## Methods

Each simulated dataset was based on a simulated randomized experiment with five dichotomous effect-coded factors  $X_{1j}, X_{2j}, \dots, X_{5j}$ , using the following data-generating model:

$$Y_{ij} = \gamma_P P_{ij} + u_j C_j + \gamma_1 X_{1j} + \gamma_3 X_{3j} + \gamma_{1,3} X_{1j} X_{3j} + e_{ij}. \quad (12)$$

Here,  $u_j \sim N(0, \tau_u^2)$ , and  $e_{ij} \sim N(0, \sigma_{e0}^2)$  for unclustered individuals, and  $e_{ij} \sim N(0, \sigma_{e1}^2)$  for clustered individuals. As in Simulation Study 1, we assume that  $\gamma_1 = \gamma_3 = \gamma_{1,3}$ , the pretest variance ( $\sigma_P^2$ ) is 1, the posttest variance after controlling for treatment and cluster effects ( $\sigma_Y^2$ ) is 1, and  $\gamma_P = 0.65$ . However, unlike the scenario used for Simulation Study 1, here the data-generating model includes a cluster-generating treatment factor  $X_{1j}$ , as well as  $C_{ij}$  –a dummy-coded version of  $X_{1j}$ , which is set to 0 if  $X_{1j} = -1$  and to 1 if  $X_{1j} = +1$ . Hence, the posttest ICC ( $\rho_Y$ ) is relevant only for the clustered individuals, and for them it is set to either 0.10 or 0.20.

In the equal variance scenario, we set  $\sigma_{e0}^2 = \sigma_{e1}^2 = 1 - \gamma_p^2 = 0.5775$  to achieve the desired pretest-posttest correlation  $\rho_{\text{pre},\text{post}} = \gamma_p/\sigma_Y^2 = .65$ , and we set  $\tau_u^2 = \rho_Y/(1 - \rho_Y)$  to achieve the desired posttest ICC of  $\rho_Y = .10$  or  $\rho_Y = .20$ . In the unequal variance scenario, we set the error variance to be twice as large for unclustered compared to clustered individuals, so that  $\sigma_{e1}^2 = \frac{2}{3}(1 - \gamma_p^2)$ , and  $\sigma_{e0}^2 = \frac{4}{3}(1 - \gamma_p^2)$ . As a result, the pretest-posttest correlation is somewhat different between clustered and unclustered individuals, namely  $\rho_{\text{pre},\text{post}} = \frac{\gamma_p}{\sqrt{\gamma_p^2 + \frac{2}{3}(1 - \gamma_p^2)}} = .723$  and  $\rho_{\text{pre},\text{post}} = \frac{\gamma_p}{\sqrt{\gamma_p^2 + \frac{4}{3}(1 - \gamma_p^2)}} = .595$ , respectively. We set  $\tau_u^2 = \gamma_p^2 + \frac{2}{3}(1 - \gamma_p^2) \frac{\rho_Y}{1 - \rho_Y}$  to achieve the desired posttest ICC in the unequal variance scenario.

Additionally, for simplicity we set  $\gamma_1, \gamma_3, \gamma_{1,3}$  in the current scenario to either 0 or 0.15 (roughly corresponding to Cohen's  $d$  of 0 or 0.3). Last, we assume that the investigator chooses to assign levels of  $X_{1j}$  either in a naively balanced way (50% clustered, 50% unclustered) or else in a seemingly unbalanced way that allocates either 60% or 70% of the individuals to clustered conditions. Note that allocating a larger portion of individuals to the clustered conditions is done in a way that increases the number of clusters in the clustered conditions, rather than the number of individuals within each cluster. Allocating more individuals to the clustered condition by increasing cluster size would be a less efficient way to improve power, because the benefit of additional observations per cluster is partially balanced out by the increased design effect associated with greater cluster sizes for a given ICC (see Baldwin et al., 2011).

**Design.** We assumed either a complete or half factorial, using the same designs as in Simulation Study 1, except that now  $X_1$  not only served as an experimental factor with a fixed effect, but also determined whether an individual is assigned to a cluster or not.

**Simulated sample size.** As before, we modeled an experiment with 300, 400, 500, or 600 total individuals available, to be assigned to clusters of size 5. Each individual has independent random probability 0.20 of dropping out. This is then taken into account just as before when calculating power.

**Assumed model for performing tests.** As before, we assume that the investigator fits a model with all main effects and two-way interactions. That is, we use Model 11 as before, except that we replace  $u_j$  with  $C_j u_j$ . The SAS code for fitting this model is shown in Appendix C.

**Other technical details.** As in Simulation Study 1, the significance of each effect was decided by a marginal ("Type III" in SAS) significance test on the appropriate regression coefficient. The assumed denominator degrees of freedom of the test for predicting power was conservatively estimated as the number of clusters minus the number of regression parameters. When actually performing the test, a Satterthwaite approximation was used. Residual error variances were allowed to differ between clustered and unclustered individuals in the analysis.

**Summary.** Table 8 summarizes all simulation study scenarios defined by the characteristics described above (192 scenarios). For each scenario, 5000 datasets were generated.

## Results

The absolute difference in Type I error between  $\gamma_1$  and either  $\gamma_3$  or  $\gamma_{1,3}$  in the  $d = 0$  scenarios did not exceed 0.012. Also, results for  $\gamma_1$  with respect to power were similar to those for  $\gamma_3$  or  $\gamma_{1,3}$ . Specifically, the absolute difference in power between  $\gamma_1$  and either  $\gamma_3$  or  $\gamma_{1,3}$  in the  $d=.30$  scenarios did not exceed 0.023 in any scenario. Therefore, only results for  $\gamma_1$  are presented and discussed. The equivalence among these effects is not immediately intuitive, given that  $\gamma_1$  is the main effect of a factor that determines clustering ( $X_1$ ), while  $\gamma_3$  is the main effect of a factor that was assigned after clustering ( $X_3$ ), and  $\gamma_{1,3}$  is the interaction between the two. Still,

as we explain in detail in Appendix B, such equivalence is reasonable mathematically, because of the factorial structure of the experiment. That is, all effects are estimated using data from all individuals; therefore all effect estimates are subject in a similar manner to the entire cluster structure of the design.

Simulated and predicted power are compared in Table 9. The difference in power between complete and fractional factorial designs was usually extremely small, but sometimes there was a slight advantage to fractional factorials, especially when  $N=300$ . This advantage has no obvious theoretical explanation but may be an artifact of randomization. With 300 individuals, 50% allocation to clustering, and cluster size 5, there are 30 clusters. For the fractional factorial design, 3 clusters were randomized to each of the 8 clustered conditions, and 6 clusters were left over. The leftover clusters were randomized independently to any clustered condition. For the complete factorial design, 1 cluster was assigned to each of the 16 clustered conditions. There were 14 leftover clusters, which were then assigned independently to any clustered condition, allowing a higher probability of poorer balance. However, this could be avoided in practice by restricted randomization. Thus, the five simulation questions can be addressed in a straightforward way for complete or fractional factorial designs together.

Regarding the first question, about validity of the null hypothesis tests, we found that for scenarios in which the true effect of Factor 1 was zero, the Type I error rate was between .039 and .058 for this effect in every scenario. In addition, coefficients that were missing from the data-generating model had Type I error rates between .040 and .064 in every scenario. It is reasonable to conclude that Type I error rate is essentially nominal, as desired.

Regarding the second question, about the feasibility of factorial designs with partial EIC, the results (in Table 9) indicate that acceptable statistical power can be obtained in such a setting.

For example, a total of 400 individuals, with half allocated to clusters of size 5, yields acceptable power (slightly above 0.8) for detecting a small to moderate effect size, given a small ICC.

Regarding the third question, about the performance of the proposed power formulas, as shown in Table 9, the power formulas appear to be very accurate in terms of predicting power for complete and fractional factorial designs with partial EIC, although the estimates tended to be slightly conservative due to the intentionally conservative degrees of freedom estimate.

Regarding the fourth question, about allocation proportion, there was usually little difference between 50%, 60%, and 70% allocation to clusters. 60% allocation to clusters was often, although not always, slightly more powerful than 50% allocation. 70% allocation was generally slightly less powerful than 50% or 60% allocation. It is reasonable to recommend either equal allocation or only slightly greater allocation to the clustered condition.

Regarding the fifth question, about the performance of the proposed power formula in cases where the error variances for unclustered and clustered individuals are not equal, we found the error variance scenarios to have very little systematic influence on power; the power formula performed well in either scenario. Consistent with the power formula in Table 5, these results indicate that power generally improves to the extent that the overall level of error variance is lower, regardless of whether the clustered and unclustered conditions differ in terms of their respective error variances. To clarify this, recall that we compared the unequal and equal error variance scenarios while holding the overall amount of error variance fixed. Specifically, we set the error variance either to .5775 for all individuals, or to .385 for clustered and .770 for unclustered individuals, two numbers whose average is .5775. Table 5 makes it clear that what matters to power is a combination of  $\sigma_{e0}^2$  and  $\sigma_{e1}^2$ . Although this combination is weighted according to allocation ( $J_1n$  versus  $J_0$ ), allocation proportions in this simulation study were set to

be equal or near-equal, and therefore  $\sigma_{e0}^2$  and  $\sigma_{e1}^2$  contributed about equally to the overall sampling variance of treatment effect. Reducing one while increasing the other, then, had little net effect on sampling variance, and therefore little effect on predicted or simulated power.

In practice, it is not clear how someone planning a partial EIC factorial experiment could predict whether error variances would be equal or not. Therefore, since it is difficult to decide on their relative sizes, and their relative sizes do not matter very much, it seems reasonable to assume equal variances for simplicity when using the power formula, and then allow unequal variances for greater robustness when analyzing the data.

## Discussion

In the current article we discussed modeling and power considerations for factorial designs with full and partial EIC. There has recently been increased interest in factorial designs as a tool for improving and developing high-quality multi-component interventions (e.g., Cook et al., 2016; Howard & Jacobs, 2016; Pellegrini et al., 2014; 2015). However, EIC is prevalent in psychological and intervention research (Baldwin et al., 2011; Roberts & Roberts, 2005), and, to our knowledge, no past literature has explained how to implement factorial designs in such a setting. Therefore, the current study serves as a bridge between these two design literatures, helping investigators plan and properly analyze data arising from factorial designs with EIC. The results of our simulation studies indicated that under reasonable scenarios of number of clusters, number of individuals within clusters, and ICC, adequate power can be achieved for detecting main effects and selected two-way interactions in factorial designs with full or partial EIC. Factorial designs with partial EIC usually offered better power than those with full EIC; however, it is possible to obtain adequate power even with full EIC.

The pattern of results obtained here is consistent with the results seen in between-cluster RCTs (Murray, 1998) and multilevel factorial designs (Dziak et al., 2012). Specifically, power increases to the extent that ICC is lower. Moreover, power increases to the extent that the number of clusters and the number of individuals within a cluster increase, with the number of clusters having more influence on power than the number of individuals within a cluster. Although our results showed that in some scenarios the number of clusters had to be rather large to obtain adequate power, this is not a result of the presence of multiple factors, but rather a result of the small to moderate effect size assumed in these scenarios. Indeed, the simulated power per factor in the five-factor screening scenario can be predicted reasonably with power formulas that assume a single factor (except for adjustment of the degrees of freedom). However, because small effect sizes were assumed in most scenarios, the sample size requirements were relatively high for some conditions. Small effect sizes were assumed here because in the context of screening experiments, the goal is to detect effects for each individual intervention component. These effects would reasonably be expected to be smaller than the effect of an entire intervention package consisting of multiple active components.

### **Complete vs. Fractional Factorials with EIC**

As expected, we found in both the full and partial EIC simulations that the complete and fractional designs were about equally powerful given an equal total sample size. Hence, the decision of whether to use complete or fractional factorial should depend on other practical, ethical and scientific considerations. Specifically, although a fractional factorial design requires the same number of individuals as a complete factorial, fewer experimental conditions are needed. Hence a fractional factorial design might be easier to implement and/or less costly than a complete factorial design (see Collins et al., 2009).

A potential disadvantage of fractional factorial designs is that they always involve some aliasing of effects. The strategy behind the use of fractional factorial designs involves deliberately aliasing effects of primary scientific interest, typically main effects and lower-order interactions, with higher-order interactions that are not of primary scientific interest and that can be assumed to be negligible in size. In these cases, the estimate of the aliased effect, which is an estimate of a combination of two or more effects, is attributed to the effect of primary interest. In cases where the assumption about the size of these higher-order interactions is incorrect, this attribution is also incorrect, and the resulting effect estimate will be an under- or over-estimate of the effect of primary interest. This, in turn, has an impact on the Type I error rate and power (Collins et al., 2009; Dziak et al., 2012). In our simulation studies we did not include substantial higher-order interactions when we generated the data. Hence, as expected, there was little systematic difference in terms of Type I error and power for the effects of interest between complete and fractional factorial experiments. In practice, software can be used to properly plan the aliasing structure of a fractional factorial design so as to reduce the risk of substantial bias (see Wu & Hamada, 2011 for a more detailed discussion of aliasing).

### **Power Planning Resources**

In the current article, we provide power planning formulas for factorial designs with either full or partial EIC. Our simulation results indicated that the proposed power planning approach provides a reasonable approximation to the actual power for both cases. While the simulation results in Tables 7 and 9 cover only a few selected scenarios, the formulas in Tables 4 and 5 can be used very widely to guide investigators in planning factorial designs with EIC. Still, the results indicated that the power formula slightly over-predicted power for factorial designs

with full EIC and slightly under-predicted power for factorial designs with partial EIC. These discrepancies were generally quite minor in practical terms.

### **Cluster Allocation in Factorial Designs with Partial EIC**

Our results with respect to cluster allocation in factorial designs with partial EIC are consistent with prior investigations of cluster allocation in RCTs with partial EIC (Baldwin et al., 2011). Specifically, our results indicate that allocating more individuals to the clustered condition provides a small increase in power compared with equal allocation.

Roberts & Roberts (2005) provided the following formula for optimal allocation of individuals in large RCTs with partial EIC:  $\frac{p}{q} = \sqrt{1 + (m - 1)ICC}$ , where  $p$  is the proportion of individuals assigned to the clustered condition,  $q$  is the proportion of individuals assigned to the unclustered condition,  $m$  is the number of individuals within each cluster, and ICC is the posttest ICC. Based on this formula, for a cluster size of 5, the optimal cluster allocation proportion ( $p$ ) is 54% clustered in the case of ICC = 0.1, and 57% clustered in the case of ICC = 0.2, regardless of sample size. This may apply to the factorial case as well, although our simulations were not precise enough to distinguish between these values. More precise estimates in specific circumstances could be achieved from specially designed simulation studies. There is probably a range of reasonable values for the proportion allocated to clusters, and these values can be chosen based on considerations other than power (e.g., ethical or practical). For example, if group treatment is highly expensive, a balanced allocation might be more warranted.

### **Randomization Plans for Factorial Designs with EIC**

When planning the randomization scheme for factorial designs with EIC, careful consideration should be given to the potential for contamination. If individuals within a cluster receive different experimental conditions, then the potential for contamination could be high (see

Dziak et al., 2012). Therefore, it will often be helpful to have a randomization plan that assures that everyone in a given cluster is also in the same condition. This can be done in various ways, depending on whether full or partial EIC is used. When planning factorial designs with full EIC, investigators can either begin by assigning individuals to clusters and then randomly assigning *clusters* to conditions, or they can equivalently begin by assigning individuals to the conditions and then assign individuals to clusters within each condition. Either method assures that all the members of a given cluster receive the same condition.

In the case of factorial designs with partial EIC, individuals cannot be assigned to clusters before they are assigned to conditions, because certain conditions are clustered and others are not. Therefore, one option would be to assign individuals first to conditions, and then randomly assign individuals to clusters within each clustered condition. Another option would be to divide the randomization scheme into three steps. First, assign individuals to the two levels of the clustering factor  $X_1$ . Second, randomly assign individuals in the *On* level (i.e., the clustering level) of  $X_1$  to clusters. Finally, randomly assign clusters in the *On* level of  $X_1$ , as well as individuals in the *Off* level of  $X_1$ , to the experimental conditions resulting from crossing the remaining (non-clustering) factors  $X_2 \dots, X_K$ . Once again, either method leads to clusters in which all members have the same treatment condition.

### **Limitations and Directions for Future Research**

The discussion of factorial designs with partial EIC in this article is limited to designs in which only one of the factors involves generating clusters. Note that this does not mean that only one factor is cluster-level or is affected by clustering, but rather that there is only a single way in which the individuals are clustered. However, other kinds of partial EIC factorial designs are possible. For example, consider the investigation of five intervention components, of which one

involves generating in-person support groups, and another involves generating online support groups via social media tools. In this case, there will be two cluster-generating factors, one aiming to assess the efficacy of in-person group support (i.e., In-Person Support = *On*) vs. no group support (i.e., In-Person Support = *Off*); and the other aiming to assess the efficacy of online group support (i.e., Online Support = *On*) vs. no online group support (i.e., Online Support = *Off*). Such an experiment is possible but beyond the scope of this paper.

Further, in the current study we assumed only one level of clustering. More complicated scenarios might occur if, for example, individuals are assigned to groups, and then multiple groups are led by each of a limited number of therapists. In these cases, there are technically two levels of clustering, as individuals are nested within groups and groups are nested within therapists. Studies involving such clustering structure use various means to address this dependency, such as balancing conditions across therapists, such that all therapists facilitate an approximately equal number of groups in each condition in an attempt to reduce therapist-level effects (e.g., Herbert et al., 2009); ignoring the therapist-level in primary analyses (Peterson, Mitchell, Crow, Crosby, & Wonderlich, 2009), sometimes after showing no differences in outcome between therapists in each treatment condition (Lecomte, Leclerc, Corbiere, Wykes, Wallace, & Spidel, 2008); or adding therapists as a covariate to control for possible significant therapist effects (e.g., Bergraff et al., 2014). The adequacy of these approaches is debated (e.g., Murray, 1998; De Jong, Moerbeek, & Van der Leeden, 2010; Wampold & Bolt, 2006). Hence, future investigations of factorial designs with EIC might focus on modeling considerations and power planning resources that accommodate more complicated designs with multiple levels of nesting.

## Appendix A

### **Justification of the Proposed Regression Approach for Factorial Designs with Partial EIC**

In this paper we proposed a new approach to modeling treatment effects in a factorial experiment with partial EIC. In that approach, we represent the cluster-generating factor twice, once with an effect-coded variable  $X_1$  (+1 for clustered and -1 for unclustered), and once with a dummy-coded  $C$  (1 for clustered and 0 for unclustered). We asserted that this approach allows investigators both to conveniently model the random cluster effect only for individuals in the clustered condition and to interpret the effects of multiple factors in a straightforward manner while using standard procedures in common analysis software (e.g., PROC MIXED in SAS; see Littell, Milliken, Stroup, Wolfinger & Schabenberger, 2006) to analyze data arising from a factorial design with partial EIC. Here, we explain in more detail why this approach does not result in a confounded model, despite the fact that the same information seems to be used twice.

We begin by proposing an intuitively reasonable model for factorial experiments with partial EIC, one that many researchers would find acceptable but which cannot easily be fit using standard software. We then consider a simple way to re-express this model using only dummy codes, which closely resembles our Model 3, based in turn on Roberts and Roberts (2005) and Bauer et al. (2008). We show that this dummy-coding approach has some inconvenient features when dealing with more than one treatment factor, which is why we elected instead to use our hybrid parameterization using a dummy code and an effect code. We describe how the results of our hybrid parameterization can be translated to and from the results of the dummy-coding approach. For simplicity, we do not consider attrition or missingness in this presentation. Also, for simplicity of notation and without loss of generality, we suppose that there are two randomized factors, that each cluster consists of three individuals per cluster, and that error

variances are the same in clustered and unclustered conditions. We ignore the pretest as in Model 5, but the basic structure of our arguments also applies to models such as 6.

To begin, consider that a reasonable model for analyzing data from a randomized experiment ought to take into account the way in which the randomization took place. In the factorial experiment with partial EIC as we described it, a typical individual is first randomly assigned to a level of factor 1. Suppose for now that factor 1 is dummy-coded and denote it as  $D_1$ . If  $D_1 = 1$ , then individual  $i$  is randomly assigned to a cluster  $j$ , which includes other individuals. If  $D_1 = 0$ , then individual  $i$  is still given a cluster number  $j$  for bookkeeping reasons, but is the only individual in this “trivial cluster” of size one. Cluster  $j$  (whether trivial or nontrivial) is then assigned to a level of the second factor  $D_2$ , also dummy-coded as 1 or 0. Note that the assignment is at the cluster level. If  $D_1 = 1$ , then clusters are assigned to the levels of the second factor to avoid contamination within a treatment group. If  $D_1 = 0$ , then assignment can still be said to be at the cluster level, because there is no difference between randomizing the individual and randomizing the one-person cluster.

Now consider modeling the cluster responses  $\mathbf{Y}_j$  for cluster  $j$  conditionally on  $D_1$  and  $D_2$ . Such a model would most naturally be expressed in parts. If  $D_1 = 0$  then  $\mathbf{Y}_j = [Y_{1j}]$  is a vector of length 1, and a reasonable linear model would be

$$(Y_{1j}|D_{1j} = 0, D_{2j} = d_2) \sim N(\lambda_1 + \lambda_2 d_2, \sigma_e^2),$$

for some regression parameters  $\lambda_1$  and  $\lambda_2$  and variance  $\sigma_e^2$ . Equivalent to the above would be

$$\text{If } D_{1j} = 0 \text{ then } Y_{1j} = \lambda_1 + \lambda_2 D_{2j} + e_{1j}, \quad (13)$$

with  $e_{1j} \sim N(0, \sigma_e^2)$ .

If  $D_1 = 1$  then  $\mathbf{Y}_j = [Y_{1j}, Y_{2j}, Y_{3j}]$  is a vector of length 3, and a reasonable multivariate normal model would be

$$\left( \begin{bmatrix} Y_{1j} \\ Y_{2j} \\ Y_{3j} \end{bmatrix} \mid X_{1j} = 1, X_{2j} = d_2 \right) \sim N \left( \begin{bmatrix} \lambda_3 + \lambda_4 d_2 \\ \lambda_3 + \lambda_4 d_2 \\ \lambda_3 + \lambda_4 d_2 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 + \tau_u^2 & \tau_u^2 & \tau_u^2 \\ \tau_u^2 & \sigma_e^2 + \tau_u^2 & \tau_u^2 \\ \tau_u^2 & \tau_u^2 & \sigma_e^2 + \tau_u^2 \end{bmatrix} \right)$$

for some  $\lambda_3$  and  $\lambda_4$ , some cluster-level variance  $\tau_u^2$ , and some individual-level variance  $\sigma_e^2$ .

Equivalent to the above would be

$$\text{If } D_{1j} = 1 \text{ then } Y_{ij} = \lambda_3 + \lambda_4 D_{2j} + u_j + e_{ij}, \text{ for } i = 1, \dots, 3, \quad (14)$$

with  $u_j \sim N(0, \tau_u^2)$  and each  $e_{ij} \sim N(0, \sigma_e^2)$ . Taken together, Expressions 13 and 14 completely specify the assumed population distribution of the responses  $\mathbf{Y}$  conditional on  $D_1$  and  $D_2$ . One could numerically optimize the log-likelihood function implied by Expressions 13 and 14 in terms of the parameters, and thus obtain maximum-likelihood or restricted maximum-likelihood estimates of the parameters. This could be done either by writing one's own code to implement an appropriate algorithm or by writing code in a general log-likelihood optimizer such as SAS PROC NLMIXED (see Littell et al., 2006). Unfortunately, this would be time-consuming and error-prone, relative to being able to use standard off-the-shelf software written for more familiar models. However, standard off-the-shelf software does not easily handle models constructed in a two-part way as in Expressions 13 and 14. Because of this dilemma, we sought a way to express Expressions 13 and 14 together in a single linear mixed effects model that could be fit more easily with more familiar procedure such as SAS PROC MIXED. Notice that Expressions 13 and 14 can be expressed simultaneously as

$$Y_{ij} = \theta_0 + \theta_1 D_{1j} + \theta_2 D_{2j} + \theta_{12} D_{1j} D_{2j} + u_j D_{1j} + e_{ij}. \quad (15)$$

Specifically,  $\lambda_1 = \theta_0$ ,  $\lambda_2 = \theta_2$ ,  $\lambda_3 = \theta_0 + \theta_1$ , and  $\lambda_4 = \theta_2 + \theta_{12}$ . There is no term for  $u_j$  alone, without being multiplied by  $D_{1j}$ , but intuitively there should not be such a term because it is not reasonable to make inferences about the cluster-level variability of people who are alone. A somewhat similar situation which can arise with conceptually nested covariates is described in Henry and Dziak (2016).

The dummy-coding notation in Model 15 is entirely adequate for running an interpretable linear mixed model. However, the disadvantage of dummy coding is that the coefficients do not correspond to ANOVA main effects or interactions and are not independent in their distributions or interpretations. For example, testing the main effect of the second factor should be equivalent to testing whether  $E(Y_{ij}|D_{2j} = 1) - E(Y_{ij}|D_{2j} = 0)$  is zero. However, from Model 15 it can be seen that  $E(Y_{ij}|D_{2j} = 1) - E(Y_{ij}|D_{2j} = 0) = \theta_2 + \frac{1}{2}\theta_{12}$ ; there is no single regression parameter for  $E(Y_{ij}|D_{2j} = 1) - E(Y_{ij}|D_{2j} = 0)$  (see, e.g., Chakraborty, Collins, Strecher & Murphy, 2009; Kugler, Trail, Dziak & Collins, 2012). This is not inherently a problem, because the test can still be done using a linear combination of parameters, but it is a source of potential confusion and inconvenience. We have observed in practice that researchers often interpret the first-order coefficients of variables as main effects, regardless of whether they are truly main effects in the ANOVA sense or not. This is problematic because the value and interpretation of the coefficients depends on whether the factors are dummy-coded or effect-coded, and the first-order coefficients of dummy-coded factors are not actually main effects but simple effects. In an attempt to avoid this risk of confusion, we consider another way to express Models 13 and 15, which represents the same assumptions as Model 15 but whose parameters are more convenient to interpret when the design involves multiple factors.

The approach we propose involves using effect coding for the two factors. We denote the effect-coded factors  $X_1$  and  $X_2$ , so that, for example,  $X_{1j} = +1$  if  $D_{1j} = 1$  and  $X_{1j} = -1$  if  $D_{1j} = 0$ ; in other words,  $X_{1j} = 2D_{1j} - 1$  and, similarly,  $X_{2j} = 2D_{2j} - 1$ . We also define the clustering indicator  $C_j = D_{1j}$ . Now let  $\gamma_0 = \theta_0 + \frac{1}{2}\theta_1 + \frac{1}{2}\theta_2 + \frac{1}{4}\theta_{12}$ ,  $\gamma_1 = \frac{1}{2}\theta_1 + \frac{1}{4}\theta_{12}$ ,  $\gamma_2 = \frac{1}{2}\theta_2 + \frac{1}{4}\theta_{12}$  and  $\gamma_{12} = \frac{1}{4}\theta_{12}$ . Then expression 15 is algebraically equivalent to

$$Y_{ij} = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \gamma_{12} X_{1j} X_{2j} + u_j C_j + e_{ij}, \quad (16)$$

which is essentially the two-factor version of Model 5. Thus either Model 15 or Model 16 will serve as a valid way of expressing Models 13 and 14 together. As argued in the main text of this paper and elsewhere, tests of the  $\gamma$  parameters correspond directly to tests of main effects and interactions in the factorial ANOVA sense.

It would alternatively be possible to simply specify that

$$Y_{ij} = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \gamma_{12} X_{1j} X_{2j} + u_j + e_{ij}$$

but constrain the variance of the  $u_j$  to be zero for the subsample with  $X_{1j} = -1$ . However, this would not be possible to do directly in many software packages, and it is logically the same as Expression 16, because Expression 16 multiplies  $u_j$  by zero whenever  $X_{1j}$  does not equal +1. Therefore, we argue that Expression 16, although initially counterintuitive, actually represents the most convenient way to write the conceptual model for this partially nested factorial design.

## Appendix B

### Derivation of the Sampling Variance Formulas for Calculating Power

Here, we provide rationales for the sampling variances given in Tables 4 and 5.

**Full EIC, No Pretest** (Model 5;  $C_j \equiv 1$ ). Consider the regression coefficient  $\gamma_k$  for the main effect of an effect-coded factor  $X_k$ . Let  $\hat{\mu}_{(X_k=L)}$  be the average of all cell means having  $X_k = L$ . For example, if there are three factors in total, then we use  $\hat{\mu}_{(X_k=+1)}$  to denote the average of the  $(+1,+1,+1)$ ,  $(+1,+1,-1)$ ,  $(+1,-1,+1)$  and  $(+1,-1,-1)$  cells. If the cell sizes (number of individuals per cell) are equal, then this is also the average of all individuals having  $X_k = L$ ; if they are unequal, then it is a weighted average (by the effective size of each cell). In either case,  $\hat{\mu}_{(X_k=L)}$  is the maximum likelihood estimate of  $E(Y|X_k = L)$  for a population with balanced allocation. The corresponding sample estimate of  $\gamma_k$  is

$$\hat{\gamma}_k = \frac{\hat{\mu}_{(X_k=+1)} - \hat{\mu}_{(X_k=-1)}}{(+1) - (-1)} = \frac{1}{2}(\hat{\mu}_{(X_k=+1)} - \hat{\mu}_{(X_k=-1)}).$$

Notice that  $\hat{\mu}_{(X_k=+1)}$  and  $\hat{\mu}_{(X_k=-1)}$  are independent, because we assume random assignment to clusters and no contamination. Therefore, the sampling variance of  $\hat{\gamma}_k$  is

$$\begin{aligned}\text{Var}(\hat{\gamma}_k) &= \text{Var}\left(\frac{1}{2}\hat{\mu}_{(X_k=+1)}\right) + \text{Var}\left(\frac{1}{2}\hat{\mu}_{(X_k=-1)}\right) \\ &= \frac{1}{4}\text{Var}(\hat{\mu}_{(X_k=+1)}) + \frac{1}{4}\text{Var}(\hat{\mu}_{(X_k=-1)}).\end{aligned}$$

By the  $\text{Var}(\cdot)$  operator here we implicitly mean  $\text{Var}(\cdot|\mathbf{X})$  (i.e., we are conditioning on the design). This is typical in linear models, especially for experiments with fixed effects factors assigned directly by the experimenter, such as those considered in the context of this paper.

To proceed further, we now assume balance in cell sizes and cluster sizes. Then, because the random effects terms in Model 5 are the same for every individual, and there is the same

number of individuals in every cell,  $\text{Var}(\hat{\mu}_{(X_k=+1)}) = \text{Var}(\hat{\mu}_{(X_k=-1)})$ , so  $\text{Var}(\hat{\gamma}_k)$  simplifies further to  $\text{Var}(\hat{\gamma}_k) = \frac{1}{2}\text{Var}(\hat{\mu}_{(X_k=+1)})$ . Two terms in Model 5 contribute to this variance: the  $u_j$  terms with variance  $\tau_u^2$  and the  $e_{ij}$  terms with variance  $\sigma_e^2$ . For purposes of computing this variance conditionally upon treatment assignment, the fixed effects terms do not matter, so we treat  $\text{Var}(\hat{\mu}_{(X_k=+1)})$  as the variance of the sum of the averages of the random terms. The cluster-level term  $u_j$  is averaged over  $J/2$  clusters having  $X_k = +1$ , where  $J$  is the total number of clusters. The individual-level term  $e_{ij}$  is averaged over  $nJ/2$  cluster members having  $X_k = +1$ , where  $n$  is the cluster size. In other words,  $\text{Var}(\hat{\mu}_{(X_k=+1)}) = \text{Var}\left(\frac{\sum_{i=1}^J u_j}{J/2}\right) + \text{Var}\left(\frac{\sum_{i=1}^{nJ} e_{ij}}{nJ/2}\right) = \frac{\tau_u^2}{J/2} + \frac{\sigma_e^2}{nJ/2}$ , such that

$$\text{Var}(\hat{\gamma}_k) = \frac{1}{2} \left( \frac{\tau_u^2}{J/2} + \frac{\sigma_e^2}{nJ/2} \right) = \frac{\tau_u^2}{J} + \frac{\sigma_e^2}{nJ},$$

as given in Table 4.

A similar argument can be made for interactions. Let  $\hat{\mu}_{(X_a=L_a, X_b=L_b)}$  be the average of all cell means having  $X_a = L_a$  and  $X_b = L_b$ . For example, if there are three factors in total, then we use  $\hat{\mu}_{(X_1=+1, X_3=+1)}$  to denote the average of the  $(+1, +1, +1)$  and  $(+1, -1, +1)$  cells. Then

$$\hat{\gamma}_{a,b} = \frac{\frac{\hat{\mu}_{(X_a=+1, X_b=+1)} - \hat{\mu}_{(X_a=+1, X_b=-1)}}{2} - \frac{\hat{\mu}_{(X_a=-1, X_b=+1)} - \hat{\mu}_{(X_a=-1, X_b=-1)}}{2}}{(+1) - (-1)}.$$

So

$$\begin{aligned} \text{Var}(\hat{\gamma}_{a,b}) &= \left(\frac{1}{4}\right)^2 \text{Var}(\hat{\mu}_{(X_a=+1, X_b=+1)}) + \left(\frac{1}{4}\right)^2 \text{Var}(\hat{\mu}_{(X_a=+1, X_b=-1)}) \\ &\quad + \left(\frac{1}{4}\right)^2 \text{Var}(\hat{\mu}_{(X_a=-1, X_b=+1)}) + \left(\frac{1}{4}\right)^2 \text{Var}(\hat{\mu}_{(X_a=-1, X_b=-1)}). \end{aligned}$$

Recall that  $\hat{\mu}_{(X_a=L_a)} = \frac{1}{2} \hat{\mu}_{(X_a=L_a, X_b=+1)} + \frac{1}{2} \hat{\mu}_{(X_a=L_a, X_b=-1)}$  (this is true by our definition even if cell sizes in the observed sample are not balanced). This implies that  $\text{Var}(\hat{\mu}_{(X_a=L_a)}) = \left(\frac{1}{2}\right)^2 \text{Var}(\hat{\mu}_{(X_a=L_a, X_b=+1)}) + \left(\frac{1}{2}\right)^2 \text{Var}(\hat{\mu}_{(X_a=L_a, X_b=-1)})$ . Thus,

$$\text{Var}(\hat{\gamma}_{a,b}) = \frac{1}{4} \text{Var}(\hat{\mu}_{(X_a=+1)}) + \frac{1}{4} \text{Var}(\hat{\mu}_{(X_a=-1)}) = \text{Var}(\hat{\gamma}_a).$$

This means that the same formula for the sampling distribution of the regression coefficient applies both to main effects and to two-way interactions here.

This finding implies that the test of an interaction will have the same power as the test of a main effect, if they have the same true effect size (expressed as an effect-coded regression coefficient). This seems to conflict with findings elsewhere (see Peterson & George, 1993; Murray, 1998), in which an interaction has less power than a main effect if they have the same true effect size, but there is no actual conflict. The latter result comes from a different metric, which treats main effects and interactions somewhat differently (see Appendix A in Dziak, Nahum-Shani, & Collins, 2012). Some authors consider the interaction to be

$$\frac{\hat{\mu}_{(X_a=+1, X_b=+1)} - \hat{\mu}_{(X_a=+1, X_b=-1)}}{2} - \frac{\hat{\mu}_{(X_a=-1, X_b=+1)} - \hat{\mu}_{(X_a=-1, X_b=-1)}}{2} = 2\hat{\gamma}_{a,b},$$

while others define it as an unscaled difference of differences:

$$(\hat{\mu}_{(X_a=+1, X_b=+1)} - \hat{\mu}_{(X_a=+1, X_b=-1)}) - (\hat{\mu}_{(X_a=-1, X_b=+1)} - \hat{\mu}_{(X_a=-1, X_b=-1)}) = 4\hat{\gamma}_{a,b}.$$

However, the effect-coded main effect is always expressed as, for example,  $2\hat{\gamma}_a$  rather than  $4\hat{\gamma}_a$ . If the interaction is “equal to the main effect” in terms of the former definition, then  $\hat{\gamma}_{a,b} = \hat{\gamma}_a$ ; but if they are equal in terms of the latter definition, it follows that  $\hat{\gamma}_{a,b} = \frac{1}{2}\hat{\gamma}_a$ . The ambiguous meaning of “equal size” here can lead to contradictory recommendations, which demonstrates the need for thoughtful consideration of the meaning of effect sizes in sample size planning.

Therefore, for both main effects and interactions, we have the variance formula

$\text{Var}(\hat{\gamma}_k) = \frac{\tau_u^2}{J} + \frac{\sigma_e^2}{nJ}$ . However, one remaining inconvenience is that it may be difficult to find a reasonable value for  $\tau_u^2$  and  $\sigma_e^2$  for power planning. Fortunately, it is only necessary to specify the posttest ICC, for the following reason. The covariance between two observations in the same cluster is  $\text{Cov}(Y_{ij}, Y_{ij'}) = \text{Cov}(u_j + e_{ij}, u_{j'} + e_{ij'}) = \tau_u^2$ . The total variance of one observation is  $\sigma_e^2 + \tau_u^2$ . This means that the intraclass correlation is  $\rho_Y = \frac{\tau_u^2}{\sigma_e^2 + \tau_u^2}$ . Therefore,  $\frac{\rho_Y}{1 - \rho_Y} = \frac{\tau_u^2}{\sigma_e^2}$ , such that

$$\frac{\tau_u^2}{J} + \frac{\sigma_e^2}{nJ} = \sigma_e^2 \left( \frac{\rho_Y}{(1 - \rho_Y)J} + \frac{1}{nJ} \right),$$

as in Table 4.

Researchers may be able to find realistic values for  $\rho_Y$  in their field of study by doing a literature review. Specifying  $\sigma_e^2$  can also be avoided if, as in Expression 8 one specifies  $\gamma$  as a standardized coefficient (i.e., as a multiple of  $\sigma_e$  rather than a raw score difference). Notice that in this setting, because there is no pretest,  $\sigma_e^2$  equals  $\sigma_Y^2$  in Expression 7 which is the posttest variance, adjusting for any cluster and treatment effects that may exist but not adjusting for pretest. Hence, when specifying a standardized coefficient,  $\sigma_e^2 = \sigma_Y^2$  cancels out and only  $\rho_Y$  is needed.

**Full EIC with Pretest as a Covariate** (Model 6;  $C_j \equiv 1$ ). Continuing to assume balance (i.e., that cluster sizes are equal and the number of individuals in each condition is equal), now let the pretest adjusted response be  $A_{ij} = Y_{ij} - \gamma_{10}P_{ij}$  and define  $\hat{\mu}_{(X_k=L)}^{(A)}$  to equal the sample estimate of  $E(A|X_k = L)$ . Then  $\text{Var}(\hat{\gamma}_k) \approx \text{Var}\left(\frac{1}{2}\left(\hat{\mu}_{(X_k=+1)}^{(A)} - \hat{\mu}_{(X_k=-1)}^{(A)}\right)\right)$ . The equality is not precise because  $\gamma_{10}$  must also be estimated and is not really a known constant; however, this is likely to be only a minor inaccuracy, similar to the way that estimation error in coefficients for

effects other than  $X_k$  was ignored in the previous derivations. Thus, the only remaining random terms are  $u_j$  and  $e_{ij}$ , and so, following the same logic as in the no-pretest scenario above, we have

$$\begin{aligned}\text{Var}(\hat{\gamma}_k) &\approx \text{Var}\left(\frac{1}{2}\left(\hat{\mu}_{(X_k=+1)}^{(A)} - \hat{\mu}_{(X_k=-1)}^{(A)}\right)\right) \\ &= \frac{1}{4}\left(\text{Var}\left(\hat{\mu}_{(X_k=+1)}^{(A)}\right) + \text{Var}\left(\hat{\mu}_{(X_k=-1)}^{(A)}\right)\right) \\ &= \frac{1}{2}\text{Var}\left(\hat{\mu}_{(X_k=+1)}^{(A)}\right) \\ &= \frac{1}{2}\left(\frac{\tau_u^2}{J} + \frac{\sigma_e^2}{nJ}\right) = \frac{\tau_u^2}{J} + \frac{\sigma_e^2}{nJ}.\end{aligned}$$

As above, this can be re-expressed in terms of correlations and marginal variances. Within a cluster, the total posttest variance without knowing the pretest (i.e., the variance conditioning on  $\mathbf{X}$  and  $u$  but not conditioning on  $P$ ) is  $\sigma_Y^2 = \gamma_{10}^2 \sigma_P^2 + \sigma_e^2$ , where  $\sigma_P^2$  is the variance of the pretest and  $\gamma_{10}$  is the egression coefficient of the pretest on the posttest. The pretest-posttest correlation conditioning on cluster membership will be

$$\rho_{\text{pre,post}} = \frac{\gamma_{10} \sigma_P^2}{\sigma_P \sigma_Y} = \frac{\gamma_{10} \sigma_P}{\sigma_Y},$$

so  $\sigma_e^2 = \sigma_Y^2 - \gamma_{10}^2 \sigma_P^2 = (1 - \rho_{\text{pre,post}}^2) \sigma_Y^2$ . Thus, at least in theory, using the pretest never hurts, although it is more helpful when the pretest-posttest correlation is larger. This agrees with the findings of Vickers (2001) in classic ANCOVA with independent data. Finally, the total posttest variance without knowing the pretest (i.e., adjusting for  $\mathbf{X}$  but not  $P$ ) is  $\sigma_Y^2 + \tau_u^2$ , so the posttest ICC is  $\rho_Y = \tau_u^2 / (\sigma_Y^2 + \tau_u^2)$ . Thus,  $\tau_u^2 = \sigma_Y^2 \rho_Y / (1 - \rho_Y)$ . Therefore,  $\text{Var}(\hat{\gamma}_k)$  can be re-expressed as

$$\text{Var}(\hat{\gamma}_k) = \frac{\tau_u^2}{J} + \frac{\sigma_e^2}{nJ} = \sigma_Y^2 \left( \frac{\rho_Y}{(1-\rho_Y)J} + \frac{1-\rho_{\text{pre,post}}^2}{nJ} \right), \quad (17)$$

as in Table 4.

Recall that without a pretest,  $\frac{\tau_u^2}{J} + \frac{\sigma_e^2}{n_J} = \sigma_Y^2 \left( \frac{\rho_Y}{(1-\rho_Y)J} + \frac{1}{n_J} \right)$ . Intuitively, this is the same as Expression 17 but with  $\rho_{\text{pre},\text{post}} = 0$ . More precisely, the reason why the formulas differ when expressed in terms of correlations, but appear not to differ when expressed in terms of variance components, is that  $\sigma_e^2$  has a slightly different interpretation when a pretest is present. When a pretest is not included,  $\sigma_e^2$  is the individual-level error in the posttest response. When a pretest is included,  $\sigma_e^2$  is the individual-level error in the pretest-adjusted posttest response.

In the full-EIC case, Model 6 appears essentially identical to the pretest-adjusted model for between-clusters experiments in which clusters exist prior to experimentation, as given in Dziak, Nahum-Shani, and Collins (2012). This is because, although the model implies that the posttest scores are clustered, it does not explicitly specify whether the pretest scores are clustered or not. However, in our power formula derivation we do use the assumption that pretest scores are unclustered. Dziak and colleagues (2012) were not able to provide an exact power formula in their context, because the relationship of the pretest cluster-level variability to the posttest cluster-level variability is important for power but is not specified in the analysis model. Dziak and colleagues (2012) therefore had to provide an approximate formula based on a related model (specifically, a three-level model in which pretest is included as a repeated measure). In the context of this paper, this limitation is removed because in the EIC setting it is assumed that pretest scores are unclustered.

Therefore, deriving a power formula was more straightforward for factorial designs with full EIC than for between-clusters factorial experiments. Deriving a power formula for partial EIC is slightly more complicated because in this setting some individuals have a different variance structure from others. However, as we explain below, a power formula can be derived by adapting the reasoning used earlier in the case of full EIC.

**Partial EIC, No Pretest** (Model 5;  $C_j$  is 0 or 1 depending on  $X_1$ ). We now continue to the partial EIC (partial nesting) scenario with no pretest. We need to determine variance formulas for the main effect coefficients  $\gamma_k$ , both for  $k=1$  (the cluster-generating factor) and  $k>1$  (the remaining factors). First consider  $k=1$ . Then  $\text{Var}(\hat{\gamma}_k) = \text{Var}\left(\frac{1}{2}(\hat{\mu}_{(X_1=+1)} - \hat{\mu}_{(X_1=-1)})\right)$ .

Therefore,

$$\text{Var}(\hat{\gamma}_k) = \left(\frac{1}{2}\right)^2 \text{Var}(\hat{\mu}_{(X_1=+1)}) + \left(\frac{1}{2}\right)^2 \text{Var}(\hat{\mu}_{(X_1=-1)}).$$

Analogously to the full-EIC case,  $\text{Var}(\hat{\mu}_{(X_1=+1)}) = \frac{\tau_u^2}{J_1} + \frac{\sigma_{e1}^2}{J_1 n}$ , where  $J_1$  is the number of nontrivial clusters, and  $n$  is the cluster size. However, cluster-level variability does not apply to those individuals with  $X_1 = -1$ , so we simply have  $\text{Var}(\hat{\mu}_{(X_1=-1)}) = \frac{\sigma_{e0}^2}{J_0}$ , where  $J_0$  is the number of unclustered individuals (trivial clusters). Combining these, we have

$$\text{Var}(\hat{\gamma}_k) = \frac{\tau_u^2}{4J_1} + \frac{\sigma_{e0}^2}{4J_0} + \frac{\sigma_{e1}^2}{4J_1 n},$$

as in Table 5.

Superficially comparing this to the corresponding full-EIC formula, the reader may notice a "4" in the denominator of this expression that was not found in Expression 17. This is simply a consequence of the use of different mathematical expressions (e.g.,  $J_1$  and  $J_0$  instead of  $J/2$  and  $J/2$  for the number of clusters per factor level) and is not necessarily an indication of more precise results for one design relative to the other.

Now suppose that the main effect of one of the other factors, say  $X_k$ , is being tested. At first, this appears to be a different case from testing  $X_1$ , because the variance structure depends only on  $X_1$  and not on  $X_k$ . However,  $\hat{\mu}_{(X_k=x_k)} = \frac{1}{2}\hat{\mu}_{(X_k=x_k, X_1=+1)} + \frac{1}{2}\hat{\mu}_{(X_k=x_k, X_1=-1)}$  and  $\text{Var}(\hat{\mu}_{(X_k=x_k)}) = \frac{1}{4}\text{Var}(\hat{\mu}_{(X_k=x_k, X_1=+1)}) + \frac{1}{4}\text{Var}(\hat{\mu}_{(X_k=x_k, X_1=-1)})$  for both  $x_k = +1$  and  $x_k = -1$ .

-1. Because we define the main effect as an average of cell means, not individuals, we *do not need* to assume either equal variances or balance between levels of  $X_1$  in order to conclude this.

Therefore, even if  $k \neq 1$ ,

$$\begin{aligned}\text{Var}(\hat{\gamma}_k) &= \frac{1}{4}\text{Var}(\hat{\mu}_{(X_k=+1)}) + \frac{1}{4}\text{Var}(\hat{\mu}_{(X_k=-1)}) \\ &= \frac{1}{16}\text{Var}(\hat{\mu}_{(X_1=+1, X_k=+1)}) + \frac{1}{16}\text{Var}(\hat{\mu}_{(X_1=+1, X_k=-1)}) \\ &\quad + \frac{1}{16}\text{Var}(\hat{\mu}_{(X_1=-1, X_k=+1)}) + \frac{1}{16}\text{Var}(\hat{\mu}_{(X_1=-1, X_k=-1)}) \\ &= \frac{1}{4}\text{Var}(\hat{\mu}_{(X_1=+1)}) + \frac{1}{4}\text{Var}(\hat{\mu}_{(X_1=-1)}) \\ &= \text{Var}(\hat{\gamma}_1).\end{aligned}$$

Thus, the variance for each main effect is equal to the variance for the first main effect. A similar argument can be made that the variance for each interaction is equal to the variance for the first main effect, although the algebra would be slightly more involved, generally requiring eight terms of the form  $\hat{\mu}_{(X_1, X_a, X_b)}$  rather than four terms of the form  $\hat{\mu}_{(X_1, X_k)}$ . Therefore, from here on we consider only  $\text{Var}(\hat{\gamma}_1)$ , because the theoretical sampling variance for the other coefficients will be the same.

**Partial EIC with Pretest as Covariate** (Model 6;  $C_j$  is 0 or 1 depending on  $X_1$ ). Finally, we consider ANCOVA with partial EIC and the adjusted outcomes  $A_{ij}$  as before. The variance

of interest is  $\text{Var}(\hat{\gamma}_k) = \frac{1}{4}\text{Var}(\hat{\mu}_{(X_1=+1)}^{(A)}) + \frac{1}{4}\text{Var}(\hat{\mu}_{(X_1=-1)}^{(A)})$ , and the random effects part of  $A_{ij}$  is  $C_j u_j + e_{ij}$ , so  $\text{Var}(\hat{\mu}_{(X_1=+1)}^{(A)}) = \frac{\tau_u^2}{J_1} + \frac{\sigma_{e1}^2}{n J_1}$ , and  $\text{Var}(\hat{\mu}_{(X_1=-1)}^{(A)}) = \frac{\sigma_{e0}^2}{J_0}$ .

Therefore,

$$\text{Var}(\hat{\gamma}_k) = \frac{\tau_u^2}{4J_1} + \frac{\sigma_{e0}^2}{4J_0} + \frac{\sigma_{e1}^2}{4J_1 n}, \tag{18}$$

as in Table 5.

In order to find a way to express this formula in terms of correlations, we make the simplifying assumption that  $\sigma_{e0}^2 = \sigma_{e1}^2 = \sigma_e^2$ . Then the posttest variance, after adjusting for cluster and treatment, is  $\sigma_Y^2 = \gamma_{10}^2 \sigma_P^2 + \sigma_e^2$ . The pretest-posttest correlation is  $\rho_{\text{pre,post}} = \gamma_{10} \sigma_P / \sigma_Y$  as before, so  $\sigma_Y^2(1 - \rho_{\text{pre,post}}) = \sigma_Y^2 - \gamma_{10}^2 \sigma_P^2 = \sigma_e^2$ . Also, as before, the posttest intraclass correlation is  $\rho_Y = \tau_u^2 / (\sigma_Y^2 + \tau_u^2)$ , so  $\tau_u^2 = \sigma_Y^2 \rho_Y / (1 - \rho_Y)$ . We conclude that

$$\text{Var}(\hat{\gamma}_k) = \sigma_Y^2 \left( \frac{\rho_Y}{4(1-\rho_Y)J_1} + \frac{1-\rho_{\text{pre,post}}^2}{4J_1 n} + \frac{1-\rho_{\text{pre,post}}^2}{4J_0} \right),$$

as in Table 5. Note that this holds only under the assumption of equal error variances, and Expression 18 is more general.

Under the same assumptions, for the no-pretest scenario, we have (as in Table 5)

$$\text{Var}(\hat{\gamma}_k) = \sigma_Y^2 \left( \frac{\rho_Y}{4(1-\rho_Y)J_1} + \frac{1}{4J_1 n} + \frac{1}{4J_0} \right).$$

## Appendix C

### Sample SAS Code for Modeling Experiments with Full and Partial EIC

Below is the SAS code used for Model 11 in the context of Simulation Study 1. Here  $j$  represents the cluster,  $x_1$  through  $x_5$  represent the factors,  $P_{ij}$  represents the pretest, and  $Y_{ij}$  represents the posttest.

```
PROC MIXED DATA=wide NOCLPRINT;
   CLASS x1 x2 x3 x4 x5 j;
   MODEL Yij = Pij x1 x2 x3 x4 x5
              x1*x2 x1*x3 x1*x4 x1*x5
              x2*x3 x2*x4 x2*x5
              x3*x4 x3*x5
              x4*x5 / DDFM=SATTERTHWAITE;
   RANDOM INTERCEPT / SUBJECT = j(x1 x2 x3 x4 x5);
   ODS OUTPUT TESTS3=OutputAncova COVPARMS=cp;
QUIT;
```

Below is the SAS code used for Model 12 in the context of Simulation Study 2. The variables have the same meaning as above, except that  $j$  can now identify either a real cluster or a trivial cluster (i.e., unclustered individual).

```
PROC MIXED DATA=wide NOCLPRINT;
   CLASS x1 x2 x3 x4 x5 j;
   MODEL Yij = Pij x1 x2 x3 x4 x5
              x1*x2 x1*x3 x1*x4 x1*x5
              x2*x3 x2*x4 x2*x5
              x3*x4 x3*x5
              x4*x5 / DDFM=SATTERTHWAITE;
   RANDOM clustered / SUBJECT = j(x1 x2 x3 x4 x5);
   REPEATED / SUB=i LOCAL=EXP(clustered) TYPE=VC; /* Remove this
REPEATED statement in order to assume equal error variances*/
   ODS OUTPUT TESTS3=OutputAncova COVPARMS=cp;
QUIT;
```

## References

- Baker, T. B., Gustafson, D. H., & Shah, D. (2014). How can research keep up with eHealth? Ten strategies for increasing the timeliness and usefulness of eHealth research. *Journal of Medical Internet Research*, 16(2), e36.
- Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, 16(2), 149-165.
- Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research*, 43, 210–236.  
doi:10.1080/00273170802034810
- Berggraf, L., Ulvenes, P. G., Øktedalen, T., Hoffart, A., Stiles, T., McCullough, L., & Wampold, B. E. (2014). Experience of affects predicting sense of self and others in short-term dynamic and cognitive therapy. *Psychotherapy*, 51(2), 246.
- Candel, M. J., & Van Breukelen, G. J. (2009). Varying cluster sizes in trials with clusters in one treatment arm: Sample size adjustments when testing treatment effects with linear mixed models. *Statistics in Medicine*, 28(18), 2307-2324.
- Chakraborty, B., Collins, L. M., Strecher, V. J., & Murphy, S. A. (2009). Developing multicomponent interventions using fractional factorial designs. *Statistics in Medicine*, 28, 2687–2708.
- Charlesworth, G., Burnell, K., Beecham, J., Hoare, Z., Hoe, J., Wenborn, J., ... & Orrell, M. (2011). Peer support for family carers of people with dementia, alone or in combination with group reminiscence in a factorial design: study protocol for a randomized controlled trial. *Trials*, 12(1), 1.

- Chebli, J. L., Blaszczynski, A., & Gainsbury, S. M. (2016). Internet-based interventions for addictive behaviours: a systematic review. *Journal of Gambling Studies*, 1-26.
- Cloitre, M., Koenen, K. C., Cohen, L. R., & Han, H. (2002). Skills training in affective and interpersonal regulation followed by exposure: A phase-based treatment for PTSD related to childhood abuse. *Journal of Consulting and Clinical Psychology*, 70(5), 1067-1074.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Collins, L. M., Dziak, J. J., Kugler, K. C., & Trail, J. B. (2014). Factorial experiments: Efficient tools for evaluation of intervention components. *American Journal of Preventive Medicine*, 47(4), 498-504.
- Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods*, 14, 202-224.
- Collins, L. M., Kugler, K. C., & Gwadz, M. V. (2016). Optimization of multicomponent behavioral and biobehavioral interventions for the prevention and treatment of HIV/AIDS. *AIDS and Behavior*, 20(1), 197-214.
- Collins, L. M., Nahum-Shani, I., & Almirall, D. (2014). Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). *Clinical Trials*, 11(4): 426-434.
- Cook, J. W., Collins, L. M., Fiore, M. C., Smith, S. S., Fraser, D., Bolt, D. M., ... & Loh, W. Y. (2016). Comparative effectiveness of motivation phase intervention components for use with smokers unwilling to quit: a factorial screening experiment. *Addiction*, 111(1), 117-128.

- Crespi, C. M. (2016). Improved designs for cluster randomized trials. *Annual Review of Public Health, 37*: 1-16. doi:10.1146/annurev-publhealth-032315-021702
- Czajkowski, S. M., Powell, L. H., Adler, N., Naar-King, S., Reynolds, K. D., Hunter, C. M., ... & Epel, E. (2015). From ideas to efficacy: The ORBIT model for developing behavioral treatments for chronic diseases. *Health Psychology, 34*(10), 971.
- Dallery, J., Riley, W.T., & Nahum-Shani, I. (2015). Research Designs to Develop and Evaluate Technology-Based Health Behavior Interventions. To appear in L. Marsch, S. Lord, and J. Dallery (Eds), *Leveraging Technology to Transform Behavioral Healthcare*. Oxford University Press.
- De Jong, K., Moerbeek, M., and Van der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects. *Psychotherapy Research, 20*, 273-284.
- Derlega, V. J., Winstead, B. A., Wong, P. T., & Hunter, S. (1985). Gender effects in an initial encounter: A case where men exceed women in disclosure. *Journal of Social and Personal Relationships, 2*(1), 25-44.
- Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London, England: Arnold.
- Dziak, J. J., & Nahum-Shani, I. (2016). Three-level modeling for factorial experiments with experimentally induced clustering (Technical Report No. 16-133). University Park, PA: The Methodology Center, Penn State.
- Dziak, J. J., Nahum-Shani, I., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. *Psychological Methods, 17*2, 153-175.

- Erez, M., & Arad, R. (1986). Participative goal-setting: Social, motivational, and cognitive factors. *Journal of Applied Psychology*, 71(4), 591.
- Gainsbury, S., & Blaszczynski, A. (2011). A systematic review of Internet-based therapy for the treatment of addictions. *Clinical Psychology Review*, 31(3), 490-498.
- Herbert, J. D., Gaudiano, B. A., Rheingold, A. A., Moitra, E., Myers, V. H., Dalrymple, K. L., & Brandsma, L. L. (2009). Cognitive behavior therapy for generalized social anxiety disorder in adolescents: A randomized controlled trial. *Journal of Anxiety Disorders*, 23, 167-177.
- Henry, K., & Dziak, J. J. (2016). A nested covariate approach to the inclusion of two-part predictors in a regression model. Manuscript submitted for publication.
- Hox, J. J., & Kreft, I. G. (1994). Multilevel analysis methods. *Sociological Methods & Research*, 22(3), 283-299.
- Howard, M. C., & Jacobs, R. R. (2016). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): two novel evaluation methods for developing optimal training programs. *Journal of Organizational Behavior* (online version).
- Jacobs, M. A., & Graham, A. L. (2016). Iterative development and evaluation methods of mHealth behavior change interventions. *Current Opinion in Psychology*, 9, 33-37.
- Karakowsky, L., & McBey, K. (2001). Do my contributions matter? The influence of imputed expertise on member involvement and self-evaluations in the work group. *Group & Organization Management*, 26(1), 70-92.

- Kasari, C., Rotheram-Fuller, E., Locke, J., & Gulsrud, A. (2012). Making the connection: randomized controlled trial of social skills at school for children with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, 53, 431-439.
- Kenny, D. A., Bolger, N., & Kashy, D. A. (2002). Traditional methods for estimating multilevel models. In D. S. Moskowitz & S. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Method and applications*. Englewood Cliffs, NJ: Erlbaum (pp. 1-24).
- Kirk, R. (2003). *Experimental design: Procedures for the behavioral sciences*. Los Angeles, CA: SAGE.
- Kramer, T. J., Fleming, G. P., & Mannis, S. M. (2001). Improving face-to-face brainstorming through modeling and facilitation. *Small Group Research*, 32(5), 533-557.
- Kugler, K.C., Trail, J.B., Dziak, J.J., & Collins, L.M. (2012). Effect coding versus dummy coding in analysis of data from factorial experiments (No. 12-120). University Park, PA: The Methodology Center, Pennsylvania State University. Accessed at <http://methodology.psu.edu/media/techreports/12-120.pdf>
- Lecomte, T., Leclerc, C., Corbiere, M., Wykes, T., Wallace, C. J., & Spidel, A. (2008). Group Cognitive Behavior Therapy or Social Skills Training for Individuals With a Recent Onset of Psychosis?: Results of a Randomized Controlled Trial. *Journal of Nervous and Mental Disease*, 196, 866-875.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schaabenberger, O. (2006). *SAS(R) for Mixed Models* (2nd ed). Cary, NC: SAS Institute Inc.

- Moerbeek, M., & Wong, W. K. (2008). Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine*, 27(15), 2850-2864. doi: 10.1002/sim.3115.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials* (Vol. 29). Oxford University Press.
- Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments*. Hoboken, NJ: Wiley.
- Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Erlbaum.
- Nackers, L. M., Dubyak, P. J., Lu, X., Anton, S. D., Dutton, G. R., & Perri, M. G. (2015). Group dynamics are associated with weight loss in the behavioral treatment of obesity. *Obesity*, 23(8), 1563-1569.
- Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin*, 133, 673-93.
- Nye, J. L. (2002). The eye of the follower information processing effects on attributions regarding leaders of small groups. *Small Group Research*, 33(3), 337-360.
- Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs. The impact of change-score versus ANCOVA models. *Evaluation Review*, 25, 3-28.

- Pals, S. L., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., & Baker, W. L. (2008). Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. *American Journal of Public Health, 98*, 1418–1424.
- Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2014). Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol. *Contemporary Clinical Trials, 38*(2), 251-259.
- Pellegrini, C.A., Hoffman, S.A., Collins, L.M., & Spring, B. (2015). Corrigendum to Optimization of remotely delivered intensive lifestyle treatment for obesity using the multiphase optimization strategy: Opt-IN study protocol. *Contemporary Clinical Trials, 45*, 468-469.
- Peters, G. J. Y., de Bruin, M., & Crutzen, R. (2015). Everything should be as simple as possible, but no simpler: towards a protocol for accumulating evidence regarding the active content of health behaviour change interventions. *Health Psychology Review, 9*(1), 1-14.
- Peterson, B., & George, S. L. (1993). Sample size requirements and length of study for testing interaction in a  $2 \times k$  factorial design when time-to-failure is the outcome (corrected). *Controlled Clinical Trials, 14*, 511-522. See also erratum in *Controlled Clinical Trials, 15*: 326.
- Peterson, C. B., Mitchell, J. E., Crow, S. J., Crosby, R. D., & Wonderlich, S. A. (2009). The efficacy of self-help group treatment and therapist-led group treatment for binge eating disorder. *American Journal of Psychiatry, 166*, 1347-54.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173–185.

- Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, 2, 153–162.
- Schulz, M. S., Cowan, C. P., & Cowan, P. A. (2006). Promoting healthy beginnings: A randomized controlled trial of a preventive intervention to preserve marital quality during the transition to parenthood. *Journal of Consulting and Clinical Psychology*, 74(1), 20-31.
- Slymen, D. J., & Hovell, M. F. (1997). Cluster versus individual randomization in adolescent tobacco and alcohol studies: illustrations for design decisions. *International Journal of Epidemiology*, 26(4), 765-771.
- Tokola, K., Larocque, D., Nevalainen, J., & Oja, H. (2011). Power, sample size and sampling costs for clustered data. *Statistics & Probability Letters*, 81, 852-860.
- Valacich, J. S., Wheeler, B. C., Mennecke, B. E., & Wachter, R. (1995). The effects of numerical and logical group size on computer-mediated idea generation. *Organizational Behavior and Human Decision Processes*, 62(3), 318-329.
- Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Medical Research Methodology*, 1:6. Accessed at <http://www.biomedcentral.com/1471-2288/1/6>
- Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research*, 16, 184-187.
- Wilson, D. K., Kitzman-Ulrich, H., Resnicow, K., Van Horn, M. L., George, S. M. S., Siceloff, E. R., ... & Coulon, S. (2015). An overview of the Families Improving Together (FIT) for weight loss randomized controlled trial in African American families. *Contemporary Clinical Trials*, 42, 145-157.

Wu, C. F. J., & Hamada, M. (2011). *Experiments: Planning, analysis, and parameter design optimization* (Vol. 552). New York, NY: Wiley,

Table 1  
Experimental Conditions in the Hypothetical  $2 \times 2 \times 2$  Factorial Design

Experimental condition	Brief description	V (Video)	T (Texts)	M (Meals)
1	Untreated	Off	Off	Off
2	M only	Off	Off	On
3	T only	Off	On	Off
4	M and T	Off	On	On
5	V only	On	Off	Off
6	V and M	On	Off	On
7	V and T	On	On	Off
8	All three	On	On	On

Table 2

Effect Coding for a  $2 \times 2 \times 2$  Factorial Design in the Weight Loss Program Example

Condition	Brief description	Main Effects				Interactions		
		V (Video)	T (Texts)	M (Meals)	V×T	V×M	T×M	V×T×M
Conditions in Complete Factorial								
1	Untreated	-1	-1	-1	+1	+1	+1	-1
2	M only	-1	-1	+1	+1	-1	-1	+1
3	T only	-1	+1	-1	-1	+1	-1	+1
4	M and T	-1	+1	+1	-1	-1	+1	-1
5	V only	+1	-1	-1	-1	-1	+1	+1
6	V and M	+1	-1	+1	-1	+1	-1	-1
7	V and T	+1	+1	-1	+1	-1	-1	-1
8	All three	+1	+1	+1	+1	+1	+1	+1
Conditions Retained in a Fractional Factorial								
2	M only	-1	-1	+1	+1	-1	-1	+1
3	T only	-1	+1	-1	-1	+1	-1	+1
5	V only	+1	-1	-1	-1	-1	+1	+1
8	All three	+1	+1	+1	+1	+1	+1	+1

Table 3  
Conditions in a Factorial Experiment with Full EIC and a Factorial Experiment with Partial EIC

Condition	Full EIC	Partial EIC	Factors		
			Clustered?	$X_1$	$X_2$
1	Yes	No	Off	Off	Off
2	Yes	No	Off	Off	On
3	Yes	No	Off	On	Off
4	Yes	No	Off	On	On
5	Yes	Yes	On	Off	Off
6	Yes	Yes	On	Off	On
7	Yes	Yes	On	On	Off
8	Yes	Yes	On	On	On

**Notes.** The table assumes that there are a total of three factors arranged in a complete factorial design. In the weight loss examples given in the text,  $X_2$ , and  $X_3$  represent the Text and Meals factors.  $X_1$  represents the Video factor in the full-EIC factorial example and the Support factor in the partial EIC factorial example.

Table 4  
Sampling Variances for Regression Coefficients in Factorial Designs with Full EIC

Pretest	Variance of regression coefficient for effect
No (Model 5, $C_j \equiv 1$ )	$\text{Var}(\hat{\gamma}) = \frac{\tau_u^2}{J} + \frac{\sigma_e^2}{Jn} = \sigma_Y^2 \left( \frac{\rho_Y}{(1 - \rho_Y)J} + \frac{1}{Jn} \right)$ <p>where <math>\sigma_Y^2 = \sigma_e^2</math></p>
Yes (Model 6, $C_j \equiv 1$ )	$\text{Var}(\hat{\gamma}) = \frac{\tau_u^2}{J} + \frac{\sigma_e^2}{Jn} = \sigma_Y^2 \left( \frac{\rho_Y}{(1 - \rho_Y)J} + \frac{1 - \rho_{\text{pre,post}}^2}{Jn} \right)$ <p>where <math>\sigma_Y^2 = \gamma_P^2 \sigma_P^2 + \sigma_e^2</math></p>

**Note.** The "variance of regression coefficient for effect" column shows how to calculate  $\text{Var}(\hat{\gamma})$  for expressions (7), (8), or (9) from the variance components in the context of each model.  $J$  represents the total number of clusters summed across all conditions, and  $n$  is the number of members per cluster. In Model 5,  $\sigma_e^2$  is the individual-level error variance, and  $\tau_u^2$  is the cluster-level variance. In Model 6,  $\gamma_P$  is the pretest effect on posttest;  $\sigma_P^2$  is the variance of the pretest;  $\sigma_e^2$  is the individual-level error variance after adjusting for pretest; and  $\tau_u^2$  is the cluster-level variance in posttest after adjusting for pretest. In each case,  $\sigma_Y^2$  represents the total posttest variance after adjusting for any cluster or treatment effects but not adjusting for pretest;  $\rho_Y$  is the ICC at posttest; and  $\rho_{\text{pre,post}}$  is the pretest-posttest correlation after adjusting for treatment and cluster membership (i.e., the within-person ICC). Note that  $\gamma_P$  is written as  $\gamma_{10}$  in Model 6, but it seemed clearer to write it as  $\gamma_P$  here to emphasize that it represents the pretest and not the first factor.

Table 5

Sampling Variances for Regression Coefficients in Factorial Designs with Partial EIC

Pretest	Variance of regression coefficient for effect
No (Model 5; $C_j = 1$ if $X_1 = 1$ , $C_j = 0$ if $X_1 = -1$ )	$\text{Var}(\hat{\gamma}) = \frac{\tau_u^2}{4J_1} + \frac{\sigma_{e0}^2}{4J_0} + \frac{\sigma_{e1}^2}{4J_1 n}$ $= \sigma_Y^2 \left( \frac{\rho_Y}{4(1-\rho_Y)J_1} + \frac{1}{4J_1 n} + \frac{1}{4J_0} \right)$ <p style="text-align: center;">if <math>\sigma_{e0}^2 = \sigma_{e1}^2</math>, where <math>\sigma_Y^2 = \sigma_e^2</math></p>
Yes (Model 6; $C_j = 1$ if $X_1 = 1$ , $C_j = 0$ if $X_1 = -1$ )	$\text{Var}(\hat{\gamma}) = \frac{\tau_u^2}{4J_1} + \frac{\sigma_{e0}^2}{4J_0} + \frac{\sigma_{e1}^2}{4J_1 n}$ $= \sigma_Y^2 \left( \frac{\rho_Y}{4(1-\rho_Y)J_1} + \frac{1 - \rho_{\text{pre},\text{post}}^2}{4J_1 n} + \frac{1 - \rho_{\text{pre},\text{post}}^2}{4J_0} \right)$ <p style="text-align: center;">if <math>\sigma_{e0}^2 = \sigma_{e1}^2</math>, where <math>\sigma_Y^2 = \gamma_P^2 \sigma_P^2 + \sigma_e^2</math></p>

**Note:** The "variance of regression coefficient for effect" column shows how to calculate  $\text{Var}(\hat{\gamma})$  for expressions (7), (8), or (9) from the variance components in the context of each model.  $J_1$ ,  $J_0$  and  $n$  represent the number of clusters, number of unclustered individuals, and number of members per cluster respectively, so that  $J_1 n$  is the number of clustered individuals. In Model 5,  $\sigma_{e0}^2$  and  $\sigma_{e1}^2$  are the individual-level variances for unclustered and clustered individuals, both designated as  $\sigma_e^2$  if they are assumed equal; and  $\tau_u^2$  is the cluster-level variance. In Model 6,  $\gamma_P$  is the pretest effect on posttest;  $\sigma_P^2$  is the variance of pretest;  $\sigma_{e0}^2$  and  $\sigma_{e1}^2$  are the individual-level error variances after adjusting for pretest; and  $\tau_u^2$  is the cluster-level variance for the clustered conditions after adjusting for pretest. In each case,  $\sigma_Y^2$  represents the total posttest variance after adjusting for any cluster or treatment effects but not adjusting for pretest;  $\rho_Y$  is the posttest ICC for clustered individuals, adjusting for treatment; and  $\rho_{\text{pre},\text{post}}$  represents the pretest-posttest correlation after adjusting for treatment for unclustered individuals, and the pretest-posttest correlation after adjusting for treatment and cluster membership for clustered individuals. Note that  $\gamma_P$  is written as  $\gamma_{10}$  in Model 6, but it seemed clearer to write it as  $\gamma_P$  here to emphasize that it represents the pretest and not the first factor.

Table 6  
Scenarios in Simulation Study 1

Independent variable	Levels
Design of experiment	Complete factorial or fractional (half) factorial
Number of individuals	300 or 400 or 500 or 600
Size of each cluster (before dropout)	5 or 10
True effect size	0 or 0.10 or 0.15 or 0.25
True intraclass correlation	0.10 or 0.20

Table 7  
Power Estimates in Simulation Study 1

$N$	$n_i$	$d = .2$			$d = .3$			$d = .5$		
		Complete	Fractional	Predicted	Complete	Fractional	Predicted	Complete	Fractional	Predicted
<i>ICC = .1</i>										
300	5	0.29	0.30	0.32	0.55	0.59	0.61	0.94	0.96	0.96
300	10	—	0.18	0.22	—	0.36	0.43	—	0.76	0.84
400	5	0.40	0.41	0.41	0.72	0.73	0.74	0.99	0.99	0.99
400	10	0.27	0.27	0.29	0.55	0.52	0.56	0.93	0.93	0.94
500	5	0.49	0.49	0.50	0.82	0.82	0.83	1.00	1.00	1.00
500	10	0.34	0.35	0.36	0.63	0.67	0.67	0.97	0.98	0.98
600	5	0.57	0.56	0.57	0.88	0.89	0.90	1.00	1.00	1.00
600	10	0.40	0.40	0.42	0.70	0.73	0.76	0.99	0.99	0.99
<i>ICC = .2</i>										
300	5	0.20	0.21	0.23	0.39	0.42	0.44	0.81	0.84	0.85
300	10	—	0.13	0.15	—	0.21	0.27	—	0.52	0.61
400	5	0.28	0.28	0.29	0.54	0.55	0.56	0.93	0.94	0.94
400	10	0.18	0.18	0.19	0.34	0.35	0.36	0.74	0.73	0.76
500	5	0.33	0.35	0.35	0.65	0.64	0.66	0.97	0.97	0.98
500	10	0.21	0.22	0.23	0.41	0.44	0.44	0.83	0.86	0.86
600	5	0.41	0.41	0.41	0.72	0.74	0.74	0.99	0.99	0.99
600	10	0.25	0.25	0.27	0.47	0.50	0.52	0.89	0.90	0.92

**Notes.**  $N$  denotes the total sample size before dropout,  $n_i$  the cluster size before dropout (so that  $N/n_i$  is the number of clusters), "Complete," "Fractional," and "Predicted" refer to the simulated power for the complete factorial, the simulated power for the fractional factorial, and the predicted power from the proposed formula.  $d$  refers to the Cohen's  $d$  for a main effect (i.e., twice the  $\gamma$  parameter), and *ICC* refers to posttest intraclass correlation. Impossible designs are marked with a dash.

Table 8  
Scenarios in Simulation Study 2

Independent variables	Levels
Design of experiment	Complete factorial or fractional (half) factorial
Number of individuals	300 or 400 or 500 or 600
Error variances ratio between unclustered and clustered individuals	1:1, or 2:1
True effect size	0 or 0.15
True intraclass correlation	0.10 or 0.20
Allocation proportion	50% or 60% or 70%

Note. The cluster size (before dropout) was fixed at 5, unlike in Simulation Study 1.

Table 9  
Power Estimates in Simulation Study 2

<i>N</i>	Alloc.	Equal variances			Unequal variances		
		Complete	Fractional	Predicted	Complete	Fractional	Predicted
<i>ICC</i> = .1							
300	50%	0.69	0.69	0.67	0.71	0.72	0.70
	60%	0.72	0.71	0.70	0.73	0.72	0.70
	70%	0.69	0.69	0.68	0.66	0.68	0.65
400	50%	0.82	0.82	0.82	0.84	0.86	0.84
	60%	0.84	0.85	0.83	0.84	0.84	0.83
	70%	0.81	0.82	0.81	0.79	0.79	0.78
500	50%	0.91	0.91	0.90	0.92	0.93	0.92
	60%	0.91	0.91	0.91	0.91	0.91	0.91
	70%	0.89	0.89	0.89	0.87	0.87	0.87
600	50%	0.95	0.95	0.95	0.96	0.96	0.96
	60%	0.95	0.96	0.95	0.95	0.96	0.95
	70%	0.94	0.94	0.94	0.92	0.93	0.93
<i>ICC</i> = .2							
300	50%	0.55	0.57	0.55	0.59	0.61	0.58
	60%	0.61	0.61	0.59	0.61	0.63	0.61
	70%	0.59	0.60	0.59	0.57	0.60	0.58
400	50%	0.71	0.72	0.70	0.75	0.75	0.74
	60%	0.73	0.74	0.73	0.77	0.78	0.75
	70%	0.73	0.73	0.73	0.73	0.73	0.72
500	50%	0.81	0.80	0.80	0.84	0.84	0.84
	60%	0.84	0.84	0.83	0.85	0.85	0.84
	70%	0.82	0.82	0.82	0.81	0.82	0.81
600	50%	0.87	0.87	0.87	0.90	0.91	0.90
	60%	0.89	0.90	0.89	0.90	0.91	0.90
	70%	0.88	0.88	0.89	0.87	0.87	0.88

**Note.** *N* denotes total sample size before dropout. "Alloc." refers to the proportion of the sample assigned to clusters. "Complete," "Fractional," and "Predicted" refer to simulated power for complete factorial, simulated power for fractional factorial, and predicted power from the proposed formula. *ICC* refers to posttest ICC. In each case, the effect size is *d* = .3 (where *d* refers to the Cohen's *d* for a main effect, which is twice the  $\gamma$  parameter) and the pre-dropout cluster size is  $n_i = 5$ .