Article

Journal of Educational and Behavioral Statistics
2022, Vol. 47, No. 6, pp. 639–665
DOI: 10.3102/10769986221104800
Article reuse guidelines: sagepub.com/journals-permissions

Pooling Interactions Into Error Terms in Multisite Experiments

Wendy Chan
University of Pennsylvania

Larry Vernon Hedges
Northwestern University

Multisite field experiments using the (generalized) randomized block design that assign treatments to individuals within sites are common in education and the social sciences. Under this design, there are two possible estimands of interest and they differ based on whether sites or blocks have fixed or random effects. When the average treatment effect is assumed to be identical across sites, it is common to omit site by treatment interactions and "pool" them into the error term in classical experimental design. However, prior work has not addressed the consequences of pooling when site by treatment interactions are not zero. This study assesses the impact of pooling on inference in the presence of nonzero site by treatment interactions. We derive the small sample distributions of the test statistics for treatment effects under pooling and illustrate the impacts on rejection rates when interactions are not zero. We use the results to offer recommendations to researchers conducting studies based on the multisite design.

Keywords: *multisite experiments; randomized block designs; pooling; ANOVA; interactions*

Introduction

Large-scale evaluation studies often involve many different sites (such as schools, clinics, or locations). In education, most of these studies are carried out using variants of two designs, the hierarchical or cluster randomized design and the multisite or (generalized) randomized block design (Spybrook and Raudenbush, 2009). In the cluster randomized design, randomization takes place at the cluster level, so that all the individuals within the same site receive the same treatment. In the language of classical experimental design, clusters are nested within treatments. With multisite designs, treatments are randomly assigned to individuals within sites, so that all sites include some individuals who receive each treatment. In the language of classical experimental design, treatments are crossed with sites. The multisite design is widely used in education and other fields because it can often obtain higher statistical power for lower cost than the hierarchical (cluster randomized) design (Hedges and Rhoads, 2009). The multisite design was found to be the most frequently used design in one survey of randomized trials supported by the U.S. Institute of Education Sciences (IES; Spybrook and Raudenbush, 2009). In recent work, Weiss et al. (2017) cite 16 large multisite designs spanning a range of levels of education. This article focuses on the analysis of the multisite design.

A consequential choice for both analysis and interpretation (which also has implications for details of the design, such as sample sizes) is the inference model or choice of estimand (i.e., the treatment effect to be estimated). One possible estimand is the average treatment effect among the sites observed in the experiment. In classical experimental design, this situation is described as saying that the sites have fixed effects. An alternative estimand is the average treatment effect in some (super)population of sites, a sample of which are observed in the experiment. In classical experimental design, this situation is described as saying that the sites have random effects. In general, the treatment effect estimate has a different distribution when the sites have fixed effects than when the sites have random effects and different test statistics are appropriate for testing the significance of (or creating confidence intervals for) treatment effects.

Regardless of which estimand is of interest, the most general analytic model would include parameters representing the differences of average treatment effects across sites (site–treatment interactions). If the average treatment effect (parameters) is identical across sites, then a more sensitive test for treatment effects can be obtained by omitting site–treatment interactions from the analytic model. This is called “pooling site–treatment interactions into the error term” in classical experimental design (see, e.g., Hines, 1996). Pooling has not been universally embraced in the statistical literature on experimental design. Scheffé (1959) advises against it, while Kendall and Stuart (1968) argue that the issues are complex and need further research, and Hines (1996) argues that power improvements obtained from pooling are generally small. Much of the literature on pooling has addressed conditional pooling strategies based on preliminary tests for interactions (see, e.g., Bancroft and Han, 1983; Bozivich et al., 1956; Han and Bancroft, 1968; Mead et al., 1973; Paull, 1950; Wolde-Tsadi and Afifi, 1980). Yet none of this work seems to address the consequences for inference of pooling when interactions are not exactly zero.

The purpose of this article is to assess the impact of pooling (i.e., omitting site–treatment interactions from the analytic model) when treatment by site interactions are not exactly zero—that is, when average treatment effect parameters are not identical across sites. First, we briefly review relevant literature on classical experimental design, from which recent work on social experiments developed. Then, we derive approximate small sample distributions of the test statistics for treatment effects for both fixed and random effects inference models and show by simulation that the analytic results are quite accurate. Throughout, we focus on the balanced multisite design, in which each site contains the same number of individuals. Focusing on the balanced design removes the ambiguity of whether the site-average treatment effect or the person-average treatment effect is the estimand (because the estimates are identical in this case). Then, we use the small sample distributions to study the impact of pooling nonnull interactions on tests for treatment effects. We assess the impact of pooling by computing the rejection rates of the test statistics based on pooling. We then use the analytic results to derive post hoc adjustments to test statistics computed using pooling. Finally, we discuss other considerations that affect the interpretation of the results presented in this study and offer recommendations for researchers conducting studies based on the multisite design.

Classical Experimental Design

Although large-scale funding for randomized field trials in education became more frequent with the creation of IES in 2002, statistical work on experimental design began 80 years before that time (see, e.g., Fisher and MacKenzie, 1922). Many of the issues that arise in contemporary discussion of educational experiments have their historical antecedents in the classical experimental design literature. The cluster randomized, randomized block designs, and many others were well known and widely used by the 1930s. More subtle issues of the target of estimation (the estimand) were also understood by the first half of the 20th century. Consider the question of whether the purpose of the experiment was to make inferences about the mean treatment effect in the observed blocks (sometimes called a finite sample or fixed effects estimand) or in a population of blocks from which the observed blocks are a sample (sometimes called a superpopulation or random effects estimand). This distinction was explicitly made by Eisenhart (1947) who described the estimands and may have been the first to use the names fixed, random, and mixed effect models. He also described the analysis of variance targeting each of these estimands and gave the sampling distribution of the *F*-test statistics for each effect. However, the distinctions and the appropriate analyses were clearly understood earlier. Daniels (1939) had previously described the analysis of fixed, random, and mixed model analyses (although not using these names).

The earliest work on computing statistical power in the analysis of variance appears to be that of Tang (1938), but there was much subsequent work on computing the relevant noncentral sampling distributions for power calculations. Scheffé (1959) provided a sophisticated explication of the analysis of variance that unified and extended much of the previous literature on experimental design and analysis. Early work focused on balanced designs where the same number of individuals are randomized to each treatment in each site, but the analysis of unbalanced designs was addressed by Windsor and Clark (1940) and considerably elaborated by Henderson (1953). The classical experimental design literature seems to have focused on the estimation and testing of mean and variance of mean block-average treatment effects, rather than person average or precision weighted treatment effects, although this distinction and its implications for estimation and precision were well known in the survey literature (see, e.g., Kish, 1965).

The emergence of interest in large-scale randomized trials in education after the formation of IES in 2002 led to a resurgence of interest in the design and analysis of randomized experiments in education. Because the power function of tests for treatment effects in both fixed and random effects models for balanced designs was available and because these designs were most efficient, much of the early work on power analysis focused on balanced designs. Raudenbush (1997) used classical results on balanced nested designs to study power and optimal design in cluster randomized trials. Raudenbush and Liu (2000) used classical results on balanced mixed model designs to study power and optimal design in multisite trials. Schochet (2008) applied the same theory to discuss power analyses of randomized trials. More recently, there has been renewed interest in estimating variance components characterizing cross-site variation in average treatment effects. For example, Raudenbush and Bloom (2015) used classical results to study the power of tests for treatment effect variance components.

The availability of software for multilevel analysis (the mixed general linear model) has enabled broader use of analyses of unbalanced designs, for which analytic results are difficult to obtain. For example, Bloom et al. (2017) studied alternative models for studying cross-site variation in treatment effects in unbalanced designs. Miratrix et al. (2021) used simulation to study the properties of various analyses in unbalanced designs.

While some authors appreciated and exploited classical results from experimental design, classical results have not always been fully exploited to understand the consequences of choices about design and analysis. The main focus of classical experimental design has been on balanced designs, which often have simpler analytic properties than unbalanced designs. Thus, they can provide analytic results that give insight into what might happen in unbalanced designs. For example, the sampling properties of finite and superpopulation estimators of average treatment effects were studied using simulation by Miratrix et al. (2021). Two of their findings were that (1) in designs with treatment effect heterogeneity, analyses targeting finite sample treatment effect estimands generally have smaller variance than those targeting superpopulation estimands (which puts individuals into an infinite population framework) and (2) the *variances* of the standard errors of treatment effect estimates are generally smaller for finite population estimands. Both of these results can be obtained analytically for balanced designs and they may be useful to help contextualize simulation results.

The analytic results would show, for example, how results vary as a function of parameters like sample size allocations and variance components.

The Analysis of Variance of the Multisite (Randomized Block) Design

In this article, we consider the analysis of the balanced multisite design, the design in which each site (block) contains the same number of individuals. We center our discussion on the balanced design because its optimal analysis (the standard analysis of variance) is well-understood. Note that while this section uses the analysis of variance framework, this approach and regression, or the mixed linear model approach, are notational variants (i.e., they give identical analyses).

Model and Notation

Suppose that the study design has m sites (blocks), and the n individuals in each site are randomly assigned to each of the treatment and the control conditions (a balanced design). Throughout, we use the terms sites and blocks interchangeably. The total number of individuals in the study is given by $N = 2mn$. Let Y_{ijk} denote the outcome measurement of the kth individual in the jth site and the ith treatment. Using the conventional notation of the analysis of variance, the model for Y_{ijk} in the multisite design is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}; i = 1, 2; j = 1, \ \ldots, \ m; k = 1, \ \ldots, \ n, \qquad (1)$$

where μ is the grand mean, α_i is the effect of the ith treatment, β_j is the effect of the jth site (block), γ_{ij} is the interaction effect of the ith treatment and the jth site, and ε_{ijk} is a residual. The residual ε_{ijk} is assumed to be independent of other residuals and to have a normal distribution with variance $\sigma_\varepsilon{}^2$. The grand mean μ and the treatment effects α_i are taken to be fixed, but unknown constants.

Let $\bar{Y}_{1j\cdot}$ $(j = 1, \ldots, m)$ and $\bar{Y}_{2j\cdot}$ $(j = 1, \ldots, m)$ be the means of the jth site in the treatment and control groups, respectively, let $\bar{Y}_{1\cdot\cdot}$ and $\bar{Y}_{2\cdot\cdot}$ be the overall (group) means in the treatment and control groups, respectively, and let $\bar{Y}_{\cdot\cdot\cdot}$ be the grand mean. The analysis of variance for this design is typically described in terms of *MSA, MSB, MSG*, and *MSW*, the mean squares for treatment, sites, the site–treatment interaction, and within cells, respectively, which are defined by

$$MSA = SSA = \frac{mn(\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot})^2}{2},$$

$$MSB = \frac{SSB}{m-1} = 2n \sum_{j=1}^{m} \frac{[(\bar{Y}_{1j\cdot} + \bar{Y}_{2j\cdot}) - (\bar{Y}_{1\cdot\cdot} + \bar{Y}_{2\cdot\cdot})]^2}{(m-1)},$$

$$MSG = \frac{SSG}{m-1} = n\sum_{j=1}^{m}\frac{[(\bar{Y}_{1j\cdot} - \bar{Y}_{2j\cdot}) - (\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot})]^2}{(m-1)},$$

and

$$MSW = \frac{SSW}{2mn - 2m} = \sum_{i=1}^{2}\sum_{j=1}^{m}\sum_{k=1}^{n}\frac{(Y_{ijk} - \bar{Y}_{\cdots})^2}{2m(n-1)}.$$

Here, the mean square for each source of variability is defined by the sum of squares for the source divided by the degrees of freedom for that source. A fundamental identity in the analysis of variance for balanced designs is that the sums of squares *SSA, SSB, SSG*, and *SSW* are statistically independent.

Fixed Versus Random Block Effects

There are different possible models for the parameters β_j (block or site effects) and γ_{ij} (interaction effects). In the *blocks fixed or fixed effects model*, β_j and γ_{ij} are assumed to be fixed, but unknown constants. *In the blocks random or mixed effects model*, β_j and γ_{ij} are assumed to be random and we define the variance of γ_{ij} to be $\sigma_\gamma^2/2$. We define the interaction variance in this manner to simplify the formulas for the distribution of the sums of squares. In mixed effects models, the block and interaction effects are also often assumed to be normally distributed, and we adopt that model here. Although the computations for the mean squares in both models (fixed and mixed effects) are identical, the choice of model has important implications for the distribution of the sums of squares, for the test statistics used to test the treatment effect, for the precision (variance) of the treatment effect estimate, and for the inferences that can be drawn.

Blocks Random Analysis

Using the model in Equation 1, the treatment effect is defined as $\alpha_1 - \alpha_2$ and an estimate of the treatment effect is given by

$$\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot}$$

When the blocks are random, the variance of the treatment effect estimate is

$$\frac{2(\sigma_\varepsilon{}^2 + n\sigma_\gamma{}^2)}{mn}, \tag{2}$$

and the standard error of the treatment effect is the square root of this variance. To test the null hypothesis of no treatment effects ($\alpha_1 = \alpha_2$), the test statistic is given by

$$F = \frac{MSA}{MSG} = \frac{(m-1)SSA}{SSG}. \tag{3}$$

When the null hypothesis is true, this test statistic has an F-distribution with 1 degree of freedom in the numerator and $m-1$ degrees of freedom in the denominator (see, e.g., Kirk, 1995). In this design, this test is the uniformly most powerful unbiased test for the treatment effect. Note that *MSG,* the mean square for the interaction, is sometimes called the "error term" in this analysis because it is the denominator of the F-statistic.

Note that the F-test given in Equation 3 has approximately the same sampling distribution as the square of the treatment effect estimate divided by the square of its standard error in a two-level hierarchical model (individuals nested within blocks) introducing treatment as a Level 1 (individual level) predictor, with treatment and block effects random and normally distributed. The treatment effect estimates are identical, but the standard errors will be slightly different because here, maximum likelihood or restricted maximum likelihood will be used to estimate the variance components in the error term. The advantage of this analytic strategy (based on regression) is that it can easily be used with unbalanced designs.

When blocks are random, the inference model refers to the average treatment effect in the population of blocks from which the observed blocks are a (random) sample. In other words, in this case, the estimand of interest is the "super population block" average treatment effect (Miratrix et al., 2021). In this design, inferences about the treatment effect are about the mean in the (super)population of blocks from which those observed are a sample. An advantage of the blocks random design is that the statistical model incorporates some degree of uncertainty associated with generalizations of the average treatment effect to a larger population. However, while some uncertainty due to the sampling of blocks is taken into account, the invocation of the sampling theory of generalization is not entirely warranted if the blocks are not a true random sample of a well-defined population.

Blocks Fixed Analysis

If blocks (and therefore block–treatment interactions) are fixed, the only random term is the residual ε_{ijk}. The test statistic for testing the null hypothesis of no treatment effect ($\alpha_1 = \alpha_2$) is

$$F = \frac{MSA}{MSW} = \frac{2m(n-1)SSA}{SSW}, \tag{4}$$

which has the F-distribution with 1 degree of freedom in the numerator and $2mn - 2m$ degrees of freedom in the denominator when the null hypothesis is true. In this design with fixed blocks, this test is the uniformly most powerful unbiased test for the treatment effect. Note that unlike the blocks random case, *MSW* is considered the "error term" in this analysis because it is the denominator of the F-statistic. The F-statistic given in Equation 4 is exactly the same as would be obtained by fitting a regression model with dummy variables for treatment, blocks, and the block–treatment interactions. The advantage of the regression analysis is that it is easy to use when the design is unbalanced.

Like the blocks random case, the treatment effect is defined here as $\alpha_1 - \alpha_2$, and an estimate of the treatment effect is

$$\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot}$$

When the blocks are fixed, the variance of the treatment effect estimate is

$$\frac{2{\sigma_\varepsilon}^2}{mn}, \tag{5}$$

and the standard error of the treatment effect is the square root of the variance.

The inference model for treatment effects when blocks are fixed refers to the average treatment effect in the set of blocks that happen to be included in the experiment. Similarly, the precision of the treatment effect estimate is based on the estimate of the average treatment effect in the specific set of blocks in the study. In general, when blocks have fixed effects, the statistical inference is only about the treatment effect in the blocks actually observed in the experiment.

Unlike the blocks random case, in the blocks fixed case, there is no statistical sampling theory warrant for generalizing about treatment effects to blocks that are not included in the experiment. The uncertainty considered in the statistical inference for the blocks fixed case is only the uncertainty associated with obtaining samples of individuals within the blocks that are included in the experiment. When scientists make generalizations beyond the blocks in the experiment, the argument is extrastatistical (e.g., "I can generalize to additional blocks because they are *sufficiently similar* to the blocks in the experiment, so that treatment effects are likely to be the same"). However, if treatment effects vary across blocks, generalizations will be unwarranted if the sample of blocks included in the experiment is not a representative sample of all blocks.

Pooling Sums of Squares

The analysis of variance above (or the corresponding regression analysis) permits testing for average treatment effects in the presence of block or block–treatment interaction effects. However, if there is no block–treatment interactions, then a more powerful test and a smaller standard error of the treatment effect estimate can be obtained by omitting the interaction effect from the analytic model. This is called "pooling the interaction sums of squares into the error term" in classical experimental design because (in either the blocks fixed or blocks random model) the error term of the test statistic becomes

$$MS_{\text{Pooled}} = SS_{\text{Pooled}}/(m - 1 + 2mn - 2m) = (SSG + SSW)/(2mn - m - 1), \tag{6}$$

the test statistic becomes

$$FP = MSA/MS_{\text{Pooled}}, \tag{7}$$

and the denominator degrees of freedom of the test statistic become

$$[m - 1 + 2m(n - 1)] = 2nm - m - 1 = N - m - 1. \tag{8}$$

Let FP denote the F-statistic based on the pooled sums of squares. Note that FP is the same test statistic for the treatment effect that we would obtain from fitting a regression model for treatment and blocks (but omitting block–treatment interactions), so that

$$FP = (b_{\text{Treatment}}/SE_{\text{Treatment}})^2,$$

where $b_{\text{Treatment}}$ is the regression coefficient estimate for the treatment effect and $SE_{\text{Treatment}}$ is its standard error.

The improvements in power and precision due to pooling arise because the expectation of SE^2 is the same with and without pooling, but the degrees of freedom in the pooled error term (the denominator of the F-statistic) increase. These improvements depend, of course, on the hypothesis that the interaction effect parameters are exactly zero. If this is not true, then the null distribution of the test statistic based on pooled error terms is not the nominal F-distribution.

A Note on Unbalanced Designs

The treatment effect estimate in the balanced design that we address in this article could have been written as

$$\bar{Y}_{1.} - \bar{Y}_{2.} = \frac{1}{m}\sum_{j=1}^{m}(\bar{Y}_{1j.} - \bar{Y}_{2j.}).$$

Thus, it is the unweighted average of the block-specific treatment effect estimates. Because each block-specific treatment effect estimate $\bar{Y}_{1j.} - \bar{Y}_{2j.}$ has the same variance in the balanced design, the overall average treatment effect could be described as either the unweighted average or (equivalently) as the precision weighted average.

If the design was unbalanced and the block sizes (or cell sizes) were unequal, then the variances of the $\bar{Y}_{1j.} - \bar{Y}_{2j.}$ would not all be the same. Thus, the unweighted and precision weighted averages of the $\bar{Y}_{1j.} - \bar{Y}_{2j.}$ would be different. For example, if blocks have fixed effects, the treatment effect estimate computed in a regression analysis (fitting only a treatment dummy and block dummies) would be the precision weighted mean. However, the precision weighted mean need not be the estimand of interest, and if it is not, then the analysis needs to be adjusted (e.g., by choosing an appropriate weighting scheme). This distinction in survey sampling is known as that between the element mean versus the group mean, where its implications for estimation and variance are well known (see, e.g., Kish, 1965).

The Small Sample Properties of Analyses That Omit Site by Treatment Interactions

In this section, we provide the small sample distribution of the test for treatment effects in the balanced multisite design when nonnull interactions are present, but they are not included in the analytic model. We use it to compute the actual significance level of tests for treatment effects that omit the interactions from the analytic model. We also give the impact on the precision of (and confidence intervals for) treatment effect estimates. Details of the derivations are given in the Appendix.

Blocks Random

If blocks have random effects, then the sampling distribution of the regression test statistic *FP* is a constant times an *F*-distribution with reduced degrees of freedom. That is,

$$FPC_R = FP\left(\frac{(N-m-1)\sigma_\varepsilon{}^2 + n(m-1)\sigma_\gamma{}^2}{(N-m-1)(\sigma_\varepsilon{}^2 + n\sigma_\gamma{}^2)}\right) \tag{9}$$

has an *F*-distribution with 1 degree of freedom in the numerator and

$$h = \frac{[(N-m-1)\sigma_\varepsilon{}^2 + n(m-1)\sigma_\gamma{}^2]^2}{(N-m-1)\sigma_\varepsilon{}^4 + n^2(m-1)\sigma_\gamma{}^4 + 2n(m-1)\sigma_\varepsilon{}^2\sigma_\gamma{}^2} \tag{10}$$

degrees of freedom in the denominator. The term FPC_R refers to the corrected pooled *F* statistic in the blocks random case. Note that if $\sigma_\gamma^2 = 0$, *FP* has an *F*-distribution with 1 and $(N-m-1)$ degrees of freedom, so that the test based on pooling has exactly the expected distribution. However, if $\sigma_\gamma^2 > 0$, then the corrected *F*-statistic $FPC_R < FP$ and $h < N-m-1$. Therefore, the critical value for FPC_R is greater than the nominal critical value (based on $N-m-1$ degrees of freedom in the denominator). This means that using *FP* with its nominal critical value will lead to a rejection rate that is higher than nominal.

Equation 9 (along with Equation 10) can be used to compute the actual rejection rate of the test conducted using *FP*. Because $FPC_R = cFP$ (the constant *c* is the second factor on the right-hand side of Equation 9), it follows that

$$P\{FP > F^{\alpha}_{N-m-1}\} = P\{cFP > cF^{\alpha}_{N-m-1}\} = P\{FPC_R > cF^{\alpha}_{N-m-1}\}. \tag{11}$$

Because the distribution of FPC_R is given by Equation 9, using F^{α}_{N-m-1} as the critical value of the *F*-distribution with 1 and $N-m-1$ degrees of freedom in Equation 11 gives the actual rejection rate of the nominal level α test.

Note that Equation 9 implies that if we knew (or were willing to impute) the values of σ_ε^2, σ_γ^2, or just $\sigma_\gamma^2/\sigma_\varepsilon^2$, we could adjust or "correct" the results of a significance test performed that omitted the block by treatment interactions from the analytic model. While the value of $\sigma_\gamma^2/\sigma_\varepsilon^2$ is unlikely to be known, the imputation of a range of plausible values might be useful as a robustness check. This is similar to the procedure used by the What Works Clearinghouse in imputing values of the intraclass correlation to adjust significance tests for clustering when the original analysis omitted the clustering effects (Hedges, 2007a).

To assess the impact of using the test statistic based on the pooled error term in the blocks random case when $\sigma_\gamma^2 > 0$, we use the distribution of FPC_R to evaluate the actual significance level (Type I error or rejection rate) of tests for the treatment effect based on *FP*. Figure 1 gives the actual significance level for nominal 0.05 level tests using *FP* for various values of m, n, and $\sigma_\gamma^2/\sigma_\varepsilon^2$. The figure shows the rejection rates obtained by the simulation of the test based on the corrected test statistic FPC_R using Equation 9 (sim corrected) and the uncorrected test statistic *FP* (sim uncorrected). It also gives the theoretical rejection rate for the uncorrected test statistic based on Equation 11 (theo uncorrected). Each column of the figure has a different value of the ratio $\sigma_\gamma^2/\sigma_\varepsilon^2$ (where $\sigma_\varepsilon^2 = 1$) and each row has a different block size *n*. Figure 1 shows that, when blocks are random, the test with no block–treatment interactions fitted gives quite liberal results, even when the interaction variance is small compared to the within cell variance (i.e., even when σ_γ^2 is small compared to σ_ε^2). For example, when $\sigma_\gamma^2 = 0.05\sigma_\varepsilon^2$, the actual significance level of a nominal 0.05 level test is about 0.10, when the number of sites $m = 10$ and $n = 10$. As $\sigma_\gamma^2/\sigma_\varepsilon^2$ or the number of sites *m* becomes larger, the test becomes more liberal. When $m = 50$, $n = 20$, and $\sigma_\gamma^2/\sigma_\varepsilon^2 = 0.50$, the rejection rate is greater than 0.50. Note that the rejection rates for the corrected test statistic FPC_R obtained by simulation are essentially identical to those given by the theory and are close to nominal, which implies that Equation 9 provides an accurate correction to the significance test. Furthermore, the rejection rates of the uncorrected test statistics obtained by theory and those obtained by simulation are essentially identical to each other and they are both higher than nominal.

In addition to the rejection rates, Figure i in the Supplemental Appendix provides the degrees of freedom *h* in Equation 10 for the corrected test statistic and the nominal degrees of freedom $N - m - 1$ for the uncorrected test statistic for the blocks random case. The figure illustrates that while the differences in degrees of freedom are small when the number of sites *m* is small, they are much larger when the number of sites increases. The larger differences are also seen when the ratio $\sigma_\gamma^2/\sigma_\varepsilon^2$ increases.

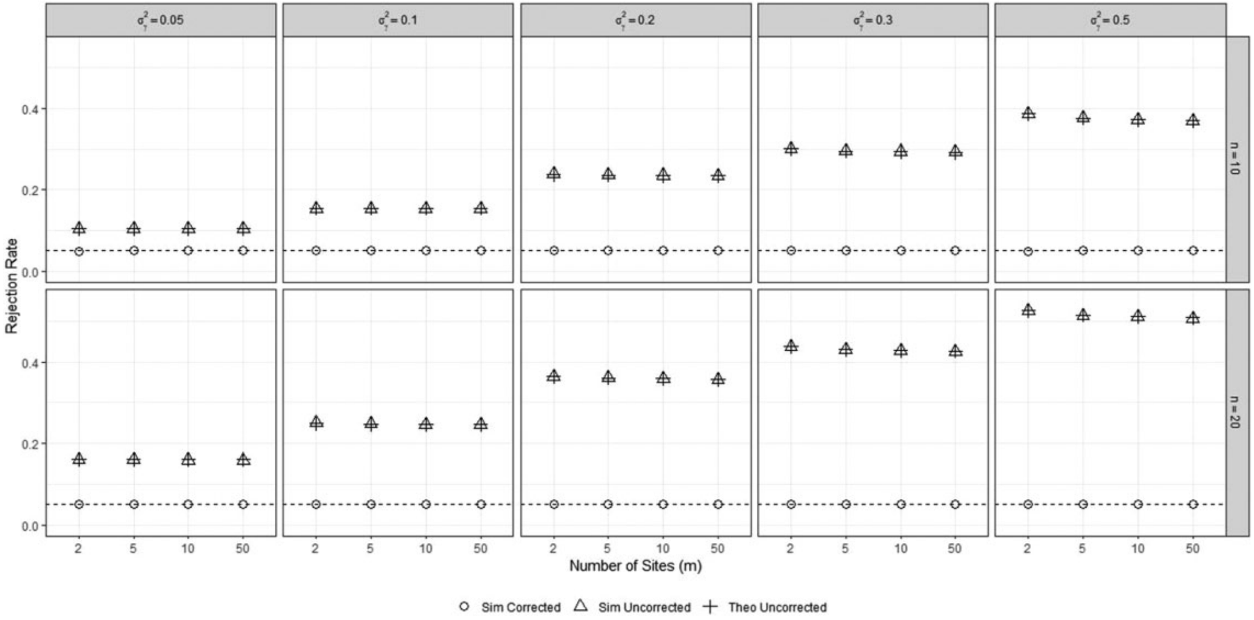

FIGURE 1. *Rejection rates for test statistics pooling interactions into the error term in the blocks random case.*

Note. *Sim corrected is the rejection rate of the corrected test statistic obtained by simulation. Sim uncorrected is the rejection rate of the uncorrected test statistic obtained by simulation. Theo uncorrected is the theoretical rejection rate of the uncorrected test statistic. The within cell variance is set to $\sigma\epsilon^2 = 1$. The simulation results are based on 50,000 replications. The dashed line is the nominal $\alpha = 0.05$ rejection rate.*

Blocks Fixed

If blocks have fixed effects, then the sampling distribution of the test statistic $FP = MSA/MS_{\text{Pooled}}$ is also a constant times an F-distribution with altered degrees of freedom. The distribution of *FP* when blocks are fixed depends on the constants $\gamma_{11},\ \gamma_{21},\ \ldots,\ \gamma_{1m},\ \gamma_{2m}$. To express the distribution succinctly, it is useful to express the sum of squared interaction parameters as the realized variances of the fixed interaction parameters. Using the identification that the mean of the effects is zero, we can define the realized variance of the fixed effects, ${\sigma_G}^2$ (note the Roman letter in the subscript), as

$$ {\sigma_G}^2 = \sum_{j=1}^{m} \frac{[{\gamma_{1j}}^2 + {\gamma_{2j}}^2]}{m-1}. \tag{12} $$

Note that the concept of variance here is a descriptive, rather than a formal, concept because ${\sigma_G}^2$ is not the variance of a random variable in this design. Additionally, we use the symbol ${\sigma_G}^2$ in the blocks fixed case to distinguish this parameter from the variance component ${\sigma_\gamma}^2$ in the blocks random case.

The corrected test statistic FPC_F in the blocks fixed case is given by

$$ FPC_F = FP\frac{(N-m-1){\sigma_\varepsilon}^2 + 2n(m-1){\sigma_G}^2}{(N-m-1){\sigma_\varepsilon}^2} = FP\left(1 + \frac{2n(m-1){\sigma_G}^2}{(N-m-1){\sigma_\varepsilon}^2}\right), \tag{13} $$

which has an F-distribution with 1 degree of freedom in the numerator and

$$ h = \frac{[(N-m-1){\sigma_\varepsilon}^2 + 2n(m-1){\sigma_G}^2]^2}{(N-m-1){\sigma_\varepsilon}^4 + 2n(m-1){\sigma_\varepsilon}^2{\sigma_G}^2} \tag{14} $$

degrees of freedom in the denominator. The term FPC_F refers to the corrected pooled test statistic in the blocks fixed case. Note that if ${\sigma_G}^2 = 0$, $FPC_F = FP$ has an F-distribution with 1 and $(N - m - 1)$ degrees of freedom, so that the nominal distribution of *FP* is exactly correct. However, if ${\sigma_G}^2 > 0$, then the corrected F-statistic $FPC_F > FP$ and $h > N - m - 1$. Since both the value of FPC_F and its associated degrees of freedom *h* are larger than in the nominal analysis, the true rejection rate of the test using *FP* is smaller than nominal. In fact, the rejection rate tends to be only modestly smaller than nominal.

Like the blocks random case, we can adjust the results of the significance test using (or imputing) the values of ${\sigma_\varepsilon}^2$, ${\sigma_G}^2$, or just ${\sigma_G}^2/{\sigma_\varepsilon}^2$. When the test statistic is not adjusted and interaction effects are present, the distribution of FPC_F can be used to evaluate the actual significance level (Type I error or rejection rate) of regression tests for the treatment effect based on *FP* in the blocks fixed case. Figure 2 gives the actual significance level for nominal 0.05 level tests for various values of $m,\ n,$ and ${\sigma_G}^2/{\sigma_\varepsilon}^2$. The figure is organized similar to Figure 1 with the important exception that the columns refer to the realized variance of the fixed effects ${\sigma_G}^2$. Figure 2 shows that when blocks have

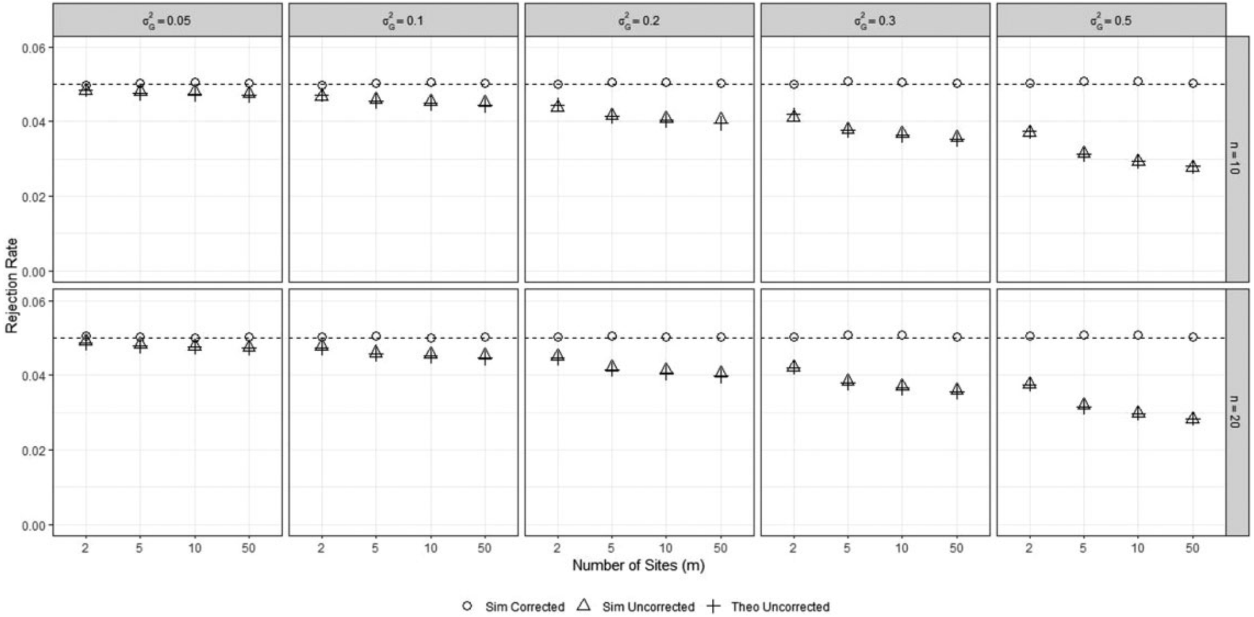

FIGURE 2. *Rejection rates for test statistics pooling interactions into the error term in the blocks fixed case.*
Note. *Sim corrected is the rejection rate of the corrected test statistic obtained by simulation. Sim uncorrected is the rejection rate of the uncorrected test statistic obtained by simulation. Theo uncorrected is the theoretical rejection rate of the uncorrected test statistic. The within cell variance is set to* $\sigma\epsilon^2 = 1$*. The simulation results are based on 50,000 replications. The dashed line is the nominal* $\alpha = 0.05$ *rejection rate.*

fixed effects, the test using a model with no block–treatment interactions fitted gives somewhat conservative results, but the actual significance levels are not far from nominal, particularly when the ratio $\sigma_G{}^2/\sigma_\varepsilon{}^2$ is small. For example, when $\sigma_G{}^2 = 0.05\sigma_\varepsilon{}^2$ and $m = n = 10$, the actual significance level of a nominal 0.05 level test is 0.048. As $\sigma_G{}^2$ becomes larger, the test becomes more conservative. In the worst case when the ratio $\sigma_G{}^2/\sigma_\varepsilon{}^2 = 0.05$ and $m = 50$, the actual significance level is less than 0.03. Figure 2 illustrates that the rejection rate of the corrected test statistic has rejection rates that are close to, but slightly larger than nominal, while the rejection rates of the uncorrected test statistic obtained by theory and by simulation are comparable to each other, but conservative overall. Note that the standard error of the rejection rates obtained by simulation is about 0.001.

Figure ii in the Supplemental Appendix provides a comparison of the degrees of freedom for the blocks fixed case. From the figure, the degrees of freedom *h* in Equation 14 and the nominal degrees of freedom $N - m - 1$ are similar overall, but the differences become more apparent when the number of sites *m* increases and the ratio $\sigma_G{}^2/\sigma_\varepsilon{}^2$ is large. However, these differences in degrees of freedom are less notable compared to those in the blocks random case.

Example

To illustrate the implications of the results, consider a balanced randomized block experiment designed to detect an effect of size $\delta = 0.2$ with 80% power under the assumption that treatment effects were homogeneous across blocks. A sample size of $m = 20$ and $n = 20$ (a total sample size of $N = 800$) would have power of over 80% to detect the treatment effect. Because the treatment effects are expected to be homogeneous, the same analysis (e.g., a regression analysis fitting a treatment effect and dummy variables for blocks, which pools the interaction variance into the error term) would be appropriate regardless of whether the blocks were considered to have fixed or random effects.

Thus far, the discussion has largely focused on hypothetical values of the interaction variance and the effect of omitting interactions on statistical inference when these variances are not zero. Here, we discuss some empirical evidence of the magnitude of these variances and consider the range of treatment effect variance component estimates found in 16 large studies by Weiss et al. (2017). The studies were selected because they had randomized block designs, they were large (and therefore likely to provide stable estimates of treatment effect variance components), and they covered a broad range of educational levels. The number of sites ranged from 9 to over 300 and each study reported several outcomes, so that there was a total of 51 heterogeneity variance component estimates. Ten (20%) of the 51 estimates were zero, but the other 80% of the estimates ranged from $\sigma_\gamma{}^2/\sigma_\varepsilon{}^2 = 0.02$ to 0.35 with a mean of 0.13 and a median of 0.12. We will examine the impact of four heterogeneity values $\sigma_\gamma{}^2/\sigma_\varepsilon{}^2 = 0.00, 0.05, 0.12$, and 0.35 that span the range of Weiss et al.'s data.

Suppose that a superpopulation estimand was of interest, so that the blocks are considered to have random effects. If the model and the assumption of homogeneity of treatment effects is correct (i.e., $\sigma_\gamma{}^2/\sigma_\varepsilon{}^2 = 0.00$), then the nominal rejection rate for the test of treatment effect using the pooled analysis is exactly correct. However, suppose that there is some heterogeneity and $\sigma_\gamma{}^2/\sigma_\varepsilon{}^2 = 0.05$ (low end of the range in Weiss et al. [2017]). Let the number of sites and site sizes $m = n = 20$. Then, Equations 9 and 10 allow us to compute the actual significance level of the $\alpha = .05$ level test. In this case, the actual significance level is about .16 (over 300% of the nominal level of significance). If the heterogeneity was the median of the nonzero values observed by Weiss et al. with $\sigma_\gamma{}^2/\sigma_\varepsilon{}^2 = 0.12$, then the actual significance level of the $\alpha = .05$ level test is about 0.27 (over 500% of the nominal level of significance). If the heterogeneity was at the high end of the range observed by Weiss et al. with $\sigma_\gamma{}^2/\sigma_\varepsilon{}^2 = 0.35$, then the actual significance level of the $\alpha = .05$ level test is about .45 (over 900% of the nominal level of significance—almost equivalent to the flip of a fair coin).

Now suppose instead that the estimand was the average treatment effect in the blocks that were observed (the finite population estimand). In this case, the effects on the significance level of pooling nonnull interactions into the error term are less meaningful. If the assumption of homogeneity of treatment effects is correct (i.e., $\sigma_G{}^2/\sigma_\varepsilon{}^2 = 0.00$), then the nominal rejection rate for the test of treatment effect using the pooled analysis is exactly correct. However, suppose that there is some heterogeneity and $\sigma_G{}^2/\sigma_\varepsilon{}^2 = 0.05$. Then, using Equations 13 and 14 to compute the actual significance level of the $\alpha = .05$ level test, we obtain .047 (only slightly less than the nominal level of significance). If the heterogeneity was a bit bigger and $\sigma_G^2/\sigma_\varepsilon^2 = 0.12$, then the actual significance level of the $\alpha = .05$ level test would be about .044, and if the heterogeneity was at the high end of the range observed by Weiss et al. with $\sigma_G{}^2/\sigma_\varepsilon{}^2 = 0.35$, then the actual significance level of the $\alpha = .05$ level test would be about .034. Thus, for blocks fixed, the test would be only slightly conservative in all cases.

Considerations in Using the Findings of This Article

The results of this article imply that failure to include interactions in the analytic model (pooling the interactions into the error term) can distort significance tests and the computed precision of treatment effects if interactions are present. This is particularly the case in the blocks random framework. However, before giving our recommendations, we call the reader's attention to other things that might be considered in interpreting the results of this study. First, while the analytic results are correct (they were checked by simulations reported in Figures 1 and 2), the levels of heterogeneity chosen for the presentation in this article could be much larger than are plausible, exaggerating the consequences in practice. Alternatively, the results may well apply, but the obvious remedy, inclusion of interactions in the analytic model, might result in so large of a reduction of design sensitivity that it is untenable, leaving the researcher in a quandary about what to do. Yet another possibility is that the use of robust standard errors obviates any problems. We examine each of these possibilities in the sections that follow.

Are the Levels of Heterogeneity Examined Implausibly Large?

To understand the implications of the analytic results in this article, it is necessary to understand what amounts of heterogeneity ($\sigma_\gamma{}^2/\sigma_\varepsilon{}^2$ or $\sigma_G{}^2/\sigma_\varepsilon{}^2$) might plausibly occur. If the effects we demonstrate only occur with implausibly large amounts of heterogeneity, they might have few practical implications. We offer two lines of evidence that the amounts of heterogeneity reported here are likely to occur in practice. One is empirical, and the other is theoretical.

The results from Weiss et al. (2017) provide empirical evidence of the magnitude of interaction variances that were observed in a sample of multisite trials. As mentioned, the range of between-blocks treatment effect variance components spanned from 0.02 to 0.35 with an average of 0.13 and a median of 0.12. This range overlaps with the values chosen in our simulation (given in Figures 1 and 2) and provides validation that the amounts of heterogeneity considered in this study are plausible in practice.

From a theoretical perspective, much of the literature on pooling involves evaluation of the plausibility that interactions are negligible via a preliminary significance test for the interaction. Pooling is only done if the preliminary test is not significant. From this perspective, an interaction that is too small to detect via a significance test is negligible—as an empirical matter, it is undetectable (at least in the particular study at hand).

Thus, we can obtain some perspective on how large the interactions that are pooled might be by considering whether the interaction could be detected by a significance test in a study that was adequately sensitive to detect treatment effects of reasonable size. Assume that a study was planned with the expectation that there were no site (block)–treatment interactions, with a significance level of .05 and power of .8. One might choose an expected effect size of $\delta = 0.2 - 0.3$ as reasonable values (the low end of the range of Cohen's small to medium effect sizes). For example, the median value of effect sizes in the What Works Clearinghouse is at the low end of this range. A study with $m = 20$ and $n = 10$ would have a minimum detectable effect size (MDES) less than 0.3 (as would a study with $m = 10$ and $n = 20$). Similarly, a study with $m = 20$ and $n = 20$ would have an MDES less than 0.2 (as would a study with $m = 40$ and $n = 10$). Thus, a range of $m = 10$–50 and $n = 10-20$ seems reasonable to expect of designs that might be used in education trials.

TABLE 1.
The Minimum Value of $\sigma_\gamma^2/\sigma_\varepsilon^2$ (or $\sigma_G^2/\sigma_\varepsilon^2$) for Which a .05 Level Significance Test for the Interaction Would Have 80% Power

	Blocks Random			Blocks Fixed		
m	*n* = 10	*n* = 15	*n* = 20	*n* = 10	*n* = 15	*n* = 20
10	.23	.15	.11	.19	.12	.09
15	.16	.11	.08	.14	.09	.07
20	.13	.09	.07	.12	.08	.06
25	.11	.07	.06	.10	.07	.05
50	.07	.05	.04	.07	.05	.04

Consider the test statistic for testing the hypothesis that $\sigma_\gamma^2 = 0$ or $\sigma_G^2 = 0$. Note that the test is the same regardless of whether sites are fixed or random, but the power function is different depending on whether sites are random or fixed. The test statistic is defined in terms of the mean squares as

$$F = MSG/MSW$$

and the results in the Appendix imply the power function of the test. While the power function of tests for interactions has been well known for some time (it is implied, e.g., by results in Searle [1971]), it has only recently been studied by applied researchers (e.g., Bloom & Spybrook, 2017; Raudenbush and Bloom, 2015).

Table 1 shows the minimum value of $\sigma_\gamma^2/\sigma_\varepsilon^2$ (and $\sigma_G^2/\sigma_\varepsilon^2$) for which the test of site–treatment interaction achieves 80% power (for significance level .05) as a function of *m* and *n*. One might say that Table 1 gives the MDES variance. The values *m* = 10–25 and *n* = 5–20 correspond to total sample sizes ranging from *N* = 200–2,000. Table 1 shows that the minimum detectable values of $\sigma_\gamma^2/\sigma_\varepsilon^2$ (for sites random) or $\sigma_G^2/\sigma_\varepsilon^2$ (for sites fixed) are 0.05 or larger for all but the largest design (with *m* = 50 and *n* = 20 and total *N* = 2,000). Thus, the smallest values of the interaction variance considered (0.05 and 0.1) would have been undetectable in experiments with a range of sample sizes that are plausible for education experiments.

Does Adding Interaction Effects Drastically Decrease Design Sensitivity?

The major reason that interactions are omitted from analytic models (pooled into the error term) is that including interaction effects reduces the sensitivity of statistical tests and decreases the width of confidence intervals because it reduces degrees of freedom in the error term. Thus, in deciding whether to include interaction effects in the analytic model, the researcher must weigh this reduction in sensitivity if interactions are included against the consequences for significance testing if they are not.

To help understand this trade-off, we examined design sensitivity in two ways. First, we examined the MDES, which is the smallest effect size that will achieve a specified statistical power given a specified significance level. Given the specified power and significance level, the MDES is a function only of m and n. Second, we computed the expected length of the confidence interval for the average treatment effect size, which is a function of only m and n and the confidence coefficient of the interval desired. In both cases, we examined the situation in which there was no interaction in the correct model, so that the analysis without interactions is a correct specification. However, because the correct model has no interaction, the nominal sampling distribution of the test statistic is correct both with and without interactions, albeit with different residual degrees of freedom.

We computed the MDES with and without interactions in the model, as a function of m and n. The specified statistical power was 80% (as suggested, e.g., by Cohen, 1977) at a specified significance level (here 5%). The largest difference was 0.16 (MDES = 2.50 without interactions versus MDES = 2.66 with interactions), when $m = 2$ and $n = 2$. For $m = 2$ and $n \geq 5$, $m = 3$ and $n \geq 7$, and $m \geq 4$ and $n \geq 5$, the difference in MDES with and without interactions is always less than 0.01. In other words, except for studies with very small numbers of blocks ($m < 4$) and very small numbers of individuals within treatments within blocks ($n < 5$), the difference between the MDES with and without interactions was negligible. Also, in all situations where the difference exceeded 0.01, the MDES was 0.74 or larger, much larger than effect sizes typically expected in education field trials.

Second, we computed the ratio of the expected length of the 95% confidence interval for the average treatment effect size without interactions in the model to that with interactions as a function of m and n. The largest value of the ratio was 1.03 for $m = 2$ and $n = 2$. For $m = 2$ and $n \geq 5$ and $m \geq 3$ and $n \geq 4$, the ratio is 1 to two decimal places. In other words, with the exception of studies with very small numbers of blocks ($m < 3$) and very small block sizes ($n < 4$), the ratio of confidence interval lengths with and without interactions fitted in the model was essentially the same. Thus, based on the calculations with the MDES and confidence interval widths, there appears to be little loss of sensitivity due to including interactions in the analytic model in most situations.

Do Robust Standard Errors Solve the Problem of Omitted Interactions?

Procedures for computing robust standard errors in linear regression models (e.g., Eichler, 1967; Huber, 1967; White, 1980) provide large sample estimates of standard errors of regression coefficients when residual variances are heteroscedastic. These methods have been incorporated into software packages and they are widely used in economics and the social sciences. It might seem reasonable to expect that these robust standard errors would correct any problems associated with the pooling of interactions that are not in the analytic model into the error term. However, the situation is slightly more complex than it might seem.

Robust variance computations are defined in terms of matrix algebra, which can obscure the actual computations involved. In a balanced design, the matrix algebra simplifies considerably. There are several variants of robust variance estimates that differ by various finite sample corrections, the details of which are beyond the scope of this article. In a balanced randomized block design, the robust variance (the square of the robust standard error) of the estimate of the average treatment effect is

$$(SSW + SSG)/(N - m - 1)mn$$

when the finite sample correction formula proposed by MacKinnon and White (1985; there called HC2), which is the default in STATA, is used. This is sometimes called the "unbiased" formula because it gives an unbiased estimate of the variance of the regression coefficient under homoscedasticity. Note that this is exactly the same (algebraically) as the variance estimate in the linear model that does not fit interactions. Thus, in balanced designs, the use of robust standard errors to account for heteroscedasticity yields exactly the same test as does pooling the interaction variance into the error term or the (exactly equivalent) regression analysis omitting interactions from the analytic model. Therefore, in a balanced randomized block design, tests based on homoscedasticity robust standard errors have the same properties as the regression or pooled analysis of variance tests.

Robust variance estimates can also be computed based on clusters and they provide consistent estimators of the variance of the treatment effect estimate when the number of clusters (the sites in this study) is large. The robust variance of the treatment effect estimate in balanced designs reduces to

$$\left(\frac{N-1}{N-m-1}\right)\frac{MSG}{mn},$$

when the default finite sample correction in STATA is used. Note that this differs by a factor of $(N - 1)/(N - m - 1)$ from the variance obtained via the analysis of variance test with blocks random. Thus, cluster robust standard errors would be an appropriate analysis if the design were balanced, even if the number of clusters (blocks/sites) is small. However, if the design is unbalanced and the number of clusters (blocks/sites) is small, the exact distribution of the test statistic computed from the robust variance estimate is more complex. In particular, it may not have close to a *t*-distribution with $(m - 1)$ degrees of freedom. This point is particularly relevant, because in many cases, the number of clusters (blocks/sites) m will not be large.

Recommendations

Because the properties of analyses depend on the estimand chosen, it is important to specify an estimand before choosing the analysis for an experiment with a multisite design (a point emphasized by Miratrix et al. [2021]). If the estimand is the mean effect in a superpopulation of sites, then the pooling of interactions into the error term (or using fixed effects analyses that do not include site–treatment interactions) should be used only if it is clear that interactions are very close to null. Even interactions that are smaller than are likely to be detectable via preliminary significance tests can seriously elevate actual significance levels. Alternatively, analyses using mixed linear models with sites and site–treatment interactions having random effects described by Raudenbush and Bryk (2002) or Bloom et al. (2017) are appropriate. Analyses using cluster robust standard errors may also be appropriate if the number of blocks *m* is large enough.

If the estimand is the mean treatment effect in the sites observed, pooling has a relatively small effect on actual significance levels, even if small site–treatment interactions are present. Alternatively, design-based approaches (Schochet, 2018), homoscedasticity consistent robust standard errors, or randomization approaches (see, e.g., Calinski and Kageyama, 2000) could be used.

Clearly specifying the estimand is important for interpretation as well as for choice of analysis. Presumably, researchers intend that their work will be interpreted as relevant to some actual educational setting or settings. Unless the experiment is intended as a case study of the particular sites observed, generalization to other sites or settings is at least implicitly intended. It is useful for the researcher to articulate the reason why the results of the study should apply to those other sites or settings. Although the literature on generalization is beyond the scope of this article, there has been research on making that logic more explicit (see, e.g., Stuart et al., 2011 or Tipton, 2013) and endorsements for doing so. For example, the American Educational Research Association's (AERA, 2006) Standards for Reporting Empirical Social Science Research in AERA Journals require that the studies should both "make clear the intended scope of generalization of the findings" and "the logic by which the findings of the study should apply within the intended scope of generalization" (p. 39). Explicit specification of the estimand and how it is linked to the intended scope of findings is one way to satisfy the AERA's standards for reporting.

Appendix

Derivations

Distribution of the sums of squares

Blocks random. If blocks are random (and therefore the block–treatment interactions are also random), the distribution of the sums of squares is given by

$$\frac{SSA}{(\sigma_\varepsilon^2 + n\gamma^2)} \sim \chi_1^2\left(\frac{mn(\alpha_1 - \alpha_2)^2}{2}\right),$$

$$\frac{SSW}{\sigma_\varepsilon^2} \sim \chi_{2mn-2m}^2,$$

$$\frac{SSB}{(\sigma_\varepsilon^2 + 2n\sigma_\beta^2)} \sim \chi_{m-1}^2,$$

and

$$\frac{SSG}{(\sigma_\varepsilon^2 + n\gamma^2)} \sim \chi_{m-1}^2,$$

where the symbols σ_β^2 and σ_γ^2 are the block and block–treatment interaction variance components, $\chi_\nu^2(\lambda)$ is the noncentral χ^2 distribution with ν degrees of freedom and noncentrality parameter λ, and $\chi_\nu^2 = \chi_\nu^2(0)$ is the central χ^2 distribution with ν degrees of freedom (see, e.g., Searle, 1971).

Blocks fixed. If blocks (and therefore block–treatment interactions) are fixed, each sum of squares (divided by a constant) has a central or noncentral χ^2 distribution given by

$$\frac{SSA}{\sigma_\varepsilon^2} \sim \chi_1^2\left(\frac{mn(\alpha_1 - \alpha_2)^2}{2}\right),$$

$$\frac{SSW}{\sigma_\varepsilon^2} \sim \chi_{2mn-2m}^2,$$

$$\frac{SSB}{\sigma_\varepsilon^2} \sim \chi_{m-1}^2\left(2n\sum_{j=1}^{m}\frac{\beta_j^2}{\sigma_\varepsilon^2}\right) = \chi_{m-1}^2\left(\frac{2n(m-1)\sigma_B^2}{\sigma_\varepsilon^2}\right),$$

and

$$\frac{SSG}{\sigma_\varepsilon^2} \sim \chi_{m-1}^2\left(n\sum_{j=1}^{m}\sigma_\varepsilon^2\right) = \chi_{m-1}^2\left(n(m-1)\sigma_G^2/\sigma_{\varepsilon 2}\right),$$

where σ_ε^2 is the within cell residual variance, β_j is the effect of block j, γ_{1j} is the interaction effect of the treatment condition and block j, γ_{2j} is the interaction effect of the control condition and block j, and $\chi_\nu^2(\lambda)$ is the noncentral χ^2 distribution with ν degrees of freedom and noncentrality parameter λ (Searle, 1971).

Distribution of the test statistics

We derive the small sample distributions of the tests for treatment effects based on FPC_R and FPC_F in the blocks random and blocks fixed case, respectively.

The derivations follow a theorem from the Appendix of Hedges (2007b), which we restate here in a slightly modified form.

Theorem. Suppose that $D \sim N(\delta, a\sigma^2)$ and S^2 is a quadratic form in normal variates independent of D and defined, so that $E(S^2) = b\sigma^2$ and $V(S^2) = 2c\sigma^4$, where a, b, and c are constants, possibly involving parameters defined by the design. Then, $F = bD^2/aS$ has the noncentral F-distribution with one degree of freedom in the numerator, b^2/c degrees of freedom in the denominator, and noncentrality parameter $\lambda = \delta^2/a$.

Blocks random. To obtain the sampling distribution of *FP* given in (7), apply the theorem using

$$D = (\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot}),\ S^2 = (SSG + SSW)/(2mn - m - 1),\ \sigma^2 = (\sigma_\varepsilon^2 + n\sigma_\gamma^2).$$

Derivation of a: Under this framework, a direct argument with $\sigma^2 = (\sigma_\varepsilon^2 + n\sigma_\gamma^2)$

shows that

$$a = \frac{Var(D)}{\sigma^2} = \frac{Var(\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot})}{\sigma_\varepsilon^2 + n\sigma_\gamma^2} = \frac{\frac{2(\sigma_\varepsilon^2 + n\sigma_\gamma^2)}{mn}}{\sigma_\varepsilon^2 + n\sigma_\gamma^2} = \frac{2}{mn}.$$

Derivation of b: The constant b is given by $E(S^2)/\sigma^2$. If $X \sim c\chi_f^2$, then $E(X) = cf$ and $Var(X) = 2c^2f$, so that we have

$$E(S^2) = E\left(\frac{SSW + SSG}{N - m - 1}\right) = \frac{E(SSW) + E(SSG)}{N - m - 1} = \frac{(2mn - 2m)\sigma_\varepsilon^2 + (m-1)(\sigma_\varepsilon^2 + n\sigma_\gamma^2)}{N - m - 1},$$

$$= \frac{(N - m - 1)\sigma_\varepsilon^2 + n(m-1)\sigma_\gamma^2}{N - m - 1},$$

Thus

$$b = \frac{(N - m - 1)\sigma_\varepsilon^2 + n(m-1)\sigma_\gamma^2}{(N - m - 1)(\sigma_\varepsilon^2 + n\sigma_\gamma^2)},$$

where $N = 2mn$.

Derivation of c: The constant c is given by$Var(S^2)/2\sigma^4$.
We have

$$Var(S^2) = Var\left(\frac{SSW + SSG}{N - m - 1}\right) = \frac{2(2mn - 2m)\sigma_\varepsilon^4 + 2(m-1)(\sigma_\varepsilon^2 + n\sigma_\gamma^2)^2}{(N - m - 1)^2},$$

Thus

$$c = \frac{(N-m-1)\sigma_\varepsilon^4 + n^2(m-1)\sigma_\gamma^4 + 2n(m-1)\sigma_\gamma^2\sigma_\varepsilon^2}{(N-m-1)^2(\sigma_\varepsilon^2 + n\sigma_\gamma^2)^2}.$$

Blocks fixed. To obtain the sampling distribution of FPC_F given in (13), apply the theorem using

$$D = (\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot}),\ S^2 = (SSG + SSW)/(2mn - m - 1),\ \text{and}\ \sigma^2 \equiv \sigma_\varepsilon^2.$$

Derivation of a: When blocks have fixed effects, a direct argument gives

$$a = \frac{Var(D)}{\sigma^2} = \frac{Var(\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot})}{\sigma_\varepsilon^2} = \frac{2}{mn}.$$

Derivation of b: The constant b is given by $E(S^2)/\sigma^2$. In the blocks fixed case, we use the fact that if $X \sim c\chi_f^2\ (\lambda)$, then $E(X) = c(f + \lambda)$ and $Var(X) = c^2(2f + 4\lambda)$. We have

$$E(S^2) = E\left(\frac{SSW + SSG}{N-m-1}\right) = \frac{1}{N-m-1}E\left(\frac{SSW + SSG}{\sigma_\varepsilon^2}\right) = \frac{(N-m-1)\sigma_\varepsilon^2 + 2n(m-1)\sigma_G{}^2}{\sigma_\varepsilon^2(N-m-1)}.$$

Thus

$$b = \frac{(N-m-1)\sigma_\varepsilon^2 + 2n(m-1)\sigma_G{}^2}{(N-m-1)\sigma_\varepsilon^2}.$$

Derivation of c: The constant c is given by $Var(S^2)/2\sigma^4$.
We have

$$Var(S^2) = Var\left(\frac{SSW + SSG}{N-m-1}\right) = \frac{1}{(N-m-1)^2}Var\left(\frac{SSW + SSG}{\sigma_\varepsilon^2}\right).$$

Since $Var(SSW)/\sigma_\varepsilon^2 = 2(2mn - 2m)$ and $Var(SSG)/\sigma_\varepsilon^2 = 2(m-1) + 4n(m-1)\sigma_G{}^2/\sigma_\varepsilon^2$, so that we have

$$Var\left(\frac{S^2}{\sigma_\varepsilon^2}\right) = 2(2mn - 2m + m - 1)\sigma_\varepsilon^2 + 4n(m-1)\sigma_G{}^2,$$

which gives

$$Var(S^2) = 2(N-m-1)\sigma_\varepsilon^2 + 4n(m-1)\sigma_G{}^2,$$

which implies

$$c = \frac{Var(S^2)}{2\sigma_\varepsilon^4} = \frac{(N-m-1)\sigma_\varepsilon^4 + 2n(m-1)\sigma_G{}^2\sigma_\varepsilon^2}{(N-m-1)^2(\sigma_\varepsilon^2)^2}.$$

Acknowledgments

We thank the editor and two anonymous reviewers for their helpful feedback on earlier versions of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This work was supported by the Institute of Education Sciences (R305B170016).

ORCID iD

Wendy Chan https://orcid.org/0000-0002-0933-9532

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40. https://doi.org/10.3102/0013189X035006033.

Bancroft, T. A., & Han, C. P. (1983). A note on pooling variances. *Journal of the American Statistical Association*, *78*, 981–983.

Bloom, H. S., Raudenbush, S.W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 917–842.

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902.

Bozivich, H., Bancroft, T. A., & Hartley, H. O. (1956). Power of analysis of variance test procedures for certain incompletely specified models. *Annals of Mathematical Statistics*, *27*, 1017–1043.

Calinski, T., & Kageyama, S. (2000). *Block designs: A randomization approach*. Springer.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press.

Daniels, H. E. (1939). The estimation of components of variance. *Supplement to the Journal of the Royal Statistical Society*, *6*(2), 186–197.

Eichler, F. (1967). Limit theorems for regressions with unequal and dependent errors. In L. M. LeCam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (vol. 1., pp. 59–82). University of California Press.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*, *3*(1), 1–21.

Fisher, R. A., & MacKenzie, W. A. (1922). Studies in crop variation II: The manurial response of different potato varieties. *Journal of Agricultural Science*, *13*, 311–320.

Han, C. P., & Bancroft, T. A. (1968). On pooling means when the variance is unknown. *Journal of the American Statistical Association*, *63*,1333–1342.

Hedges, L. V. (2007a). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*, 151–179.

Hedges, L. V. (2007b). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370.

Hedges, L. V., & Rhoads, C. (2009). *Statistical power analysis in education research (NCSER 2010-3006)*. National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, *9*(2), 226–252.

Hines, W. G. S. (1996). Pragmatics of pooling in ANOVA tables. *The American Statistician*, *50*(2), 127–139.

Huber, P. (1967). The behavior of maximum-likelihood estimates under nonstandard conditions. In L. M. LeCam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 221–233). University of California Press.

Kendall, M. G., & Stuart, A. (1968). *The advanced theory of statistics (Vol. 3 Design and analysis and time series*, 2nd ed.). Griffin.

Kirk, R. (1995). *Experimental design* (3rd Edition). Brooks Cole.

Kish, L. (1965). Sampling organizations and groups of unequal sizes. *American Sociological Review*, *30*, 564–572.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325.

Mead, R., Bancroft, T. A., & Han, C. P. (1973). Power of analysis of variance test procedures for incompletely specified models. *Annals of Statistics*, *3*, 797–808.

Miratrix, L., Weiss, M., & Henderson, B. (2021). An applied researchers guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, *14*, 270–308. https://dash.harvard.edu/handle/1/37366188.

Paull, A. E. (1950). On a preliminary test for pooling mean squares in the analysis of variance. *Annals of Mathematical Statistics*, *21*, 539–556.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185.

Raudenbush, S. W., & Bloom, H. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475–499.

Raudenbush, S. W., & Bryk, A. S. (2002) *Hierarchical linear models*. Sage Publications.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213.

Scheffé, H. (1959). *The analysis of variance*. Wiley.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87.

Schochet, P. Z. (2018). Design-based estimators for average treatment effects for multi-armed RCTs. *Journal of Educational and Behavioral Statistics*, *43*(5), 568–593.

Searle, S. R. (1971). *Linear models*. John Wiley.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–316.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A*, *174*(2), 369–386.

Tang, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Statistical Research Memoirs*, *2*, 126–149.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite trials. *Journal of Research on Educational Effectiveness*, *10*, 843–876.

White, H. (1980). A heteroscedastity-consistent covariance matrix and a direct test for heteroscedasticity. *Econometrica*, *48*, 817–838.

Windsor, C.R., & Clark, G.L. (1940). A statistical study of variation in the catch of plankton nets. *Journal of Marine Research*, *3*, 1–34.

Wolde-Tsadi, G., & Afifi, A. A. (1980). Comparison of sometimes pool, sometimes switch, and never pool procedures in the 2-way ANOVA random effects model. *Technometrics*, *22*(3), 367–373.

Authors

WENDY CHAN is an assistant professor in the Human Development and Quantitative Methods Division in the Graduate School of Education at the University of Pennsylvania; email: wechan@upenn.edu. Her research focuses on model-based and nonparametric methods to improve generalizations from small studies in education.

LARRY VERNON HEDGES is the Board of Trustees Professor of Statistics and Education and Social Policy at Northwestern University; email: l-hedges@northwestern.edu. His research interests center on the development and application of statistical methods for the social, medical, and biological sciences. His major research areas within statistics focus on the development of statistical methods for meta-analysis and the design and analysis of experiments.

Manuscript received December 14, 2020
Revision received May 3, 2022
Accepted May 9, 2022