Multilevel Factorial Experiments for Developing Behavioral Interventions: Power, Sample Size, and Resource Considerations

Article *in* Psychological Methods · February 2012

DOI: 10.1037/a0026972 · Source: PubMed

John Joseph Dziak
Pennsylvania State University
32 PUBLICATIONS **604** CITATIONS

Linda M Collins
Pennsylvania State University
135 PUBLICATIONS **8,798** CITATIONS

Published in final edited form as:
Psychol Methods. 2012 June ; 17(2): 153–175. doi:10.1037/a0026972.

Multilevel Factorial Experiments for Developing Behavioral Interventions: Power, Sample Size, and Resource Considerations[†]

John J. Dziak,
The Methodology Center, The Pennsylvania State University, University Park, PA

Inbal Nahum-Shani, and
Institute for Social Research, University of Michigan, Ann Arbor, MI

Linda M. Collins
The Methodology Center and the Department of Human Development and Family Studies, The Pennsylvania State University, University Park, PA

Abstract

Factorial experimental designs have many potential advantages for behavioral scientists. For example, such designs may be useful in building more potent interventions, by helping investigators to screen several candidate intervention components simultaneously and decide which are likely to offer greater benefit before evaluating the intervention as a whole. However, sample size and power considerations may challenge investigators attempting to apply such designs, especially when the population of interest is multilevel (e.g., when students are nested within schools, or employees within organizations). In this article we examine the feasibility of factorial experimental designs with multiple factors in a multilevel, clustered setting (i.e., of multilevel multifactor experiments). We conduct Monte Carlo simulations to demonstrate how design elements such as the number of clusters, the number of lower-level units, and the intraclass correlation affect power. Our results suggest that multilevel, multifactor experiments are feasible for factor-screening purposes, because of the economical properties of complete and fractional factorial experimental designs. We also discuss resources for sample size planning and power estimation for multilevel factorial experiments. These results are discussed from a resource management perspective, in which the goal is to choose a design that maximizes the scientific benefit using the resources available for an investigation.

Keywords

multilevel modeling; cluster-randomized experiment; factorial experiment; fractional factorial experiment; intraclass correlation

[†]All three authors contributed equally.

Correspondence may be sent to Linda M. Collins, The Methodology Center, Penn State, 204 E. Calder Way, Suite 400, University Park, PA 16801; LMCollins@psu.edu.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/met

Introduction

Behavioral, educational, social, and health scientists have been increasingly interested in building and evaluating intervention programs (Midgley, 2006; Rychetnik, Frommer, Hawe, & Shiell, 2002). These interventions often include multiple components (i.e., multiple aspects of the intervention program itself as well as aspects of the program delivery or the implementation; see Collins et al., 2011), that may be interdependent in their functions and effects (Allore, Tinetti, Gill, & Peduzzi, 2005; Kwan, Hand, Dennis, & Sandercock, 2004; Welton, Caldwell, Adamopoulos, & Vedhara 2009). The traditional approach to intervention development involves constructing an intervention *a priori* and then evaluating it in a standard randomized controlled trial (RCT) with one treatment group and one control group. This approach provides vital information about the efficacy of the overall intervention as a package (Flay et al., 2005). However, it provides little information about which components of the intervention are contributing to the overall effect, or how the program might be improved. Questions about the differential effects of components might be important for answering theoretical questions, designing new treatment packages, or optimizing existing ones for effectiveness or cost-effectiveness (Ahn & Wampold, 2001; Allore et al., 2005; Chakraborty, Collins, Strecher, & Murphy, 2009; Collins, Murphy & Strecher 2007; Resick et al., 2008).

Recognizing both the strengths of the two-condition RCT for addressing questions about the overall effectiveness of an intervention as a package, and also its limitations for addressing questions about individual intervention components (West & Aiken, 1997; West, Aiken, & Todd, 1993), several authors (e.g., Chakraborty et al., 2009; Collins, Dziak & Li, 2009; Collins, Murphy, Nair, & Strecher 2005; Collins et al., 2011; Collins, et al., 2007) have suggested an approach that is used in many industrial and engineering settings. In such settings, in which there are often a large number of factors that might influence a dependent variable of interest, it is common to conduct *screening experiments* (Myers & Montgomery, 1995; Dean & Lewis, 2002; Wu & Hamada, 2000) as an important first step before further investigation. The purpose of a screening experiment is to identify which factors are *active* (have a substantial influence on the response variable) and merit further investigation in a subsequent experiment (Dean & Lewis, 2002; Wu & Hamada, 2000).

In behavioral interventions, the goal of a screening experiment is to determine which of several possible components of a proposed intervention (e.g., the presence or absence of different intervention components) have effects of practical significance. This information could be used, for example, in designing a follow-up experiment aimed at finding the optimum combination of levels of these components to comprise an efficacious and efficient intervention. It could also be used directly in proposing a cost-effective, streamlined intervention to be evaluated in a standard two-condition RCT. Screening experiments offer important new opportunities for behavioral scientists to (a) identify active components and assess their contribution to the potency of an intervention; (b) test out several new intervention components at once, and keep only those with the most potential in a later confirmatory intervention trial; (c) improve cost-effectiveness by studying which components might need to be removed, improved or replaced; and (d) better understand how an intervention operates (Collins et al., 2009; 2011). Screening experiments are based on a different mindset from traditional RCTs, in which the goal is to confirm the superiority of the new intervention over placebo or over existing practice, and so they might require different approaches to design and analysis.

In engineering applications, screening experiments are implemented using factorial designs, which allow for several independent variables and the interactions between them to be investigated efficiently and simultaneously (Myers & Montgomery 1995; Wu & Hamada

2000). Because investigators conducting screening experiments are mainly interested in identifying active factors, they are likely to be interested primarily in main effects (defined as the average effect of the factor across all combinations of levels of other factors), as well as in large or theoretically important second-order interactions. Third- and higher-order interactions are less likely to be helpful in choosing intervention components, because they might contain less practical and readily interpretable information.

One complication in implementing factorial experiments is that participants in the behavioral sciences are often nested or clustered within existing social or administrative units such as workplaces, schools, clinics or hospitals. Statistically, this structure involves non-independence among subjects (e.g., students) within upper-level units (e.g., schools). When there is no multilevel structure, statistical power is affected mainly by the effect size, the chosen Type I error rate, and the sample size. By contrast, in the case of multilevel data, power is also dependent on the number of clusters (J), the number of lower-level units within each cluster (n), and the intraclass correlation (ICC), which reflects the degree of dependence among observations within clusters. A large ICC can lead to considerably lower power than would be expected based on the total sample size, particularly if J is small or n is large. Because factorial experiments involve a relatively large number of experimental conditions and J might be only modest, investigators considering multilevel factorial experiments are likely to be quite concerned about whether they can feasibly maintain adequate statistical power with the resources available.

The purpose of the present article is to demonstrate that factorial experiments can be powerful and feasible even when the population of interest is clustered. We provide some ideas that may be helpful for future practical and methodological research in this area, including *ad hoc* power formulas and Monte Carlo simulation results that may be helpful in deciding whether a factorial design is feasible. We begin by briefly reviewing complete (also known as full) as well as fractional factorial designs. We then discuss two common approaches for multilevel experimentation: the between-clusters approach, in which clusters (e.g., schools or clinics) are assigned as whole units to experimental conditions; and the within-clusters approach, in which lower-level units (e.g., individual students or patients) are assigned individually to different experimental conditions. We show why power limitations become especially important for between-clusters experiments. Finally we bring these literatures together by discussing power-planning resources for multilevel factorial experiments, and use simulation studies to explore design elements that are likely to affect the power of multilevel factorial experiments.

Review of Complete and Fractional Factorial Experiments

Consider the following hypothetical example (modeled very loosely on Wachelka & Katz, 1999). Suppose an investigator wishes to develop an intervention program designed to reduce test anxiety among elementary school students, and there are three factors of theoretical interest to the investigator, each with two levels, which could be called *On* (experimental) and *Off* (control) for convenience. The factors are whether or not the student (1) is trained in a muscle relaxation technique (Relax), (2) is trained to recognize and dispute irrational beliefs (Cognitive), and (3) is trained in using guided imagery to improve self-confidence (Imagery).

In this example a complete factorial design would be a 2×2×2 (or 2^3 for short) factorial experiment, and would involve 8 experimental conditions. Table 1 shows the conditions of a complete factorial design along with the effect-coded design matrix of the resulting experiment (coding On=+1, Off=-1). For each factor, four of the eight conditions have the Off level and four have the On level. For each factor, there is a control group of four rows and an experimental group of four rows. This balance at the condition level makes the experiment more efficient (i.e., allows greater power for testing the main effects). We use the term *level* when referring to the value of one of the independent variables (e.g., Cognitive=On), and the term *condition* or *cell* when referring to a combination of levels (e.g., Relax=Off, Cognitive=On, Imagery=Off). We use the term *main effect* to mean the difference between factor levels, averaging across conditions (e.g., the average difference in response between the four conditions with Relax=On and the four with Relax=Off; see Myers & Well 2003, p. 302).

A complete factorial design with K dichotomous factors requires 2^K conditions, which is sometimes infeasible. Fractional factorial designs are an alternative that offers many of the advantages of a complete factorial design, while requiring considerably fewer experimental conditions (Kirk, 1995; Wu & Hamada, 2000). Fractional factorial designs are a variation upon factorial designs, involving the use of a carefully chosen subset of the experimental conditions of a complete factorial design. In other words, only certain conditions from the complete factorial are implemented. Obviously, the choice of which conditions to implement has important consequences for inference; it is done strategically, often using software, to allow estimation of the effects of primary interest. There may be many alternative fractional factorial designs to choose from, particularly when the number of factors is large.

Consider our test anxiety 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 1. This subset preserves the property that the main effects are orthogonal to each other and 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). For each factor, there is a control group of two rows and an experimental group of two rows. This balance makes 2, 3, 5, 8 more powerful, at least for estimating main effects, than if one were to use a subset like 1, 2, 3, 5, which examines each component separately versus a control condition. The 2, 3, 5, 8 study would be described as a 2^{3-1} fractional factorial, indicating that this particular fractional factorial design is 2^{-1} = ½ fraction of the complete 2^3 factorial; for this reason it is also called a half factorial. The main effects of interest can thus be tested without implementing all eight conditions. In general, when there are more than eight conditions, fractional factorial designs can involve implementing only, for example, half, one quarter or one eighth of the total number of conditions.

Although more economical in terms of experimental conditions, the fractional factorial does have some weaknesses compared to the complete factorial. When not all of the conditions of the complete factorial are implemented, not all possible effects can be estimated, because a design with C conditions cannot estimate more than C effects. Therefore, certain effects (whether main effects or interactions) are aliased (i.e., deliberately confounded) with one or more other effects that the investigator will assume to be negligible. For example, a given main effect might be confounded with a two, three, or four-way interaction, depending on the specific subset of conditions being implemented. In order to interpret the estimate of the main effect, one must assume that the interaction (or interactions) with which it is confounded is negligible. This may result in loss of scientific information, although that can be minimized by careful selection of the subset of conditions to implement, and hence of which effects will be confounded with which other effects. A thorough introduction to fractional factorial designs, how to determine which designs to use, and which effects are aliased with which other effects, is beyond the scope of this paper. Readers are referred to Kirk (1995) or Collins et al. (2009) in the behavioral sciences, and to Myers and Montgomery (1995) or Wu and Hamada (2000) in engineering. Software such as SAS PROC FACTEX (SAS Institute, 2004) can be used to find a subset of conditions that satisfies an investigator's specifications for number of conditions, number of factors, and amount of aliasing (e.g., one can specify that no main effect be confounded with any two- or three-way interaction, but only higher-order interactions). The software provides design matrices (interpretable as tables of conditions to be implemented) and listings of which interactions are aliased with each given effect of interest. These can be used whether or not the population has a multilevel structure.

Within-Clusters and Between-Clusters Multilevel Experiments

In the case of a multilevel population, a question arises concerning the level of assignment to conditions. Investigators may randomize individual subjects (within-clusters approach) or whole clusters (between-clusters approach) to conditions. In a *within-clusters* experiment, different members of the same cluster are independently assigned to different conditions, so cluster can be *crossed* with treatment. Such studies are sometimes called multisite trials (e.g., Chakravorti & Grizzle, 1975; Moerbeek & Teerenstra, 2011; Pituch & Miller, 1999; Raudenbush & Liu, 2000). For example, patients within the same clinic can still be individually assigned to different drugs or therapies (see Apfel et al., 2003 for a factorial example).

In the *between-clusters* approach, all members of a given cluster are assigned to the same condition, such that clusters are *nested* within conditions. Such designs are often called cluster-, group, community- or site-randomized trials. Baldwin, Murray, and Shadish (2005); Donner and Klar (2000); Moerbeek, van Breukelen, and Berger (2000); Moerbeek and Teerenstra (2011); Murray (1998); Raudenbush (1997); and Murray, Varnell and Blitstein (2004) review options for design and analysis. There may be both logical and practical reasons for a between-clusters approach. Sometimes a treatment can be delivered only to a cluster as a whole. For example, curriculum reforms or drug abuse education programs often have to be performed at the level of the classroom or even the whole school (Flay & Collins, 2005). A related reason for a between-clusters approach is that the intervention might operate by changing social behavior within clusters. In industrial and organizational psychology, interventions aiming to affect an individual through group or team processes are often administered to groups or teams as a whole (e.g., Gaudine & Saks, 2001). In medicine, some interventions are aimed at improving whole clinics or practices (Abernethy et al., 2006). Another reason for using a between-clusters approach is the danger of contamination that might be posed by a within-clusters approach, namely that effects of a treatment on one participant might diffuse to participants in other treatment groups because they are in the same cluster. If contamination occurs, then the estimates of the effects of interest can be biased and hypothesis tests can have poor power (Slymen & Hovell, 1997).

Power for the effects of interest depends on the level of assignment. Consider the following models for experiments having within- or between-clusters assignment. For simplicity, suppose there is only one dichotomous factor of interest. Our notation is adapted from that of Raudenbush and Liu (2001) and Spybrook, Raudenbush, Congdon, and Martinez (2009). Let Y_{ij} be the outcome of the i^{th} lower-level unit (e.g., student) within the j^{th} cluster (e.g., school), and X_{ij} be +1 if the subject receives the experimental condition and -1 for the control condition. For a within-clusters experiment one can model subject-level responses as

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

where the e_{ij}, which represent a combination of random student variability and measurement error, are N(0, σ^2) and are assumed independent for each individual. For simplicity we assume a random intercept at the cluster level but no variation in treatment effect between clusters. Thus

$$\begin{array}{ll}\text{Level 2:} & \beta_{0j}=\gamma_{00}+u_{0j}\\ & \beta_{1j}=\gamma_{10}\end{array}$$

where the random school effects u_{0j} are N(0, τ^2) and independent for each cluster. The independent additive error at the cluster level corresponds to an equicorrelated (exchangeable) correlation structure. γ_{00} represents an overall grand mean and γ_{10} is the regression coefficient for Y on X. Combining the above, the model is

$$Y_{ij}=\gamma_{00}+\gamma_{10}X_{ij}+u_{0j}+e_{ij}. \quad (1)$$

The between-clusters approach leads to a similar model. One can assume

$$\begin{array}{ll}\text{Level 1:} & Y_{ij}=\beta_{0j}+e_{ij}\\ \text{Level 2:} & \beta_{0j}=\gamma_{00}+\gamma_{01}X_j+u_{0j}\end{array}$$

so that

$$Y_{ij}=\gamma_{00}+\gamma_{01}X_j+u_{0j}+e_{ij}. \quad (2)$$

where, as before, the e_{ij} are assumed independent N(0, σ^2) and the u_{0j} are assumed independent N(0, τ^2). Models (1) and (2) are quite different in terms of their interpretations as multilevel models. For within-clusters assignment, the regression coefficient for X_{ij} is called γ_{10} and interpreted as a subject-level parameter; but for between-clusters assignment, the regression coefficient for X_j is called γ_{01} and interpreted as a cluster-level parameter. However, algebraically the regression equation (2) is only a special case of the regression equation (1) in which all X_{ij} within a cluster equal the same value X_j. The expected value of Y_{ij} is γ_{00} + $\gamma_{10}X_{ij}$ in (1) or γ_{00} + $\gamma_{01}X_j$ in (2). The parameter γ_{10} or γ_{01} expresses the expected increase in a subject's Y value when that subject's X value increases by one unit. Equivalently, γ_{10} or γ_{01} expresses half the expected change in Y when X is taken from low (-1) to high (+1), because +1−(−1) = 2. Thus γ_{10} or γ_{01} expresses ½ times the effect of the treatment; the multiplier ½ does not change the power of the significance test because the standard error is also rescaled.

Power Implications of Between- and Within-Clusters Assignment

To demonstrate how between- and within-clusters assignment differ in power, consider the design effect, D. This is the ratio of the variance of the treatment effect estimate in the clustered population, to the variance this quantity would have had if the subjects were independent (Kish, 1965; Murray, 1998; Murray & Blitstein, 2003; Snijders & Bosker, 1999). Thus, if the test of the treatment effect would have required a sample size of N subjects for a given level of power under classic conditions, the fact that subjects are clustered increases the needed sample size to approximately DN. Thus, the larger D is, the larger the sample size needed for adequate power.

The design effect for a significance test for the treatment effect γ_{01} in Model (1) or γ_{10} in Model (2) can be roughly approximated by

$$D=1+(n-1)\rho_X\rho_{Y|X} \quad (3)$$

where ρ_X and $\rho_{Y|X}$ are the ICCs for X and for Y conditional on X (Scott & Holt, 1982; Neuhaus & Segal, 1993). Regardless of the level of treatment assignment, $\rho_{Y|X}$ = τ^2 / (σ^2 +

τ^2) (Raudenbush, 1997; Murray, 1998; Spybrook et al., 2009). However, ρ_X depends strongly on the level of assignment. In a within-clusters experiment, in which either subjects are assigned to experimental conditions independently of cluster membership or conditions are stratified within clusters, $\rho_X = 0$. Therefore, $D = 1$. In a between-clusters experiment, $\rho_X = 1$ because the independent variable is the same for every member of the cluster, leading to the formula $D = 1 + (n-1)\ \rho_{Y|X}$ (Kish, 1965; Murray, 1998; Murray & Blitstein, 2003). For example, suppose $n = 100$ and $\rho_{Y|X} = .1$. Then the design effect is near 1 for a within-clusters approach but 10.9 for a between-clusters approach. Thus the between-clusters approach requires a total sample size ($N = Jn$) which is over ten times larger here than the within-clusters approach would require, to achieve a given level of power.

Accomodating More Factors in the Model

Models (1) and (2) extend naturally to complete or fractional factorial experiments with several factors. Model (1) for a within-clusters factorial experiment generalizes to

$$Y_{ij}=\gamma_{00}+\sum_{k=1}^{K}\gamma_{k0}X_{kij}+u_{0j}+e_{ij}, \tag{4}$$

and Model (2) for a between-clusters factorial experiment generalizes to

$$Y_{ij}=\gamma_{00}+\sum_{k=1}^{K}\gamma_{0k}X_{kj}+u_{0j}+e_{ij}. \tag{5}$$

where $\gamma_{01},\dots,\gamma_{0K}$ or $\gamma_{10},\dots,\gamma_{K0}$ are the regression coefficients for the K effects of interest (main effects and interactions). As before, the e_{ij} are assumed independent N(0, σ^2) and the u_{0j} are assumed independent N(0, σ^2). Here, each factor is assumed to be dichotomous and represented by an effect-coded independent variable. Interactions in these models can be represented by including products of effect-coded predictors (Myers & Well, 2003), and likewise factors with more than two levels can be represented with additional codes.

Accomodating a Pretest in the Model

The power and precision of the tests of interest in Models (4) or (5) can potentially be improved by adding other covariates besides the treatment variables, particularly a pretest (Murray & Blitstein, 2003; Raudenbush, Martinez, & Spybrook 2007). A pretest might be at the subject level or the cluster level. Intuitively, if subject-level outcomes are of interest, then the subject-level pretest should be more informative than a cluster-level aggregate. Furthermore, cluster-level covariates use up cluster-level degrees of freedom (*df*), which are scarce, while subject-level covariates use up only subject-level *df*, which are relatively plentiful (Murray, 1998). However, cluster-level pretests are sometimes easier to obtain, and Bloom, Richburg-Hayes, and Black (2007) argued that in educational studies with school-level interventions, a cluster-level pretest used as a covariate was sometimes as effective as a subject-level pretest. In this article we focus on a subject-level pretest (e.g., an assessment given to each student at the beginning of the intervention period, rather than an aggregate measure of past performance for each school). One could incorporate a subject-level pretest either as a covariate, or as a repeated measure in a three-level model.

Two-Level, Covariate-Adjusted Approach—To incorporate pretest as a covariate as in classic ANCOVA for within-clusters assignment, suppose

$$\text{Level 1:}\quad Y_{ij}=\beta_{0j}+\xi_j P_{ij}+\sum_{k=1}^{K}\beta_{kj}X_{kij}+e_{ij}$$

$$\begin{aligned}\text{Level 2:}\quad \beta_{0j}&=\gamma_{00}+u_{0j}\\ \xi_j&=\gamma_{\xi 0}\\ \beta_{kj}&=\gamma_{k0}\qquad k=1,\ldots,K\end{aligned}$$

where P_{ij} is the pretest. Then, analogously to Model (4), the combined model becomes

$$Y_{ij}=\gamma_{00}+\gamma_{\xi 0}P_{ij}+\sum_{k=1}^{K}\gamma_{k0}X_{kij}+u_{0j}+e_{ij}. \tag{6}$$

Model (5) can be augmented with a covariate in the same way to obtain

$$Y_{ij}=\gamma_{00}+\gamma_{\xi 0}P_{ij}+\sum_{k=1}^{K}\gamma_{0k}X_{kj}+u_{0j}+e_{ij}. \tag{7}$$

As before, the u_j and e_{ij} represent random additive cluster and subject effects assumed to be independent N(0, τ^2) and N(0, σ^2), although they are now random effects on the covariate-adjusted Y rather than the raw Y.

Three-Level Approach—Rather than considering the pretest as a covariate, an alternative would be to consider both the pretest and posttest as observations nested within the subject. In a three-level hierarchical linear model approach (Bryk & Raudenbush 1988, Raudenbush & Liu 2001), observations within the subject over time can be modeled as a linear growth curve

$$\text{Level 1:}\quad Y_{tij}=\pi_{0ij}+\pi_{1ij}T_{tij}+e_{tij}$$

where T represents time and the e_{tij} are mutually independent N(0, σ^2). Because we assume only a total of two observations per subject, the pretest and posttest, it would not be reasonable to try to model quadratic or other nonlinear effects here.

First consider the case of *within-clusters* assignment. Here the treatment effect is represented at Level 2 (i.e., the level of the individual subject).

$$\begin{aligned}\text{Level 2:}\quad \pi_{0ij}&=\beta_{00j}+\sum_{k=1}^{K}\beta_{0kj}X_{kij}+r_{0ij}\quad r_{0ij}\sim\mathrm{N}(0,\tau_{\pi 0}^{2})\\ \pi_{1ij}&=\beta_{10j}+\sum_{k=1}^{K}\beta_{1kj}X_{kij}+r_{1ij}\quad r_{1ij}\sim\mathrm{N}(0,\tau_{\pi 1}^{2})\end{aligned}$$

$$\begin{array}{lll} \text{Level 3:} & \beta_{00j}=\gamma_{000}+u_{00j} & u_{00j}\sim \mathrm{N}(0,\tau^2_{\beta 0}) \\ & \beta_{10j}=\gamma_{100}+u_{10j} & u_{10j}\sim \mathrm{N}(0,\tau^2_{\beta 1}) \\ & \beta_{0kj}=\gamma_{0k0},\beta_{1kj}=\gamma_{1k0} & k=1,\ldots,K \end{array}$$

The r_{1ij} effects can be interpreted as a subject-level random slope for time, or as a time-by-subject interaction. However, with only a pretest and posttest it is impossible to empirically distinguish the r_{1ij} from the measurement errors e_{tij}, so for simplicity of the model we absorb the former into the latter, effectively assuming $\tau^2_{\pi 1}=0$. The combined model is then

$$Y_{tij}=\left(\gamma_{000}+u_{00j}+r_{0ij}+\sum_{k=1}^{K}\gamma_{0k0}X_{kij}\right)+\left(\gamma_{100}+u_{10j}+\sum_{k=1}^{K}\gamma_{1k0}X_{kij}\right)T_{tij}+e_{tij}. \tag{8}$$

Here γ_{000} and γ_{100} represent an average intercept and slope for time, r_{0ij} is a random effect for subject, u_{00j} is an independent random effect for cluster, and u_{10j} is a independent random cluster-by-time interaction. γ_{1k0} is half of the population mean difference in the pretest-to-posttest gain score, between subjects receiving the high and low levels of X_{kij}, averaging over levels of the other factors.

For between-clusters factorial experiments, the treatment variables are at Level 3.

$$\begin{array}{lll} \text{Level 2:} & \pi_{0ij}=\beta_{00j}+r_{0ij}, & r_{0ij}\sim \mathrm{N}(0,\tau^2_{\pi 0}) \\ & \pi_{1ij}=\beta_{10j}+r_{1ij}, & r_{1ij}\sim \mathrm{N}(0,\tau^2_{\pi 1}) \end{array}$$

$$\begin{array}{lll} \text{Level 3:} & \beta_{00j}=\gamma_{000}+\sum_{k=1}^{K}\gamma_{00k}X_{kj}+u_{00j}, & u_{00j}\sim \mathrm{N}(0,\tau^2_{\beta 0}) \\ & \beta_{10j}=\gamma_{100}+\sum_{k=1}^{K}\gamma_{10k}X_{kj}+u_{10j}, & u_{10j}\sim \mathrm{N}(0,\tau^2_{\beta 1}) \end{array}$$

Combining the levels and once again setting $\tau^2_{\pi 1}=0$, we obtain

$$Y_{tij}=\left(\gamma_{000}+u_{00j}+r_{0ij}+\sum_{k=1}^{K}\gamma_{00k}X_{kj}\right)+\left(\gamma_{100}+u_{10j}+\sum_{k=1}^{K}\gamma_{10k}X_{kj}\right)T_{tij}+e_{tij}. \tag{9}$$

As before, Model (9) is closely analogous to Model (8), with γ_{00k} taking the role of γ_{0k0} and γ_{10k} taking the role of γ_{1k0}. The time variable T_{tij} in (8) or (9) distinguishes between pretest and posttest; it is treated here as categorical. In this paper we follow Raudenbush & Liu (2001) and Spybrook et al. (2009) in centering time, so that the pretest is indicated as T_{tij} = -½ and the posttest is indicated as T_{tij} = + ½; see Appendix A for details.

The parameters of interest for testing the efficacy of the treatment components are likely to be $\gamma_{110},\ldots\gamma_{1K0}$ or $\gamma_{101},\ldots\gamma_{10K}$. For example, in the within-clusters case, γ_{1k0} expresses the effect of X_{kij} on the average change in the response from pretest to posttest, so a significant positive value means that setting X_{kij} to +1 increases the expected pretest-to-posttest gain (or reduces pretest-to-posttest loss) relative to setting X_{kij} to -1. The analogous statement holds for γ_{10k} and X_{kj} in the between-clusters case.

The other set of regression parameters, $\gamma_{010},\ldots\gamma_{0K0}$ or $\gamma_{001},\ldots\gamma_{00K}$, are a required part of the model, but they are not readily interpretable because of the centering of time.

Specifically, when time is centered, then γ_{0k0} or γ_{00k} expresses the effect of X_{kij} or X_{kj} on the *average* of the pretest and the posttest, which is not generally of interest in a randomized experiment. We chose to center time despite this disadvantage, because the expression $Y_{tij} = \pi_{0ij} + \pi_{1ij}T_{tij} + e_{tij}$ with random effects in π_{1ij} forces the cluster-level error variance to depend on the magnitude of T, so that (without further *ad hoc* changes to the parameterization) a 0/1 coding would have required making the strong assumption that cluster-level error variance is considerably higher at posttest than at pretest, and this limitation would have made it more difficult to provide useful power formulas.

In summary, the pretest can be incorporated using covariate-adjusted Models (6) or (7), or three-level Models (8) or (9). Janega et al. (2004), Murray (1998), and Murray and Blitstein (2003) further discuss the use and relative merits of these models in a cluster-randomized experiment.

Questions of Interest Concerning Sample Size and Power

To discuss the feasibility of multilevel factorial screening designs, it is important to determine whether a multilevel, multifactor design can offer acceptable power, given realistic values for J and n. This might depend on the nature of the factorial design (complete or fractional) and the type of assignment (between-clusters or within-clusters). A special question about power arises when there are not enough clusters to conduct a between-clusters complete factorial experiment. Specifically, it is of interest whether fractional factorial designs can offer acceptable power here despite the infeasibility of complete factorial designs. For example, suppose an investigator wishes to study five factors but has access to only 25 clusters. A 5-factor complete factorial design involves 2×2×2×2×2 = 32 cells, so it cannot be implemented. A 2^{5-1} fractional factorial design would require only 16 conditions, so 25 clusters are enough to assign one or two clusters to each condition. Thus, a fractional factorial is at least possible (although perhaps not with high power), even though the complete factorial would be impossible.

Another unresolved issue is how well power for multilevel factorial designs can be predicted. Power formulas are known in the literature for clustered RCTs having one dichotomous factor (e.g., Murray, 1998; Raudenbush, 1997; Raudenbush & Liu 2000). Can the existing power formulas for one dichotomous factor generalize to designs with K dichotomous factors, and does their ability to predict power depend on the nature of the factorial design (complete or factorial)? In the following sections we investigate power in multilevel complete and fractional factorial designs, using first heuristic arguments and then Monte Carlo simulations.

Sample Size and Power Planning for Multilevel Factorial Experiments

Few power planning resources for within- and between- clusters factorial designs are currently available. However, formulas exist for simpler designs, such as unclustered factorials or single-factor clustered designs. In this section we show how these formulas can be adapted to predict power for within- and between- clusters factorial designs. For simplicity we focus only on dichotomous effect-coded factors, but the ideas here could be extended to other cases. Also for simplicity, we assume that for each factor, assignment probabilities to each level are equal (.50 and .50), leading to approximately balanced sample sizes in each condition. For within-clusters designs, we assume assignment probabilities for conditions do not differ among clusters.

Power for Within-Clusters Effects

In considering how to predict power for within-clusters factorial designs, it is useful to consider the orthogonality or recycling property (see Collins et al., 2009). This property suggests that for the purposes of detecting a single main effect of a given size for a given factor, the fact that other factors are also being investigated in the experiment does not have a large deleterious effect. Thus, existing formulas for a two-condition RCT (i.e., a one-factor factorial) can be used in an approximate way for predicting power for the main effects of dichotomous factors when there are more than one factor. Because both complete and fractional factorial designs have this balance property, one might conjecture that they offer similar power for a given total sample size (in the absence of problems caused by inappropriate aliasing; see Wu & Hamada, 2000). From the existing literature (e.g., Spybrook et al., 2009), the power for the main effect of a single dichotomous factor is approximately the probability that a noncentral F-distribution with noncentrality parameter λ and with 1 and ν degrees of freedom exceeds the critical value for the F-test (e.g., about 3.94 if α = .05 and ν = 100). For within-clusters assignment, ν can be counted on the subject rather than cluster level, so that it equals the total sample size $N = J\bar{n}$ minus the number of coefficients in the regression model. The noncentrality parameter λ equals $\frac{\gamma^2}{\mathrm{Var}(\widehat{\gamma})}$ where γ represents the parameter of interest and Var($\hat{\gamma}$) is its sampling variance (i.e., the square of its standard error).

Following the earlier discussion of the design effect, one can conjecture that clustering should not have a very large effect on power in the within-clusters case, assuming that assignment probabilities are the same regardless of cluster. Expression (3) suggests that $\rho_{Y|X}$ matters little, because each ρ_X is 0. Then for approximate power planning purposes, one could assume that the conditional ICC of Y is 0; that is, the design effect is 1. If this approximation is reasonable, it might be easy to find Var($\hat{\gamma}$. For example, in Model (4), where the parameter of interest is γ_{k0}, the sampling variance would be approximately $\sigma^2_{\mathrm{tot}}/N$ where $\sigma^2_{\mathrm{tot}}=\tau^2+\sigma^2$. This formula expresses the variance of half the difference between the means of two equally sized samples being compared in an independent samples t-test or ANOVA. It is also the variance of a regression coefficient of an effect-coded predictor assuming balanced assignment and orthogonality with other predictors (see, e.g., Neter, Kutner, Nachtsheim, & Wasserman, 1996). In other words, because clustering is expected to have only a minor effect on power for within-clusters factors, one could treat the two-level model roughly as if it had only one level.

For Model (8), the parameter of interest is γ_{1k0} and it can be estimated by $\widehat{\gamma}_{1k0}=\frac{1}{2}\left(\left(\overline{Y}_{+}^{+}-\overline{Y}_{-}^{+}\right)-\left(\overline{Y}_{+}^{-}-\overline{Y}_{-}^{-}\right)\right)$, where $\overline{Y}_{+}^{+}$ represents the average posttest for the X_k = +1 level, $\overline{Y}_{-}^{+}$ represents the average pretest for the for the X_k = +1 level, $\overline{Y}_{+}^{-}$ represents the average posttest for the X_k = −1 level, and $\overline{Y}_{-}^{-}$ represents the average pretest for the X_k = −1 level. Assuming that both the cluster effects and the levels of the other factors approximately balance between treatment levels, and that N/ 2 subjects are in each of the 2 levels of X_k, then the variance of $\overline{Y}_{+}^{+}-\overline{Y}_{-}^{+}$ or of $\overline{Y}_{+}^{-}-\overline{Y}_{-}^{-}$ can be approximated by $2\sigma^2$ / (N/ 2) = $4\sigma^2$ / N. Thus the variance of γ_{1k0} is approximately $2\sigma^2$ / N. This is because, in Model (8), the u_{00j} and r_{0ij} cancel out between pretest and posttest, and we assume the u_{10j} cancel out approximately between experimental and control members of each cluster, so the e_{tij} are the only variance component left. A worked example using this formula is shown in the online Appendix D.

Power for Between-Clusters Effects

For *between-clusters* factorial experiments under Models (5), (7), or (9), one cannot just ignore the ICC without severely overestimating power for a given total sample size. However, properly adjusted power formulas (see Donner & Klar, 2000; Murray, 1998;

Raudenbush 1997; Raudenbush & Liu, 2001) and software (Liu, Spybrook, Congdon, Martinez, & Raudenbush 2009) are available in cluster-randomized RCTs that compare a single experimental condition to a single control condition. These two-condition RCTs can be viewed as factorial designs with $K = 1$. The idea of the orthogonality of factors in a balanced factorial design suggests that the same formulas be used to calculate approximate power for a given factor when $K > 1$ as well, simply acting as if K were still 1, except for adjusting the error degrees of freedom to reflect the larger model.

We can still use the noncentral F distribution with degrees of freedom 1 and ν and noncentrality parameter $\lambda = \gamma^2$ / Var($\hat{\gamma}$). However, now ν would be the total number of *clusters* minus the number of cluster-level regression coefficients. This means that even assuming perfect orthogonality, adding more factors to an experimental design could still have at least a slight negative effect on power, because between-clusters effects use up the cluster-level *df*, already a limited resource.

Drawing on existing literature, the noncentrality parameter λ is found as follows. For the two-level Model (5), testing the significance of $\hat{\gamma}_{0k}$, assuming balanced numbers and sizes of clusters in each level of the factor, $\mathrm{Var}(\widehat{\gamma_{0k}})=\frac{\tau^2}{J}+\frac{\sigma^2}{Jn}$ so $\lambda=\frac{J\gamma_{0k}^2}{\tau^2+\sigma^2/n}$ (e.g., Raudenbush 1997, Spybrook et al., 2009, pp. 54-55). For the three-level Model (9), again assuming balance, $\mathrm{Var}\,(\widehat{\gamma_{10k}})=\frac{\tau_{\beta 1}^2}{J}+\frac{2\sigma^2}{Jn}$ so $\lambda=\frac{J\gamma_{10k}^2}{\tau_{\beta 1}^2+2\sigma^2/n}$ (see Appendix A; Murray & Blitstein, 2003; Raudenbush & Liu, 2000; Spybrook et al., 2009, pp. 120-121). A worked example using this formula is shown in the online Appendix D.

Unfortunately it is difficult to directly calculate power for the covariate-adjusted Model (7) with formulas like these. This is because including the pretest variable effectively changes both Level 1 and Level 2 variance parameters. Without prior empirical knowledge or restrictive assumptions about the distribution of the pretest and its relationship with the posttest, one cannot predict the degree to which adjusting for the pretest will reduce, or perhaps increase, the variance and conditional ICC of Y (Murray, 1998). The more elaborate three-level Model (9) avoids this problem by specifying the joint distribution of pretest and posttest. However, Model (7) by itself is not rich enough to allow for a power formula. Investigators planning to use Model (7) might consider using the power formula from Model (9) as an approximation (even though the models involve different, and not fully compatible, parameterizations and assumptions; Allison, 1990).

Another issue is that in practice the clusters do not have equal size, and this has especially important consequences in the between-clusters case. One could use the (arithmetic) mean cluster size $\bar{n}$ in place of n in the formulas. If so, then following Eldridge, Ashby, and Kerry (2006), we recommend also adjusting the cluster-level variances by $\left((CV_n)^2+1\right)$, where CV_n is some reasonable guess at the value of the coefficient of variation $SD(n)/\bar{n}$ of the cluster sizes to be observed. This helps to take into account the deleterious effect on power of having unbalanced cluster sizes. In particular, the sampling variance for the treatment effect becomes $\frac{\left(CV_n^2+1\right)\tau^2}{J}+\frac{\sigma^2}{Jn}$ in (5) and becomes $\frac{\left(CV_n^2+1\right)\tau_{\beta 1}^2}{J}+\frac{2\sigma^2}{Jn}$ for (9).

The above pertains to power for a main effect. To obtain power for a two-way interaction, one could use the same formulas but with an adjusted effect size. Specifically, the power to detect an interaction of size d is approximately equal to the power to detect a main effect of size $d / 2$ (Murray, 1998; Montgomery, Peters, & Little, 2003; Myers & Well 2003, p. 305); see Appendix A.

The power formulas proposed in this section are summarized in Table 2. In the table as in the text, we continue to assume dichotomous factors. This table shows two equivalent forms for the noncentrality parameters. One expresses them in terms of the regression parameter γ for the effect of interest for factor *k*. The second form expresses them in terms of a standardized difference analogous to d in Cohen (1988). This is calculated by setting $d = 2\gamma / \sigma_{\text{tot}}$, which gives the expected mean difference in Y (adjusted for the pretest if appropriate) between two levels of factor *k*, as a ratio to the overall error variance σ_{tot} (see Appendix B).

Monte Carlo Simulation Study: Methods

We use Monte Carlo simulations to explore (a) whether it is feasible to obtain acceptable statistical power for detecting main effects and two-way interactions in a screening experiment, using multilevel factorial designs with sample sizes commonly available in practice; and (b) how adequately the power formulas proposed here for within- and between-clusters factorial designs, can predict power. These questions must be addressed for each of four experimental approaches: (1) complete factorial design, within-clusters assignment; (2) complete factorial design, between-clusters assignment; (3) fractional factorial design, within-clusters assignment; and (4) fractional factorial design, between-clusters assignment. Within each of these four design categories, we simulated 5000 datasets for each of several scenarios, representing various values of J, $\bar{n}$, and ICC. The total number of scenarios considered was 60; that is, 2 designs × 2 values of J × 2 values of $\bar{n}$ × 3 values of ICC=24 within-clusters scenarios, plus 2 values of J × 2 values of $\bar{n}$ × 3 values of ICC =12 complete factorial between-clusters scenarios, plus 4 values of J × 2 values of $\bar{n}$ × 3 values of ICC =24 fractional factorial between-clusters scenarios. These scenarios are described below and summarized in Table 3.

Cluster Sizes and Counts

We set J for the within-clusters assignment scenario to equal 5 or 10. These values seem realistic given multisite within-clusters trials such as IMPACT (Fielding, Mason, Knight, Klesges, & Pelletier, 1995; $J = 5$), TORDIA (Brent et al., 2008; $J = 6$), or the 2 × 2 factorial CREATE (Frasure-Smith et al., 2006; Lespérance et al., 2007; $J = 9$). We set J for the between-clusters assignment to equal 25, 30, 40, and 50. The low end of this range seems realistic given between-clusters RCTs such as REACT (Murray, Feldman, & McGovern, 2000), STARS (Williams et al., 2007), or the Sydney Ambulance Service Trial (Gomel, Oldenburg, Simpson, & Owen, 1993; Murray, 1998). The high end of this range would be more comparable to a very large cluster-randomized trial such as the Tri-Ministry Study (Hundert et al., 1999; $J = 60$) or the three-condition Child and Adolescent Trial for Cardiovascular Health (Luepker et al., 1996; Murray, 1998; $J = 96$). For the between-clusters scenarios, we set $\bar{n}$ to equal 20 (e.g., a classroom or small clinic) or 100 (e.g., a grade in a school, a department in a hospital or a sizable worksite). For the within-clusters scenarios, because J was smaller, we let the lower value of $\bar{n}$ be 50 instead of 20 to avoid scenarios with impractically small total sample sizes.

To improve realism, the cluster sizes were not assumed constant. However, it was difficult to determine in general what a realistic shape would be for the distribution of cluster sizes. When the clusters were classrooms, Dee and West (2011) observed a coefficient of variation of .24. If cluster sizes are normally distributed, then the coefficient of variation will not be above approximately 0.5 (van Breukelen, Candel, & Berger, 2006). However, when the clusters were medical practices and clinics, Eldridge et al. (2006) reported coefficients of variation that were often as large as .65, and very strong right skew. In contrast, some previous work on power has seemingly ignored variability in cluster size, as if to assume the coefficient of variation was zero. It is not feasible to simulate every possible case, so we decided to generate the size of each cluster randomly from a discrete uniform distribution from $\bar{n} / 2$ to $3\bar{n} / 2$. Thus, if $\bar{n}$ for the scenario is set to 20, then values from 10 to 30 are possible and equally likely, but values less than 10 or greater than 30 are impossible. This range seemed plausible although somewhat arbitrary. It implies a CV_n of approximately .29 due to the properties of the uniform distribution (Casella & Berger 1990).

Simulated Experimental Design and Data-Generating Model

For each simulated dataset, we generate five factors (independent variables). These factors might correspond to potential intervention components, as in the test anxiety prevention example. The factors were assumed to be tested using a complete or fractional factorial experimental design, as shown in Table 4, with a total pool of J clusters. The response Y of each subject (perhaps on a measure of test anxiety or some other problematic behavior or symptom, in which case we assume that the coding is reversed to be able to call a helpful effect positive) on a pretest and a posttest was simulated using Model (8) for within-clusters scenarios or Model (9) for between-clusters scenarios. We chose Models (8) and (9) because they are very conducive to a Monte Carlo simulation; they provide an explicit joint distribution of the pretest and posttest given the parameters. Models (4) or (5) would lead to unrealistically poor performance because they do not allow the use of a pretest. Models (6) and (7) by themselves do not provide enough information for simulations because they do not specify the distribution of the pretest. Thus, Models (8) and (9) were used. The parameter values chosen for implementing Models (8) and (9) are described below.

Intraclass Correlations

Under Models (8) or (9), the ICC of Y at posttest, conditioning on the X variables and not adjusting for the pretest, is equal to the ICC of Y at pretest. This conditional ICC is related to power, with higher ICCs expected to be associated with lower statistical power, at least for between-clusters factors. This is because the more the observations within a cluster are strongly related, the less independent information is available for testing hypotheses. Because ICC varies depending on the nature of the variable being investigated, we implemented three different ICC scenarios, each of which might be realistic in different cases.

- **Low** ($\rho_{Y|\mathbf{X}}$ = .05). ICC values between .01 and .10 are often found in health-related school intervention settings (Ma & Klinger, 2000; Murray & Blitstein, 2003; Siddiqui, Hedeker, Flay, & Hu, 1996; Slymen & Hovell 1997). An ICC between .02 and .10 for medical outcome measures within clinics or practices is common (Campbell, Fayers, & Grimshaw, 2005). Murray et al. (2006) used ICC values of .001 to .05 in simulations.
- **Medium** ($\rho_{Y|\mathbf{X}}$ = .15). Schochet (2008) considered .15 to be a reasonable ICC value for power computations with school achievement variables. ICC values of .15 to .25 within schools have been reported in educational achievement tests (Spybrook et al., 2009). Values around .15 were also typical for school achievement in rural schools in Hedges and Hedburg (2007a). In an example in Singer (1998), the within-school ICC was .18, although it was reduced to .06 after adjusting for cluster-level socioeconomic status.
- **High** ($\rho_{Y|\mathbf{X}}$ = 30). Most reported ICCs within schools, workplaces or hospitals seem to be below this value. Hedges and Hedburg (2007b) reported unadjusted

ICCs around .30 in their review of educational achievement tests in mostly urban schools.

Appropriate values for the variance parameters to fulfill each of these three scenarios were calculated as described in Appendix B. The total variance at pretest or posttest was standardized at 1.

Fixed-Effects Parameters

Under Model (8) or (9), the intercept γ_{000} and the average change over time γ_{100} do not involve the effect of interest and do not affect the power of its test. This is because they apply to all participants regardless of conditions, so they cancel out when comparing condition means. Because their values did not matter to the questions of interest, they were both arbitrarily set to zero, which roughly imitates a situation in which both pretest and posttest are centered around their grand means.

The regression parameters for the X variables were chosen to simulate the following situation: at pretest, the effects of all the factors and interactions were 0. At posttest, X_1, X_3, and X_5 each had effect sizes of $d = +0.40$ relative to the measurement-level error standard deviation σ, represented by effect-coded regression coefficients of $.2\sigma$. This corresponded to about d=.2 relative to the overall standard deviation σ_{tot}. For example, in the between-clusters cases, $\gamma_{101} = 0.20\sigma$ with +1/-1 coding; see Appendix A. Depending on the scenario, σ was about .5 or .6 times σ_{tot} because σ^2 was about .2 to .3 times σ_{tot}; see Appendix B. Thus, $\gamma_{101} = 0.20\sigma \approx 0.10\sigma_{\text{tot}}$, and so the main effect as a ratio to σ_{tot} was $2\gamma_{101}$. The main effect in raw units is also $2\gamma_{101} = 0.40\sigma \approx 0.20\sigma_{\text{tot}}$, because σ_{tot} was fixed at 1.

Specifically, 0.40σ was $.2306\sigma_{\text{tot}}$, $.2182\sigma_{\text{tot}}$, or $.1980\sigma_{\text{tot}}$ in the low, medium or high ICC scenarios, respectively (see Table 3 and Appendix B). Thus, in terms of the Cohen (1988) benchmarks, the effect sizes could be called medium (around d=.4) or low (around d=.2) in relation to the individual error standard deviation or the total standard deviation, respectively.

We set the remaining γ coefficients as follows. X_2 and X_4 had coefficients of zero. $X_1{\times}X_2$ and $X_1{\times}X_3$ had coefficients of $+0.1\sigma$ and -0.1σ respectively, $X_1{\times}X_3{\times}X_5$ had a coefficient of $+0.05\sigma$, and $X_1{\times}X_2{\times}X_3{\times}X_5$ had a coefficient of $+0.025\sigma$. All other interactions had zero effect.

Having some effect sizes be nonzero and others be zero allows us to assess both Type I and Type II error rates. We decided to set coefficients for complex interactions to be progressively smaller than for simpler ones by factors of 2 each time; this was somewhat arbitrary but consistent with the sparsity principle, which is a heuristic or practical working assumption proposing that most of the interesting variability in a system can be attributed to a few main effects and a few two-way interactions, and that most other effects are likely to be relatively small (see Collins et al., 2009, Wu & Hamada, 2000, for discussion).

Complete or Fractional Factorial Design

To generate each dataset, we first constructed the design matrix $\mathbf{X}$ as appropriate for the scenario, then generated the random effects, and then computed the resulting Ys. For the simulated complete factorial experiments, each dataset included 32 conditions corresponding to the cells of a 2×2×2×2×2 factorial. For the simulated fractional factorial experiments, 16 conditions were selected using SAS PROC FACTEX (SAS Institute, 2004), as noted in bold in Table 4. In this particular design, each main effect was aliased (i.e., confounded by design) with the four-way interaction of the other four factors, and each two-way interaction was aliased with the interaction of the other three factors (see Wu &

Hamada, 2000). Specifically, the inactive Factor 4 was therefore aliased with the active interaction between Factors 1, 2, 3 and 5, and the interaction between Factors 2 and 4 was aliased with the active interaction between Factors 1, 3, and 5. For example, the products of the columns in the design matrix for Factors 1, 2, 3, and 5 are equal (in all 16 rows of the fractional factorial design) to the column for Factor 4, so it is impossible for that interaction and the main effect of that factor both to be in the model. In fact, to estimate or test the main effect of Factor 4, it is necessary to assume that the 1×2×3×5 interaction is zero. Because this interaction is not, in fact, zero in the true data-generating model of the simulation, it is expected that the estimate and test for Factor 4 will be biased, possibly leading to higher Type I error rate or poorer power. However, because the coefficient for the interaction is very small, it is also expected that this bias should be slight. Thus, for the fractional factorial conditions, we are simulating a situation in which the assumptions motivating the investigator's choice of design are almost, but not quite, satisfied.

Assignment to Conditions (Between- or Within- Clusters)

For the within-clusters designs, each individual in each cluster was independently randomly assigned to one of the conditions in the design. For the between-clusters designs, clusters were assigned to conditions in a restricted way that assured that no two conditions differed in size by more than one cluster. For example, in the $J = 25$, fractional factorial condition, there are 25 clusters and 16 conditions, so at least one cluster must be in each condition, and the remaining 9 are randomized to any condition, subject to no condition receiving more than 1 of them, that is, more than 2 clusters total. The sets of conditions used to test each main effect, implied by Table 4, are shown in Figure 1.

Analysis of Simulated Data

We performed the simulations using SAS and R, implementing them in both to allow cross-checking of results. Analyses of simulated datasets were done using PROC MIXED in SAS (see Singer, 1998) and the *nlme* package in R (R Development Core Team 2010, Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2010). We found during earlier simulations that for between-clusters designs with 100 members and more than 30 clusters, it was computationally difficult to fit Model (9) in either SAS or R for the thousands of simulated datasets. That is, it was sometimes computationally difficult to fit (9) and test whether $\gamma_{10k} = 0$, but straightforward to fit Model (7) and test whether $\gamma_{0k} = 0$; the meanings of the two tests are similar in this context although not identical. Therefore, even though the data were generated under the three-level Models (8) or (9), we analyzed them under the adjusted two-level covariate-adjusted Models (6) and (7). Also, in keeping with the interpretation of a screening study, we analyzed the data using a model that only included main effects and two-way interactions, even though (as described above) the data did in fact include a few small higher-order interactions. In both of these ways, we were simulating a situation in which the assumptions of the analysis model did not agree exactly with reality, despite the model's being close enough to reality to still be potentially useful. For simulating the outcomes of studies in the real world, this approach might be more realistic than assuming a situation in which the analysis model is exactly true. The power and Type I error results shown in the following tables are results from R software for the significance test of the effect of X_k on the pretest-adjusted Y in Models (6) and (7), rather than the effect of X_k on the change score in (8) or (9). However, for within-clusters designs as well as for between-clusters designs with only 20 members, we also fit a three-level model based on Model (8) or (9), in order to compare the results of the two modeling strategies; power did not differ much between the methods (see Appendix C).

Empirical Estimation of Power

To study the power of a test empirically, we fit the same model to each of many simulated datasets and recorded the proportion of successful rejections of the null hypothesis under α = .05. The simulated power to detect main effects and the power to detect interactions were recorded as the proportion of datasets in which the relevant treatment effect was detected as significant at the two-sided α = .05 level, averaged over the three nonzero main effects (for Factors 1, 3 and 5), or the two nonzero interactions (1×2, 1×3), accordingly. Similarly, the empirical Type I error rate for main effects or for interactions was the averaged proportion of false rejections of the null hypothesis for each of the null main effects (2 and 4), or for the null two-way interactions, respectively.

Monte Carlo Simulation Study: Results

Within-Clusters Assignment

Table 5 shows the simulated power, along with Type I error rates, for main effects in within-clusters experiments under different conditions of ICC, J and $\bar{n}$. The predicted power is also shown for comparison. As expected, larger J and $\bar{n}$ were associated with greater power. Also, simulated and predicted power usually agreed fairly well, although simulated power was slightly higher than predicted for unclear reasons (by up to about .05). For complete factorial designs, Type I error was at or below its nominal level. For fractional factorial designs, Type I error rates were slightly higher than nominal, presumably because of aliasing. Power in the within-clusters case was dependent mainly on the total number of subjects; for example, 5 clusters with about 100 members each, provided power equivalent to 10 clusters with about 50 members each.

Table 6 shows the observed and predicted power and Type I error rates for two-factor interactions in the within-clusters scenarios. The power values were considerably lower than for main effects. Simulated and predicted power agreed fairly well in most cases, although simulated power was sometimes higher by up to .10.

Between-Clusters Assignment

Table 7 shows the observed and predicted power and Type I error rate for main effects in between-clusters experiments. Type I error rate was at the nominal level for the complete factorial experiments, but slightly higher than nominal for the fractional factorial experiments, presumably because of bias caused by aliasing. Power was negatively related to ICC and positively related to J and $\bar{n}$. The number of clusters was more important than the number of members per cluster, especially in the high-ICC scenario; there, 50 clusters with about 20 members each ($N \approx 1000$) provided slightly more power than 40 clusters with about 100 members each ($N \approx 4000$). Higher ICC increased the number of clusters required to obtain adequate power, and if the ICC was very high adequate power was not obtained even with 50 clusters. Observed and predicted power agreed well.

Table 8 shows the power and Type I error rate for interactions in between-clusters experiments. In general, power to detect interactions was quite poor unless J and $\bar{n}$ were large and the ICC was close to zero. The observed and predicted power agreed well.

Discussion

The simulation results indicate that under reasonable scenarios of number of clusters, number of individuals within clusters, and ICC, it is often possible to conduct a multifactor factorial experiment with adequate power for detecting main effects, even when the population of interest is multilevel. As expected, within-clusters assignment offered better power than between-clusters assignment, but results suggest that under certain conditions it is possible to obtain excellent power for main effects in a between-clusters factorial design.

Although in some scenarios J had to be rather large to obtain adequate power, this is not a direct result of the presence of multiple factors. In fact, the simulated power per factor with five factors could be calculated well by using formulas which assumed there was only one factor, except for a df adjustment. Instead, the sample size requirements were a result of the small effect size assumed ($d \approx .2$ in terms of σ_{tot}). In a screening context, it is important to detect per-component effects that might be considerably smaller than the effect of a whole intervention. Thus it would not have been reasonable to assume a larger effect size in these simulations. Bloom et al. (2007) discuss calculation of minimum detectable effect sizes in between-clusters RCTs.

Most of the findings in the simulations agreed with past literature, but one seems initially surprising:the *increase* in Type I error rate with *higher* sample sizes (J or $\bar{n}$) in fractional factorial designs. The fact that the Type I error was often above the nominal .05 can be attributed to aliasing, but it might be surprising that having more data seems to make the inflation in the Type I error rate worse, not better. This phenomenon occurs because the bias caused by aliasing remains constant as J and $\bar{n}$ increase, although the random variability decreases. The variance estimates for the significance tests take only random variability into account and cannot measure bias. Thus, the size of the aliased coefficient relative to the standard error increases, increasing the chance of false positive results caused by bias. In other words, with larger N one has larger power to detect the (spurious) effect but, because of the aliasing, one still has no additional ability to discover that it is spurious.

Power from a Resource Management Perspective

A resource management perspective assumes that investigators wish to make decisions which keep cost (financial and otherwise) reasonable, while advancing a specific scientific agenda consisting of clearly prioritized research questions. This requires consideration of whether and how each research question can be answered with available resources and, conversely, how to make best use of available resources to answer these research questions. Consideration of statistical power is an important part of this process.

Increasing α level can increase power, at the cost of increasing Type I error risk. From a resource management perspective this tradeoff might be worthwhile in some cases, such as a factor screening experiment aimed mainly at identifying active intervention components. Here it might be sensible to tolerate a greater probability of detecting false main effects, in order to improve the ability to detect true main effects. Investigators reluctant to use a higher α might consider using α=.05 for published inference from the screening phase experiment, but keep for further study any component which is significant at the .15 or .20 level. (Alternatively, they might evaluate the components using a cost-benefit analysis, using Bayesian ideas like those in O'Hagan, Stevens & Montmartin, 2001, when planning a revised version of the intervention.) Although it can increase power, changing α does not increase the precision of estimation, so it is not a substitute for adequate sample size. However, moving intervention science forward might require making the best decisions possible under less than ideal circumstances.

Given a particular number of subjects, the same experimental design could have adequate or inadequate power, depending on the research question of interest. To illustrate how the goal of an experiment can determine its efficiency, consider a team of investigators interested in five dichotomous factors with $J = 40$ clusters available. First suppose that they wish to compare all 2^5 conditions directly to each other to empirically determine which is the best. Essentially, this involves studying all possible between-treatment comparisons or all interactions, as if to treat the study not as a 2×2×2×2×2 factorial, but rather as if it were a RCT with 32 separate conditions. This analysis would be doomed to failure. With only one or two clusters per condition (e.g., per row in Table 3), direct pairwise comparisons between *conditions* would not be feasible. For a two-condition, between-clusters RCT, about eight to ten clusters per condition is a practical minimum, and one cluster per condition is completely inadequate (see Donner & Klar, 2000; Murray, Varnell, & Blitstein, 2004; Staines, Cleland, & Blankertz, 2006; Varnell, Murray, & Baker, 2001). By contrast, suppose the the investigators wished to compare the high and low *levels* of each factor, on average across the other factors, that is, to test main effects. Each *level* of each factor has 20 clusters, so comparisons of *levels* is much more likely to have adequate power.

Two-way interactions could also be estimated in this hypothetical 2^5 factorial experiment, although perhaps with lower precision. In our simulations, power to detect interactions was poorer than power for the main effects. This was partly because the regression coefficients for interactions in the simulated scenarios were set to be smaller than those for the main effects. In practice, investigators might want to consider using an α level higher than .05 for interactions, to obtain greater power for detecting modest-sized interactions.

Investigators might be concerned about the limitations of doing a study that has sufficient power to detect main effects but poorer power for interactions or for pairwise contrasts between conditions. Whether this objection is enough to make factorial experiments infeasible in a given situation depends on one's goals and assumptions. Interactions are important for understanding how components work together, for reducing bias in estimating individual condition means, and for choosing optimal conditions (McAlister, Straus, Sackett, & Altman, 2003; Myers & Montgomery, 1995, pp. 106-7). However, especially in a screening context and when effect coding is used, main effects can still convey useful information even when less information is available about interactions. Screening experiments, especially fractional factorials, are often justified partly by a sparsity principle, a practical working assumption that most of the interesting variability in a system can be attributed to a few main effects and a few two-way interactions, and that most other effects are likely to be relatively small (see Collins et al., 2009, Wu & Hamada, 2000, for discussion). This is somewhat similar to the observation of McDonald (1997) that although a parsimonious model might not be exactly true in a hypothesis-testing sense, it might provide a good enough approximation to produce useful conclusions; for example, a small but nonzero fourth-order interaction does not necessarily mean that a model with only main effects and second-order interactions is grossly wrong. Main effects can carry useful information even in the presence of interactions, particularly when effect coding is used rather than dummy coding.

A secondary finding in the between-clusters case was that increasing the number of clusters is much more beneficial than increasing the number of members per cluster. This is consistent with existing literature (e.g., Moerbeek, van Breukelen & Berger, 2000; Moerbeek & Teerenstra, 2011). All else being equal, a sample with many small clusters provides more information than one with a few large clusters. More discussion of how to balance the costs and benefits of adding more clusters, versus adding more members to clusters, is found in Raudenbush (1997).

Power Planning Resources

In the current study we also discussed power planning resources for within- and between-clusters factorial designs. Our simulation results indicated that the formulas proposed for the within and the between-clusters assignments provide reasonable approximations to the actual power, and hence may guide investigators aiming to design multilevel factorial experiments.

For the within-clusters experiments, we proposed that investigators could obtain an initial estimate of power for a complete or fractional within-clusters factorial experiment by simply using classic power formulas for ANOVA or ANCOVA. This would be very convenient; tables and software for these methods have been available since Cohen (1988) and are relatively familiar. However, for reasons described in the Limitations section below, if an estimate of the variability of treatment effects across clusters is available, it would be better to adapt the formulas of Raudenbush and Liu (2000) instead; this may require future research.

For between-clusters experiments with pretests, we presented a power formula for Model (9), the three-level model. The other alternative, the two-level Model (7) adjusted for pretest, requires prior empirical knowledge or restrictive assumptions about the joint distibution of the pretest and posttest to estimate power (Murray, 1998). Although the power formula for Model (9) is based on different assumptions, investigators planning to use Model (7) could consider using this formula for an initial approximation (see Appendix C). If the pretest is cluster-level rather than subject-level, then formulas in Spybrook et al. (2009, p. 58) may be useful. SAS macros for calculating power based on Table 2 for either within- or between-clusters designs, and either complete or fractional factorials, are available at http://methodology.psu.edu/multilevelfactorial/ as well as in the online Appendix D.

Advantages and Disadvantages of Complete and Fractional Factorial Designs

In general, complete and fractional factorial designs were found to be equally powerful in the simulations given the same J and $\bar{n}$. In real life, various other considerations would also become important in choosing between them. For example, the fractional factorial design requires the same number of *subjects* (i.e., it does not give extra power for a given J and $\bar{n}$), but it does not require as many *conditions* (rows in the design matrix). This could make a fractional factorial cheaper and easier to implement than a complete factorial, particularly if each treatment combination requires overhead costs for creating materials, training staff, monitoring treatment quality and protocol fidelity, and so on (Collins et al., 2009). Fractional factorial designs might also be helpful in practical, ethical or political ways; for example, if carefully constructed, they might allow the investigator to omit a no-treatment group (–,–…,–), so that everyone gets at least one active treatment. In many cases, perhaps especially for studies in hospitals and schools, ethical considerations require that no condition in an experimental design be expected *a priori* to be worse than standard practice. Furthermore, statistical considerations require that the treatments for every condition be implemented faithfully, so that the factors are meaningful and the effects are detectable. Having fewer total conditions might make it easier to monitor concerns like these.

On the other hand, aliasing must be considered as a potential disadvantage of fractional factorials. The aliasing of two nonnegligible interactions with nonactive effects for the fractional factorial design increased the Type I error rates in our simulation. This is only one possible consequence of aliasing. Aliasing can cause biases in different directions, so it can either spuriously inflate or deflate test statistics. Depending on the size and direction of bias caused by the specific interactions involved in the aliasing, the power can be made lower or spuriously higher. In the simulation study, Factor 4 was aliased with the interaction of Factors 1, 2, 3, and 5. Thus, to test the main effect of Factor 4, it was necessary to assume that the interaction of Factors 1, 2, 3, and 5 was negligible. However, it was not quite zero (+0.0125). This biased the estimates for the main effect of Factor 4 away from zero slightly, slightly increasing the probability of a spurious significant result (i.e., increasing Type I error rates from .05 to perhaps .07, depending on the condition). Although the exact effects of aliasing are hard to predict, it is not difficult to use software to determine which interactions will be aliased with which effects under a given fractional factorial design, and this may make it possible to choose a fractional factorial design that is judged likely to offer an acceptable risk or amount of bias. Aliasing is described further in, for example, Wu and Hamada (2000).

Limitations and Directions for Future Research

While the focus of the current study was on power for multilevel factor screening experiments for components of intervention packages, our results can also be useful for other multilevel factorial studies in which main effects or large two-way interactions are of primary interest. In the behavioral sciences, factorial designs have typically been used to detect two- or three-way interactions between factors (e.g., Erez, Gopher, & Arzi, 1990). Because such studies are often aimed at testing interactions or comparing individual cells rather than testing main effects, the required sample size for detecting the effects of interest is likely to be very large, especially for between-clusters factorial designs (Murray, 1998; Donner & Klar,2000). Moreover, relaxing the alpha level in these studies might not be an option, unlike screening experiments in which researchers might consider an alpha level larger than .05 to increase the chance that active factors will be detected. Still, the results and formulas for main effects and two-way interactions obtained in the current study will presumably apply as well to main effects and two-way interactions in other multilevel factorial studies outside a screening context.

Models (4) through (9) can be extended in useful and straightforward ways. For instance, investigators could add additional covariates at either the cluster or individual level (Murray & Blitstein, 2003; Raudenbush, 1997). Investigators can also use a generalized linear mixed model framework if their response is non-normal, as in the case of binary or count outcomes (see Hannan & Murray, 1996; Peters, Richards, Bankhead, Ades, & Sterne, 2003; Yasui et al., 2004), rather than using the linear Models (4) through (9).

Models (4) through (9) imply that factors do not interact with the cluster-level or person-level random effects, (i.e., that effects are the same in each cluster), but it is possible for this assumption to be violated in important ways (Kraemer, 2000; Raudenbush & Liu, 2000). For example, suppose that one of the factors is a new skills training program in a school setting. The effect size of this factor could depend on how faithfully it is implemented in each school, how culturally relevant it is for the students there, or whether that school already has a similar program (in which case the new one might be redundant and ineffective). These issues cannot be described in terms of the random cluster intercept or slope effects alone. They might be addressed by adding random treatment-by-school interaction terms to the model, but that would be complicated if there were many factors, and it would usually not be practical with between-clusters data. Therefore, in between-clusters experiments, analysts might have to ignore cluster-by-treatment interactions, or try to work around them (model them indirectly by using fixed covariates or interactions with fixed covariates, and then absorb the rest into the cluster-level random slope for time or into the error term). It is easier to model and test such interactions in within-clusters experiments (Raudenbush & Liu, 2000; Moerbeek & Teerenstra, 2011). In either a within- or a between-clusters study, one can also model interactions of treatments with observable individual-level characteristics such as demographic variables or baseline risk indicators, which can also be very important for studying differential treatment effects (e.g., Conduct Problems Prevention Research Group, 2007). These non-randomized variables can be included in the multilevel regression model much as if they were additional randomized factors. One limitation is that including interactions of the factors with many predictors might lead to overparameterizing the model and/or losing statistical power, especially if the distributions of the predictors are very skewed or heavily clustered or the predictors are heavily intercorrelated.

Although our proposed power formulas often agreed well with Monte Carlo results, they have some limitations. Our approach of ignoring clustering in within-clusters designs might not be realistic if there is contamination between conditions within a cluster, and it also does not take into account cluster-by-treatment interactions, which could be modeled following Raudenbush and Liu (2000). Thus, its prediction that clustering does not reduce power at all in within-clusters designs might be too simplistic. Specifically, higher treatment-by-site variance corresponds to more noise and lower power for detecting the average treatment effect (see Raudenbush and Liu 2000), and we assumed zero treatment-by-site variance in both our theoretical formulas and simulations, so our results might be overly optimistic in this sense. In addition, floor or ceiling effects are possible in either within- or between-clusters designs, especially if clusters are small, ICC is high, and the distribution of Y is binary or skewed instead of normal. Such floor or ceiling effects may act somewhat like interactions. For example, participants who already have very high skill levels due to experiences before the study, or due to the effects of other factors in the experiment, might not experience much further benefit from a skills training component.

Our simulations were also unrealistic in some ways. For example, the random components and errors were generated from normal distributions. However, empirical data will likely have skew, missing observations, potential confounding variables, perhaps poor measurement, and so on. More research is needed on how best to take these considerations into account when planning experiments. Also, our formulas and simulations considered power and Type I error only for each component separately, rather than experimentwise Type I or Type II error. For example, if an experimenter wanted to have an 80% probability of detecting *all* effects of size .2 or greater (without missing any), this would require a larger sample size than an 80% power for each component considered separately. Furthermore, the power formulas do not account for aliasing in the fractional factorial case; this limitation is unavoidable without assumptions about the size and nature of the aliasing. Finally, we focused only on power for null hypothesis tests, not on other goals such as precision of estimates and confidence intervals. These might be very important in planning interventions, perhaps even more important than power for hypothesis tests. However, the power to detect an effect and the precision to estimate it are closely related, and ideas discussed in this paper should generalize from one to the other. Researchers may also choose to do their own simulations, more specific or more advanced than those in this paper, to address specific questions.

In this article we have compared only within- to between-clusters factorial designs, but it is possible to have experiments in which some factors are between-clusters and others are within-clusters, analogous to repeated-measures studies in which some variables are within-subjects and others are between-subjects. This might be important if some factors can be assigned at the individual level while others can be assigned only at the group level. Factors which must be applied at the cluster level will, of course, have to be treated as between-clusters, but if a factor can be practically assigned at either level, there is a potential power advantage for assigning it at the individual level. Experiments might also have three levels of nesting (e.g., student within classroom within school), in which case relevant power formulas could presumably be adapted from, for example, Teerenstra, Moerbeek, van Achterberg, Pelzer, and Borm (2008), and Spybrook et al. (2009). Furthermore, sometimes it is difficult to characterize whether a factor is within- or between-clusters. Investigators may assign independent individuals to conditions but then deliver these conditions in a group discussion or group therapy setting that generates clustering (e.g., Adarves-Yorno, Postmes, & Haslam,2007; Antonio, Chang, Hakuta, Kenny, Levin, & Milem, 2004; Roberts and Roberts, 2005). This topic requires more methodological research.

Recommendations

If an investigator is interested only in comparing a predetermined program with standard practice, then a two-condition RCT is likely to be the most easy, direct, natural, and powerful choice. However, if one is interested in studying the effectiveness of different program components, then a factorial experiment or some reduced form thereof becomes important to consider, even when the population is clustered. Based on our simulation findings and the resource management perspective, we offer the following recommendations for carrying out factorial or fractional factorial experiments in a clustered setting. First, complete factorial and fractional factorial experiments both have advantages and disadvantages. The former may be more costly to conduct, whereas the latter poses the risk of aliasing. Second, if the risk of contamination between differently treated units is low, and the treatment works at the individual rather than the cluster level, assigning lower-level units (e.g., students) individually to conditions has a considerable advantage in power for a given total sample size over assigning clusters (e.g., schools) to conditions. Assignment of whole clusters to treatments may still be necessary if the treatment is designed to affect the cluster as a whole, or if contamination is a concern (Slymen & Hovell 1997), but a larger investment in resources (e.g., high J) will be required for successful results. Third, for cases in which the investigator chooses to assign clusters to conditions but does not have access to an adequate number of clusters to allow a complete factorial design, fractional factorial designs may make a screening study with multiple factors possible.

The idea of cluster-randomized factor screening trials is an example of the challenges and potential rewards of bringing together perspectives from different disciplines. Research methods commonly applied in areas such as engineering may offer new perspectives and insights for studying complex systems and for optimizing the effectiveness and cost-effectiveness of processes, including therapies and interventions (e.g., Rivera, Pew, & Collins, 2007). However, expertise from the social and health sciences is needed to adapt these methods to the practical, ethical and statistical challenges of studying humans and changing, contextually nested social systems. If brought together, these different disciplinary perspectives may shed new light on complex behavioral and social phenomena and advance research in behavioral interventions to promote health.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NIDA grants P50 DA10075-14 and K05 DA018206.

The authors thank Bethany C. Bray, Donna L. Coffman, Suellen Hopfer, Stephanie T. Lanza, Tina Meyers, YoungKyoung Min, Rachel Smith, Xianming Tan, and especially Ayala Cohen and Megan Piper, for very helpful discussions, reviews, and comments on earlier drafts. They also thank Kari Kugler and Jessica Trail for their investigations into the handling of CLASS variables in factorial designs in SAS. They thank Amanda Applegate for her expert proofreading and editing help.

Data analysis was done using the R 2.11 and 2.12 and SAS 9.1 and 9.2 packages. The R software is copyright 2010 by The R Foundation for Statistical Computing. SAS software is copyright 2002-2010 by SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Appendix A

Coding of Categorical Variables

Effect Coding and Dummy Coding

We represent the On/Off levels for each factor using effect coding (e.g., -1 for Off and +1 for On) rather than dummy coding (e.g., 0 for Off and +1 for On). To clarify the difference, consider the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$ with two dichotomous factors. With dummy coding, the coefficient β_1 expresses the difference between the two levels of X_1, at the reference level (Off) of X_2, so only half of the data (those participants assigned to the Off level of X_2) informs β_1. However, with effect coding, the coefficient β_1 expresses the average difference between the levels of X_1 across the levels of X_2, and all of the data informs β_1. If in addition the assignment of units to conditions is perfectly balanced, effect coding allows the coefficients to be treated as orthogonal contrasts (Myers & Well, 2003, pp. 614-623). In practice, it is very unlikely that factors would be perfectly orthogonal because of unequal sample sizes in each condition, but effect coding still facilitates straightforward interpretation of coefficients.

The regression coefficient for X_k in +1/-1 coding represents ½ the difference in expected Y between $X_k = +1$ and $X_k = -1$, i.e., ½ the main effect of factor k (Moerbeek & Teerenstra, 2011). The regression coefficient for $X_k X_{k'}$ represents ¼ the interaction between factors k and k'. For example, in a 2×2 factorial, let μ_{lm} be the mean of the condition having level l for the row factor and m for the column factor. The main effect of the first factor is a difference between row means, for example, $ME=\frac{\mu_{11}+\mu_{12}}{2}-\frac{\mu_{21}+\mu_{22}}{2}=\frac{1}{2}\mu_{11}+\frac{1}{2}\mu_{12}-\frac{1}{2}\mu_{21}-\frac{1}{2}\mu_{22}$. Under the coding +1 and -1, however, the regression coefficient is half this difference: $\gamma=\frac{1}{4}\mu_{11}+\frac{1}{4}\mu_{12}-\frac{1}{4}\mu_{21}-\frac{1}{4}\mu_{22}=\frac{1}{2}ME$. In contrast, the interaction is expressed as a difference between pairwise differences, $IXN = (\mu_{11} - \mu_{21}) - (\mu_{12} + \mu_{22}) = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$ but the regression coefficient for the interaction has a similar scale to that of the main effect, $\gamma=\frac{1}{4}\mu_{11}-\frac{1}{4}\mu_{12}-\frac{1}{4}\mu_{21}+\frac{1}{4}\mu_{22}=\frac{1}{4}IXN$, because both are based on half of a difference in cell averages (whether between row and row, column and column, or diagonal and diagonal). This scaling difference, in which an IXN must be twice the size of a ME in order to have the same γ coefficient, also explains why the sampling error and required sample size for a two-way IXN is said to be $2^2 = 4$ times that of a comparable ME (Murray, 1998; Montgomery et al., 2003; Myers & Well 2003, p. 305). The power for a given regression coefficient is the same regardless of whether it represents a main effect or an interaction (assuming balanced cell sizes and effect coding), but an IXN of a given magnitude nonetheless implies a regression coefficient half the size of a ME of that magnitude and is therefore harder to detect.

The Scale of the X Variables

Raudenbush (1997), Raudenbush and Liu (2000; 2001) and Spybrook et al. (2009) do something similar to effect coding but represent the levels of X as +½ and -½ . Because the difference between high and low is now 1 unit instead of 2, the regression parameter is directly interpretable as the treatment effect. For example, in a case without a pretest, their γ is the raw difference between treatment groups. If the total variance, e.g., $\tau^2 + \sigma^2$ in (4) or (5), is standardized to 1, then their γ then becomes identical to Cohen's d. Murray and Blitstein (2003) similarly treat the levels of X as though they were one unit apart.

However, we use +1 and -1 as the levels of X in order to be more comparable with older work on factorial and fractional factorial designs; thus the regression coefficient is half the effect size. Representing interactions as products of effect codes could become more complicated and less intuitive if the codes had fractional values. Furthermore, it would not remove the discrepancy of scale between main and interaction effects.

Because Raudenbush (1997) and others, including Spybrook et al. (2009), code the levels of X as +½ and -½ while we code them as +1 and -1, our γ parameters are half the size of the corresponding γ parameters in their notation. This implies that their variance of γ is multiplied by 2^2 relative to ours. This does not matter for power, because the multipliers cancel out in the expression for the noncentrality parameter (i.e., both numerator and denominator are multiplied by 4). However, the distinction is important when comparing formulas.

Centering Time

Time (pretest versus posttest) also needed to be coded, but was treated differently from the treatment factors. As mentioned in the text, we coded time as $T_{tij} = -$ ½ for pretest and $T_{tij} =$ + ½ for posttest, following Raudenbush and Liu (2001). Centering time can help to reduce collinearity, as in Raudenbush and Liu (2001), but it can make the notation more complicated.

A simpler approach in our linear growth case could have been to let the pretest be $T_{tij} = 0$ and the posttest be $T_{tij} = 1$, somewhat like Bryk and Raudenbush (1988). Under this simpler approach, the γ_{1k0} parameters would retain the same interpretation, but the γ_{0k0} parameters would now have a very straightforward meaning: the treatment effect on the pretest. Under the assumption of random assignment, the treatment effect on the *pretest* (taken before the treatment is provided) should of course be zero. Thus one could either *a priori* constrain $\gamma_{010} = \ldots = \gamma_{0K0} = 0$ to simplify the model, or perhaps test whether $\gamma_{010} = \ldots = \gamma_{0K0} =$ as a check on internal validity. However, a possible disadvantage to this approach would be that the random cluster-by-time effects u_{10j} are multiplied by 0 for the pretest and 1 for the posttest. Thus, even if the treatment has no effect, the 0/1 coding requires the strong assumption that the ICC and overall variance of the response are both higher at posttest than at pretest. This leads to a conceptually different model which is not pursued in this article.

Appendix B

Selection of Variance Parameters

To simulate data from Model (8) or Model (9), it is necessary to choose population values for four variance parameters: $\tau^2_{\pi 0}$, $\tau^2_{\beta 0}$, $\tau^2_{\beta 1}$, and σ^2 (recall that $\tau^2_{\pi 1}$ was set to zero for identifiability-related reasons). These parameters seem very abstract, and there is relatively little literature to suggest realistic values for them directly. However, one can choose reasonable values for $\tau^2_{\pi 0}$, $\tau^2_{\beta 0}$, $\tau^2_{\beta 1}$, and σ^2 indirectly by following four steps.

1. By plugging $T_{tij}=-\frac{1}{2}$ or $T_{tij}=+\frac{1}{2}$ into (8) or (9) it can be seen that the marginal variance of Y_{tij} at pretest or posttest is $\sigma^2_{\text{tot}}=\text{Var}\left(u_{00j}+\frac{1}{2}u_{10j}+r_{0ij}+e_{tij}\right)=\tau^2_{\beta 0}+\frac{1}{4}\tau^2_{\beta 1}+\tau^2_{\pi 0}+\sigma^2$. For simplicity, we assume that the data is standardized in such a way that this marginal variance σ^2_{tot} is 1. σ^2_{tot} can be interpreted as follows: If there are no treatment effects, then σ^2_{tot} would be close to the naïve sample variance of Y at either pretest or posttest, ignoring information about who belonged to what cluster.

2. By plugging $T_{tij}=-\frac{1}{2}$ or $T_{tij}=+\frac{1}{2}$ into (8) or (9), it can also be seen that the ICC either of the pretest or of the posttest unadjusted for the pretest, and assuming no treatment effects, is

$$\rho_{\text{pre}}=\frac{\text{Cov}\left(Y_{tij},Y_{t i' j}\right)}{\text{Var}(Y_{tij})}=\frac{\tau^2_{\beta0}+\frac{1}{4}\tau^2_{\beta1}}{\tau^2_{\beta0}+\frac{1}{4}\tau^2_{\beta1}+\tau^2_{\pi0}+\sigma^2}, \tag{10}$$

which by step 1 is just $\tau^2_{\beta0}+\frac{1}{4}\tau^2_{\beta1}$. Thus, we set $\tau^2_{\beta0}+\frac{1}{4}\tau^2_{\beta1}$ to equal the unadjusted ICC chosen for the scenario; that is, .05 for low, .15 for medium or .30 for high.

3. By first finding the pretest Y_{-ij} and posttest Y_{+ij} under (8) or (9) and then finding the difference $Y_{+ij}-Y_{-ij}$ (i.e., the change or gain score), it can be seen that the random part of these change scores is $u_{10j}+e_{+ij}-e_{-ij}$. Thus, assuming no treatment effects, their variance is $\tau^2_{\beta1}+2\sigma^2$ and their ICC is $\rho_{\text{change}}=\tau^2_{\beta1}/\left(\tau^2_{\beta1}+2\sigma^2\right)$. The ICC of the change scores might be more important for the power of the tests of interest than the ICC of the pretest or posttest alone (see Murray & Blitstein, 2003, for the distinction). Although there is little empirical information available, it is possible that ρ_{change} might be considerably less than ρ_1, because some of the cluster-level variance shared by the posttest and pretest might be removed by taking the difference. Therefore, we somewhat arbitrarily set $\rho_{\text{change}}=\frac{1}{2}\rho_{\text{pre}}$; that is, .025 for low, .075 for medium or .15 for high.

4. The correlation of the pretest and posttest conditional upon cluster (i.e., Corr(Y_{-ij}, Y_{+ij}) within a given cluster j), can be shown to be $\rho_{\text{pre,post}}=\tau^2_{\pi0}/\left(\tau^2_{\pi0}+\sigma^2\right)$. This could be thought of as a pretest-posttest reliability in the absence of treatment, or as a subject-level ICC. It is somewhat analogous to α_p in Raudenbush and Liu (2001). There is some empirical information about pretest-posttest reliability, particularly in educational fields, although it obviously depends heavily upon the variable and population. For example, Randall and Engelhard (2010) reported several pretest-posttest correlations, averaging about .60. Schochet (2008) considered $\rho^2_{\text{pre,post}}=.50$ or .70 on achievement tests as reasonable values; that is, $\rho_{\text{pre,post}}$ around .7 or .8. In the context of a somewhat different model, Hedges and Hedburg (2007a) similarly reported results suggesting a typical $\rho^2_{\text{pre,post}}$ around .60 on math or reading achievement tests. However, other measures with poorer psychometric properties or very skewed distributions might have lower $\rho_{\text{pre,post}}$. Therefore, we considered a $\rho_{\text{pre,post}}$ of .65 to be reasonable and set $\frac{\tau^2_{\pi0}}{\tau^2_{\pi0}+\sigma^2}=.65$.

As a result of these four steps, we had four constraints on four unknowns. Thus, we could use algebra to back-calculate the needed variance components as $\sigma^2=(1-\rho_{\text{pre,post}})(1-\rho_{\text{pre}})$, $\tau^2_{\pi0}=\rho_{\text{pre,post}}(1-\rho_{\text{pre}})$, $\tau^2_{\beta1}=2\sigma^2\frac{\rho_{\text{change}}}{1-\rho_{\text{change}}}$, $\tau^2_{\beta0}=1-\frac{1}{4}\tau^2_{\beta1}-\tau^2_{\pi0}-\sigma^2$. In more generality, if σ^2_{tot} is not assumed to be 1, simply multiply each of these four expressions by the desired σ^2_{tot}.

For the scenarios in the simulation, the variance components were therefore chosen to be $\sigma^2=.3325$, $\tau^2_{\pi0}=.6175$, $\tau^2_{\beta0}=.0457$, $\tau^2_{\beta1}=.0171$ for low ICC, $\sigma^2=.2975$, $\tau^2_{\pi0}=.5525$, $\tau^2_{\beta0}=.1379$, $\tau^2_{\beta1}=.0482$ for medium and $\sigma^2=.2450$, $\tau^2_{\pi0}=.4550$, $\tau^2_{\beta0}=.2784$, $\tau^2_{\beta1}=.0864$ for high ICC.

The main effects were set to be $.4\sigma=.4\sqrt{\sigma^2}$ which comes to 0.2306 σ^2_{tot}, 0.2182 σ^2_{tot}, or 0.1980 σ^2_{tot}. Because σ^2_{tot} was set to one, the effects are also 0.2306, 0.2182, or 0.1980 in raw units. It might seem strange that although we chose to standardize the variances in terms of σ^2_{tot}, not σ^2, in deriving the power formulas here, we nonetheless held the effect sizes constant in terms of σ^2, not σ^2_{tot}, in the simulation. It seemed that standardizing by σ^2_{tot} was more intuitive for a power formula because, regardless of model, it could always be interpreted as the marginal population variance of the response variable in the absence of treatment effects, while the interpretation of σ^2 is slightly different in Models (4), (5), (6), and (9). However, from previous simulations we found that in the within-clusters cases, holding effect sizes constant across scenarios in terms of σ^2_{tot} led to results which were difficult to compare across conditions. The higher the ICC for fixed σ^2_{tot}, the higher τ^2 and hence the lower σ^2 must be. In the within-clusters cases power depended only on τ^2 and not on σ^2. Thus, unless we held effects constant with respect to σ^2 in the within-clusters scenarios, we obtained the paradoxical result that higher ICC seemed to spuriously increase power, because it was essentially confounded with increased effect size in terms of the most important variance parameter.

Appendix C

Comparison of Performance Between Methods

The simulated data were analyzed in both SAS and R, using models (6) or (7) for within and between clusters, respectively. Except for between-clusters scenarios with cluster size 100, the data were also analyzed using Model (8) or (9) in SAS. The two three simulated power estimates available in each scenario agreed very well with each other although there were small discrepancies. Discrepancies among the methods in simulated power for main effects or for interactions were always less than .03 and most were less than .01 in absolute value. The three scenarios with largest discrepancies were all high-ICC, between-clusters, $\bar{n}$ =20, cases. In the J=40, complete factorial case, power was .402 for the adjusted two-level model in R, .402 for the adjusted two-level model in SAS, and .4225 for the three-level model in SAS. In the J =50, fractional factorial case, power estimates were .515, .5235, and .544; and in the J=50, complete factorial case, power estimates were .492, .4985, and .514. The scenario with largest absolute discrepancy among the methods in simulated power for interactions was the low-ICC, J = 25, $\bar{n}$ =20, fractional factorial between-clusters case, with power estimates of .181, .196, and .175. Last, absolute discrepancies among scenarios in Type I error rate for main effects or for interactions were always less than .02 and most were below .005. The largest absolute discrepancy among methods for either of these error rates was in the low-ICC, J = 25, $\bar{n}$ =20, fractional factorial between-clusters case, with simulated Type I error rates of .0377, .0524, and .0366 for main effects and .0405, .0519, and .0377 for interactions. Because the differences were generally very small, in order to save space only the adjusted two-level R results are shown in Tables 5 through 8.

References

Abernethy AP, Currow DC, Hunt R, Williams H, Rowett D, Roder-Allenand G, Phillips PA, et al. A pragmatic 2×2×2 factorial cluster randomized controlled trial of educational outreach visiting and case conferencing in palliative care — Methodology of the Palliative Care Trial. Contemporary Clinical Trials. 2006; 27:83–100. [PubMed: 16290094]

Adarves-Yorno I, Postmes T, Haslam SA. Creative innovation or crazy irrelevance? The contribution of group norms and social identity to creative behavior. Journal of Experimental Social Psychology. 2007; 43:410–416.

Ahn H, Wampold B. Where oh where are the specific ingredients? A meta-analysis of component studies in counseling and psychotherapy. Journal of Counseling Psychology. 2001; 48:251–257.

Allison PD. Change scores as dependent variables in regression analysis. Sociological Methodology. 1990; 20:93–114.

Allore HG, Tinetti ME, Gill TM, Peduzzi PN. Experimental designs for multicomponent interventions among persons with multifactorial geriatric syndromes. Clinical Trials. 2005; 2:13–21. [PubMed: 16279575]

Antonio AL, Chang MJ, Hakuta K, Kenny DA, Levin S, Milem JF. Effects of racial diversity on complex thinking in college students. Psychological Science. 2004; 15:507–10. [PubMed: 15270993]

Apfel CC, Korttila K, Abdalla M, Biedler A, Kranke P, Pocock SJ, Roewer N. An international multicenter protocol to assess the single and combined benefits of antiemetic interventions in a controlled clinical trial of a 2×2×2×2×2×2 factorial design (IMPACT). Controlled Clinical Trials. 2003; 24:736–751. [PubMed: 14662280]

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

Bloom HS, Richburg-Hayes L, Black AR. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educational Evaluation & Policy Analysis. 2007; 29:30–59.

Brent D, Emslie G, Clarke G, Dineen Wagner K, Rosenbaum Asarnow J, Keller M, Zelazny J, et al. Switching to another SSRI or to venlafaxine with or without cognitive behavioral therapy for adolescents with SSRI-resistant depression: The TORDIA randomized controlled trial. Journal of the American Medical Association. 2008; 299:901–913. [PubMed: 18314433]

Bryk AS, Raudenbush SW. Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. American Journal of Education. 1988; 97:65–108.

Campbell M, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: The case of implementation research. Clinical Trials. 2005; 2:99–107. [PubMed: 16279131]

Casella, G.; Berger, RL. Statistical inference. Belmont, CA: Duxbury; 1990.

Chakraborty B, Collins LM, Strecher VJ, Murphy SA. Developing multicomponent interventions using fractional factorial designs. Statistics in Medicine. 2009; 28:2687–2708. [PubMed: 19575485]

Chakravorti SR, Grizzle JE. Analysis of data from multiclinic experiments. Biometrics. 1975; 31:325–338. [PubMed: 1100133]

Cohen, J. Statistical power analysis for the behavioral sciences. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

Collins LM, Baker TB, Mermelstein RJ, Piper ME, Jorenby DE, Smith SS, Schlam TR, Cook JW, Fiore MC. The Multiphase Optimization Strategy for engineering effective tobacco use interventions. Annals of Behavioral Medicine. 2011; 41:208–226. [PubMed: 21132416]

Collins LM, Dziak JJ, Li R. Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. Psychological Methods. 2009; 14:202–224. [PubMed: 19719358]

Collins LM, Murphy SA, Nair V, Strecher V. A strategy for optimizing and evaluating behavioral interventions. Annals of Behavioral Medicine. 2005; 30:65–73. [PubMed: 16097907]

Collins LM, Murphy SA, Strecher V. The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New methods for more potent e-health interventions. American Journal of Preventive Medicine. 2007; 32:S112–S118. [PubMed: 17466815]

Conduct Problems Prevention Research Group. Fast Track randomized controlled trial to prevent externalizing psychiatric disorders: Findings from grades 3 to 9. Journal of the American Academy of Child and Adolescent Psychiatry. 2007; 46:1250–1262. [PubMed: 17885566]

Dean AM, Lewis SM. Comparison of group screening strategies for factorial experiments. Computational Statistics and Data Analysis. 2002; 39:287–297.

Dee TS, West MR. The non-cognitive returns to class size. Educational Evaluation and Policy Analysis. 2011; 33:23–46.

Donner, A.; Klar, N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.

Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International Journal of Epidemiology. 2006; 35:1292–1300. [PubMed: 16943232]

Erez M, Gopher D, Arzi N. Effects of goal difficulty, self-set goals and monetary rewards on dual task performance. Organizational Behavior and Human Decision Processes. 1990; 47:247–269.

Fielding JE, Mason T, Knight K, Klesges R, Pelletier KR. A randomized trial of the IMPACT worksite cholesterol reduction program. American Journal of Preventive Medicine. 1995; 11:120–123. [PubMed: 7632447]

Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, Ji P, et al. Standards of evidence: Criteria for efficacy, effectiveness and dissemination. Prevention Science. 2005; 6:151–175. [PubMed: 16365954]

Flay BR, Collins LM. Historical review of school-based randomized trials for evaluating problem behavior prevention programs. Annals of the American Academy of Political and Social Science. 2005; 599:115–146.

Frasure-Smith N, Koszycki D, Swensen JR, Baker B, van Zyl LT, Laliberté M, Lesperance F, et al. Design and rationale for a randomized, controlled trial of interpersonal psychotherapy and citalopram for depression in coronary artery disease (CREATE). Psychosomatic Medicine. 2006; 68:87–93. [PubMed: 16449416]

Gaudine AP, Saks AM. Effects of an absenteeism feedback intervention on employee absence behavior. Journal of Organizational Behavior. 2001; 22:15–29.

Gomel M, Oldenburg B, Simpson JM, Owen N. Work-site cardiovascular risk reduction: A randomized trial of health risk assessment, education, counseling, and initiatives. American Journal of Public Health. 1993; 83:1231–8. [PubMed: 8362997]

Hannan PJ, Murray DM. Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed-model and the logistic mixed-model analyses in simulated community trials with a dichotomous outcome variable at the individual level. Evaluation Review. 1996; 20:338–352. [PubMed: 10182208]

Hedges LV, Hedberg EC. Intraclass correlation values for planning group-randomized experiments in rural education. Journal for Research in Rural Education. 2007a; 22(10):1–15.

Hedges LV, Hedburg EC. Intraclass correlation values for planning cluster-randomized trials in education. Educational Evaluation and Policy Analysis. 2007b; 29:60–87.

Hundert J, Boyle MH, Cunningham CE, Duku E, Heale J, McDonald J, Racine Y, et al. Helping children adjust—a Tri-Ministry Study: II. Program effects. Journal of Child Psychology and Psychiatry. 1999; 40:1061–73. [PubMed: 10576536]

Janega JB, Murray DM, Varnell SP, Blitstein JL, Birnbaum AS, Lytle LA. Assessing intervention effects in a school-based nutrition intervention trial: Which analytic model is most powerful? Health Education and Behavior. 2004; 31:756–774. [PubMed: 15539546]

Kirk, R. Experimental design: Procedures for the behavioral sciences. 3. Pacific Grove, CA: Brooks/Cole; 1995.

Kish, L. Survey sampling. New York: Wiley; 1965.

Kraemer HC. Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. Schizophrenia Bulletin. 2000; 26:533–541. [PubMed: 10993394]

Kwan J, Hand P, Dennis M, Sandercock P. Effects of introducing an integrated care pathway in an acute stroke unit. Age and Ageing. 2004; 33:362–367. [PubMed: 15047573]

Lespérance F, Frasure-Smith N, Koszycki D, Laliberté M, van Zyl LT, Baker B, Guertin M-C, et al. Effects of citalopram and interpersonal psychotherapy on depression in patients with coronary artery disease: The Canadian Cardiac Randomized Evaluation of Antidepressant and Psychotherapy Efficacy (CREATE) trial. Journal of the American Medical Association. 2007; 297:367–379. [PubMed: 17244833]

Liu X, Spybrook J, Congdon R, Martinez A, Raudenbush SW. Optimal design for longitudinal and multilevel research, V.2.0. 2009 [Software].

Luepker RV, Perry CL, McKinlay SM, Nader PR, Parcel GS, Stone EJ, Verter J, et al. Outcomes of a field trial to improve children's dietary patterns and physical activity: The Child and Adolescent Trial for Cardiovascular Health (CATCH). Journal of the American Medical Association. 1996; 275:768–776. [PubMed: 8598593]

Ma X, Klinger DA. Hierarchical linear modelling of student and school effects on academic achievement. Canadian Journal of Education. 2000; 25:41–55.

McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: A systematic review. Journal of the American Medical Association. 2003; 289:2545–2553. [PubMed: 12759326]

McDonald, RP. Goodness of approximation in the linear model. In: Harlow, LL.; Mulaik, SA.; Steiger, JH., editors. What if there were no significance tests?. Mahwah, NJ: Erlbaum; 1997. p. 199-220.

Midgley G. Systemic intervention for public health. Journal of Public Health. 2006; 3:466–472.

Moerbeek, M.; Teerenstra, S. Optimal design in multilevel experiments. In: Hox, JJ.; Roberts, JK., editors. Handbook of advanced multilevel analysis. New York: Routledge; 2011. p. 257-284.

Moerbeek M, van Breukelen GJP, Berger MPF. Design issues for experiments in multilevel populations. Journal of Educational and Behavioral Statistics. 2000; 25:271–284.

Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Medical Research Methodology. 2003; 3:26. [PubMed: 14633287]

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

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

Murray DM, Feldman HA, McGovern PG. Components of variance in a cluster-randomized trial analysed via a random-coefficients model: The Rapid Early Action for Coronary Treatment (REACT) trial. Statistical Methods in Medical Research. 2000; 9:117–133. [PubMed: 10946430]

Murray DM, Hannan PJ, Pals SP, McCowen RG, Baker W, Blitstein JL. A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a cluster-randomized trial. Statistics in Medicine. 2006; 25:375–388. [PubMed: 16143991]

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

Myers, JL.; Well, AD. Research design and statistical analysis. 2. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2003.

Myers, RH.; Montgomery, DC. Response surface methodology: process and product optimization using designed experiments. New York: Wiley; 1995.

Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. Applied linear statistical models. Boston: McGraw-Hill; 1996. Neuhaus JM, Segal MR. Design effects for binary regression models fitted to dependent data. Statistics in Medicine. 1993; 12:1259–1268. [PubMed: 8210825]

O'Hagan A, Stevens JW, Montmartin J. Bayesian cost-effectiveness analysis from clinical trial data. Statistics in Medicine. 2001; 20:733–53. [PubMed: 11241573]

Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JAC. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. International Journal of Epidemiology. 2003; 32:840–846. [PubMed: 14559762]

Pinheiro J, Bates D, DebRoy S, Sarkar D. The R Development Core Team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-97. 2010

Pituch KA, Miller JW. Strengthening multisite educational interventions: An illustration with multilevel modeling. Educational Research & Evaluation. 1999; 5:62–75.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

Randall J, Engelhard G. Performance of students with and without disabilities under modified conditions: Using resource guides and read-aloud test modifications on a high-stakes reading test. Journal of Special Education. 2010; 44:79–93.

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

Raudenbush SW, Liu X. Statistical power and optimal design for multisite randomized trials. Psychological Methods. 2000; 5:199–213. [PubMed: 10937329]

Raudenbush SW, Liu X. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. Psychological Methods. 2001; 6:387–401. [PubMed: 11778679]

Raudenbush SW, Martinez A, Spybrook J. Strategies for improving precision in cluster-randomized experiments. Educational Evaluation and Policy Analysis. 2007; 29:5–29.

Resick PA, Galovski TE, O'Brien Uhlmansiek M, Scher CD, Clum GA, Young-Xu Y. A randomized clinical trial to dismantle components of cognitive processing therapy for posttraumatic stress disorder in female victims of interpersonal violence. Journal of Consulting and Clinical Psychology. 2008; 76:243–258. [PubMed: 18377121]

Rivera DE, Pew MD, Collins LM. Using engineering control principles to inform the design of adaptive interventions. Drug and Alcohol Dependence. 2007; 88:S31–S40. [PubMed: 17169503]

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

Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. Journal of Epidemiology and Community Health. 2002; 56:119–127. [PubMed: 11812811]

SAS Institute. SAS/QC 9.1 user's guide. Cary, NC: Author; 2004.

Schochet, PZ. The late pretest problem in randomized control trials of education interventions (NCEE 2009-4033). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education; 2008.

Scott AJ, Holt D. The effect of two-stage sampling on ordinary least squares methods. Journal of the American Statistical Association. 1982; 77:848–854.

Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based smoking prevention study: Outcome and mediating variables by gender and ethnicity. American Journal of Epidemiology. 1996; 144:425–433. [PubMed: 8712201]

Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. Journal of Educational and Behavioral Statistics. 1998; 24:323–355.

Slymen DJ, Hovell MF. Cluster versus individual randomization in adolescent tobacco and alcohol studies: Illustrations for design decisions. International Journal of Epidemiology. 1997; 26:765–771. [PubMed: 9279608]

Snijders, TAB.; Bosker, RJ. Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage; 1999.

Spybrook, J.; Raudenbush, SW.; Congdon, R.; Martinez, A. Optimal Design for longitudinal and multilevel research: Documentation for the Optimal Design software V.2.0. 2009. Retrieved from www.wtgrantfoundation.org

Staines GL, Cleland CM, Blankertz L. Counselor confounds in evaluations of vocational rehabilitation methods in substance dependency treatment. Evaluation Review. 2006; 30:139–170. [PubMed: 16492996]

Teerenstra S, Moerbeek M, van Achterberg T, Pelzer BJ, Borm GF. Sample size calculations for 3-level cluster randomized trials. Clinical Trials. 2008; 5:486–495. [PubMed: 18827041]

Van Breukelen GJP, Candel MJ, Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Statistics in Medicine. 2007; 26:2589–2603. [PubMed: 17094074]

Varnell SP, Murray DM, Baker WL. An evaluation of analysis options for the one-group-per condition design: Can any of the alternatives overcome the problems inherent in this design? Evaluation Review. 2001; 25:440–453. [PubMed: 11480307]

Wachelka D, Katz R. Reducing test anxiety and improving academic self-esteem in high school and college students with learning disabilities. Journal of Behavior Therapy and Experimental Psychiatry. 1999; 30:191–8. [PubMed: 10619543]

Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. Mixed treatment comparison meta-analysis of complex interventions: Psychological interventions in coronary heart disease. American Journal of Epidemiology. 2009; 169:1158–1165. [PubMed: 19258485]

West, SG.; Aiken, LS. Toward understanding individual effects in multicomponent prevention programs. In: Bryant, K.; Windle, M.; West, S., editors. The science of prevention: Methodological advances from alcohol and substance abuse research. Washington, D. C.: American Psychological Association; 1997. p. 167-209.

West SG, Aiken LS, Todd M. Probing the effects of individual components in multiple component prevention programs. American Journal of Community Psychology. 1993; 21:571–605. [PubMed: 8192123]

Williams A, Hagerty BM, Andrei AC, Yousha SM, Hirth RA, Hoyle KS. STARS: Strategies to assist Navy recruits' success. Military Medicine. 2007; 172:942–9. [PubMed: 17937357]

Wu, CFJ.; Hamada, M. Experiments: Planning, analysis and parameter design optimization. New York, NY: Wiley; 2000.

Yasui Y, Feng Z, Diehr P, McLerran D, Beresford SA, McCulloch CE. Evaluation of community-intervention trials via generalized linear mixed models. Biometrics. 2004; 60:1043–52. [PubMed: 15606425]

Main Effect 1

(Low)					(High)			
1	***2***	***3***	4		17	***18***	***19***	20
5	6	7	***8***	versus	***21***	22	23	***24***
9	10	11	***12***	↔	***25***	26	27	***28***
13	***14***	***15***	16		29	***30***	***31***	32

Main Effect 2

1	***2***	***3***	4		9	***10***	***11***	12
5	6	7	***8***	versus	***13***	14	15	***16***
17	18	19	***20***	↔	***25***	26	27	***28***
21	***22***	***23***	24		29	***30***	***31***	32

Main Effect 3

1	***2***	***3***	4		5	***6***	***7***	8
9	10	11	***12***	versus	***13***	14	15	***16***
17	18	19	***20***	↔	***21***	22	23	***24***
25	***26***	***27***	28		29	***30***	***31***	32

Main Effect 4

1	***2***	***5***	6		3	***4***	***7***	8
9	10	13	***14***	versus	***11***	12	15	***16***
17	18	21	***22***	↔	***19***	20	23	***24***
25	***26***	***29***	30		27	***28***	***31***	32

Main Effect 5

1	***3***	***5***	7		2	***4***	***6***	8
9	11	13	***15***	versus	***10***	12	14	***16***
17	19	21	***23***	↔	***18***	20	22	***24***
25	***27***	***29***	31		26	***28***	***30***	32

Figure 1.

Subsets of conditions in Table 4 used to test each main effect. Bolded conditions are retained in the fractional factorial.

Table 1

Effect Coding for the 2 × 2 × 2 Factorial Design in the Relaxation, Cognitive, and Imagery Component Example

Condition	Components	R	C	I	R×C	R×I	C×I	R×C×I
Conditions in Complete Factorial								
1	Untreated	-1	-1	-1	+1	+1	+1	-1
2	**I** only	-1	-1	+1	+1	-1	-1	+1
3	**C** only	-1	+1	-1	-1	+1	-1	+1
4	**I** and **C**	-1	+1	+1	-1	-1	+1	-1
5	**R** only	+1	-1	-1	-1	-1	+1	+1
6	**R** and **I**	+1	-1	+1	-1	+1	-1	-1
7	**R** and **C**	+1	+1	-1	+1	-1	-1	-1
8	All three	+1	+1	+1	+1	+1	+1	+1
Conditions Retained in Fractional Factorial								
2	**I** only	-1	-1	+1	+1	-1	-1	+1
3	**C** only	-1	+1	-1	-1	+1	-1	+1
5	**R** only	+1	-1	-1	-1	-1	+1	+1
8	All three	+1	+1	+1	+1	+1	+1	+1

Note. **R** = **Relaxation**, **C** = **Cognitive**, **I** = **Imagery**.

Table 2

Summary of Proposed Power Formulas

			Noncentrality Parameter In Terms Of	
Pretest?	**Assignment**	**Model**	**In Terms Of Regression Parameter γ**	**In Terms Of $d = 2\gamma / \sigma_{tot}$**
No	Within	(4)	$\frac{N\gamma_{k0}^2}{\sigma_{\text{tot}}^2}$	$\frac{Nd^2}{4}$
	Between	(5)	$\frac{N\gamma_{0k}^2}{\sigma^2+\overline{n}\left(\text{CV}_n^2+1\right)\tau^2}$	$\frac{Nd^2}{4\left(1-\rho_{Y\vert\mathbf{X}}\right)+4\overline{n}\left(\text{CV}_n^2+1\right)\rho_{Y\vert\mathbf{X}}}$
Yes	Within	(6)	$\frac{N\gamma_{k0}^2}{\left(1-R_{\text{pre,post}}^2\right)\sigma_{\text{tot}}^2}$	$\frac{Nd^2}{4\left(1-R_{\text{pre,post}}^2\right)}$
		(8)	$\frac{N\gamma_{1k0}^2}{2\sigma^2}$	$\frac{Nd^2}{8\frac{\sigma^2}{\sigma_{\text{tot}}^2}}$
	Between	(9)	$\frac{N\gamma_{10k}^2}{2\sigma^2+\overline{n}\left(\text{CV}_n^2+1\right)\tau_{\beta 1}^2}$	$\frac{Nd^2}{8\frac{\sigma^2}{\sigma_{tot}^2}+4\overline{n}\left(\text{CV}_n^2+1\right)\frac{\tau_{\beta 1}^2}{\sigma_{tot}^2}}$

Notes. CV$_n$ is the coefficient of variation of cluster sizes. For models (4), (5), (8) and (9) compare expressions (4.8), (7.9), (4.12) and (11.15) in Spybrook et al. (2009). For model (5), $\rho_{Y\vert\text{X}}=\tau^2/(\tau^2+\sigma^2)=\tau^2/(\sigma_{tot}^2)$. Power prediction for Model (7) is not straightforward.

Table 3

Design conditions in Monte Carlo Simulation

Assignment strategy	Between-clusters or within-clusters
Experiment design	Fractional factorial or complete factorial
Number of clusters J	5 or 10 (for within-clusters)
	25, 30, 40, or 50 (for between-clusters)
Mean cluster size $\bar{n}$	50 or 100 (for within-clusters)
	20 or 100 (for between-clusters)
Unadjusted posttest ICC	Low (.05 unadjusted posttest; .025 change-score),
	Medium (.15 unadjusted posttest; .075 change-score), or
	High (.30 unadjusted posttest; .075 change-score)

Table 4

2^5 Factorial Design Used in the Simulation

Condition	F1	F2	F3	F4	F5	Condition	F1	F2	F3	F4	F5
1	-1	-1	-1	-1	-1	**17**	**+1**	**-1**	**-1**	**-1**	**-1**
2	**-1**	**-1**	**-1**	**-1**	**+1**	18	+1	-1	-1	-1	+1
3	**-1**	**-1**	**-1**	**+1**	**-1**	19	+1	-1	-1	+1	-1
4	-1	-1	-1	+1	+1	**20**	**+1**	**-1**	**-1**	**+1**	**+1**
5	**-1**	**-1**	**+1**	**-1**	**-1**	21	+1	-1	+1	-1	-1
6	-1	-1	+1	-1	+1	**22**	**+1**	**-1**	**+1**	**-1**	**+1**
7	-1	-1	+1	+1	-1	**23**	**+1**	**-1**	**+1**	**+1**	**-1**
8	**-1**	**-1**	**+1**	**+1**	**+1**	24	+1	-1	+1	+1	+1
9	**-1**	**+1**	**-1**	**-1**	**-1**	25	+1	+1	-1	-1	-1
10	-1	+1	-1	-1	+1	**26**	**+1**	**+1**	**-1**	**-1**	**+1**
11	-1	+1	-1	+1	-1	**27**	**+1**	**+1**	**-1**	**+1**	**-1**
12	**-1**	**+1**	**-1**	**+1**	**+1**	28	+1	+1	-1	+1	+1
13	-1	+1	+1	-1	-1	**29**	**+1**	**+1**	**+1**	**-1**	**-1**
14	**-1**	**+1**	**+1**	**-1**	**+1**	30	+1	+1	+1	-1	+1
15	**-1**	**+1**	**+1**	**+1**	**-1**	31	+1	+1	+1	+1	-1
16	-1	+1	+1	+1	+1	**32**	**+1**	**+1**	**+1**	**+1**	**+1**

Notes. Bolded conditions were included in the 2^{5-1} fractional (half) factorial.

F1 = Factor 1, etc.

Table 5

Observed and Predicted Power and Type I Error for Main Effects in Simulated Within-Clusters Experiments (∝=.05)

		Complete Factorial			Fractional Factorial		
		Power		T1E	Power		T1E
#Clusters	#Members	Obs.	Pred.	Obs.	Obs.	Pred.	Obs.
Low ICC							
5	50	0.647	0.605	0.049	0.645	0.605	0.057
	100	0.914	0.884	0.049	0.914	0.884	0.058
10	50	0.921	0.884	0.054	0.919	0.884	0.064
	100	0.998	0.994	0.053	0.997	0.994	0.070
Medium ICC							
5	50	0.655	0.605	0.050	0.653	0.605	0.058
	100	0.917	0.884	0.052	0.916	0.884	0.064
10	50	0.921	0.884	0.050	0.921	0.884	0.062
	100	0.998	0.994	0.049	0.998	0.994	0.066
High ICC							
5	50	0.652	0.605	0.049	0.646	0.605	0.053
	100	0.923	0.884	0.051	0.918	0.884	0.059
10	50	0.924	0.884	0.055	0.921	0.884	0.063
	100	0.997	0.994	0.050	0.998	0.994	0.077

Note. ICC = Intraclass correlation, T1E = Type I Error. Power predictions are based on the formula for Model (8) in Table 2.

Table 6

Observed and Predicted Power and Type I Error for Interactions in Simulated Within-Clusters Experiments (∝=.05)

		Complete Factorial			**Fractional Factorial**		
		Power		**T1E**	**Power**		**T1E**
#Clusters	**#Members**	**Obs.**	**Pred.**	**Obs.**	**Obs.**	**Pred.**	**Obs.**
Low ICC							
5	50	0.215	0.200	0.052	0.221	0.200	0.057
	100	0.395	0.351	0.051	0.399	0.351	0.060
10	50	0.403	0.351	0.052	0.399	0.351	0.061
	100	0.692	0.608	0.050	0.685	0.608	0.071
Medium ICC							
5	50	0.220	0.200	0.052	0.221	0.200	0.054
	100	0.402	0.351	0.051	0.397	0.351	0.061
10	50	0.399	0.351	0.049	0.382	0.351	0.061
	100	0.677	0.608	0.049	0.675	0.608	0.073
High ICC							
5	50	0.216	0.200	0.050	0.216	0.200	0.054
	100	0.393	0.351	0.048	0.409	0.351	0.062
10	50	0.393	0.351	0.051	0.387	0.351	0.060
	100	0.679	0.608	0.050	0.675	0.608	0.072

Note. ICC = Intraclass correlation, T1E = Type I Error. Power predictions are based on the formula for Model (8) in Table 2.

Table 7

Observed and Predicted Power and Type I Error for Main Effects in Simulated Between-Clusters Experiments (∝=.05)

		Complete Factorial			Fractional Factorial		
		Power		T1E	Power		T1E
#Clusters	#Members	Obs.	Pred.	Obs.	Obs.	Pred.	Obs.
Low ICC							
25	20				0.594	0.618	0.038
	100				0.885	0.897	0.056
30	20				0.744	0.733	0.048
	100				0.964	0.959	0.062
40	20	0.875	0.867	0.046	0.887	0.867	0.057
	100	0.993	0.993	0.044	0.993	0.993	0.065
50	20	0.943	0.936	0.047	0.959	0.936	0.059
	100	0.999	0.999	0.046	0.999	0.999	0.074
Medium ICC							
25	20				0.379	0.398	0.053
	100				0.493	0.523	0.054
30	20				0.504	0.493	0.054
	100				0.639	0.635	0.055
40	20	0.635	0.638	0.049	0.645	0.638	0.053
	100	0.777	0.783	0.049	0.787	0.783	0.056
50	20	0.744	0.744	0.049	0.767	0.744	0.061
	100	0.871	0.874	0.047	0.889	0.874	0.058
High ICC							
25	20				0.236	0.252	0.051
	100				0.262	0.292	0.053
30	20				0.310	0.312	0.050
	100				0.355	0.363	0.050
40	20	0.402	0.416	0.051	0.414	0.416	0.053
	100	0.464	0.481	0.050	0.469	0.481	0.051
50	20	0.492	0.507	0.048	0.515	0.507	0.055

		Complete Factorial			Fractional Factorial		
		Power		T1E	Power		T1E
#Clusters	**#Members**	**Obs.**	**Pred.**	**Obs.**	**Obs.**	**Pred.**	**Obs.**
	100	0.559	0.581	0.046	0.594	0.581	0.056

Note. ICC = Intraclass correlation, T1E = Type I Error. Power predictions are based on the formula for Model (9) in Table 2.

Table 8

Observed and Predicted Power and Type I Error for Interactions in Simulated Between-Clusters Experiments (∝=.05)

		Complete Factorial			Fractional Factorial		
		Power		T1E	Power		T1E
#Clusters	#Members	Obs.	Pred.	Obs.	Obs.	Pred.	Obs.
Low ICC							
25	20				0.181	0.206	0.040
	100				0.348	0.369	0.057
30	20				0.250	0.253	0.050
	100				0.472	0.458	0.062
40	20	0.343	0.337	0.043	0.359	0.337	0.059
	100	0.575	0.597	0.043	0.618	0.597	0.069
50	20	0.424	0.413	0.046	0.448	0.413	0.063
	100	0.693	0.704	0.043	0.751	0.704	0.078
Medium ICC							
25	20				0.129	0.137	0.050
	100				0.161	0.173	0.054
30	20				0.163	0.163	0.053
	100				0.207	0.211	0.055
40	20	0.213	0.212	0.048	0.222	0.212	0.053
	100	0.269	0.279	0.050	0.282	0.279	0.057
50	20	0.261	0.258	0.048	0.276	0.258	0.054
	100	0.343	0.342	0.047	0.359	0.342	0.058
High ICC							
25	20				0.089	0.099	0.048
	100				0.106	0.109	0.050
30	20				0.116	0.114	0.053
	100				0.128	0.127	0.053
40	20	0.132	0.141	0.050	0.144	0.141	0.054
	100	0.143	0.160	0.050	0.149	0.160	0.052
50	20	0.164	0.167	0.050	0.177	0.167	0.053

		Complete Factorial			Fractional Factorial		
		Power		T1E	Power		T1E
#Clusters	#Members	Obs.	Pred.	Obs.	Obs.	Pred.	Obs.
	100	0.186	0.191	0.048	0.192	0.191	0.053

Note. ICC = Intraclass correlation, T1E = Type I Error. Power predictions are based on the formula for Model (9) in Table 2.