Background / Context: Description of prior research and its intellectual context.

The United States educational system has a natural hierarchical structure (students are nested in classrooms, which are nested in schools, which are nested in school districts, etc.) Evaluation researchers usually utilize these natural groupings when planning randomized experiments. However, the decision regarding which level in the hierarchy to utilize as the unit of randomization is not obvious. On the one hand, there are concerns about the possibility of contamination in designs where subjects receiving the “experimental” condition are in close proximity to subjects in the “control” condition. Contamination (or spillover) occurs when features are the experimental condition that are intended to be experienced only by “experimental” subjects spill over so that “control” subjects receive certain features of the experimental treatment. The possibility of contamination is often cited as a reason for choosing a higher level as the unit of randomization for a study (eg., Donner and Klar, 2000; Bloom, Richburg-Hayes and Black, 2007; Shadish, Cook and Campbell, 2002).

On the other hand, it is well known that randomizing at a higher level (eg. randomizing schools instead of classrooms or classrooms instead of students) will substantially decrease the statistical power and precision of the study (eg., Schochet, 2008; Hedges and Rhoads, 2009; Raudenbush, Martinez and Spybrook, 2007). Rhoads (2011) explored the implications of contamination for experimental design in situations where there is only one level of clustering to account for (students nested within schools). Rhoads compares designs that randomly assign entire schools (cluster randomized, or CR, designs) to designs that randomly assign individuals within schools (randomized block, or RB, designs). Rhoads compares the two designs with respect to two different criteria. The first, based on statistical power, chooses the design that can achieve a given type II error rate with the smallest sample size of clusters. The second criterion is based on mean squared error and chooses the design that minimizes mean squared error (MSE).

Rhoads (2011) finds that the threat of contamination should not necessarily lead experimenters to opt for a cluster randomized design. Depending on the values of relevant design parameters (i.e., the ICC, within cluster sample size, the heterogeneity in treatment effects across clusters, and the number of clusters in the experiment) the statistical power of a RB design remains higher than the power of a CR design even when contamination causes the effect size to decrease by as much as 10-60%. Similarly, from the standpoint of mean squared error, the RB design is preferred to the CR design for many design situations of practical interest.

Purpose / Objective / Research Question / Focus of Study: Description of the focus of the research.

The current paper extends the results found in Rhoads (2011) to experimental designs with multiple levels of nesting. Relevant designs are compared with respect to their mean squared error and with respect to the sample size required (at the highest level) to achieve a fixed type II error rate. The paper also includes a short discussion of the (small) literature providing empirical
estimates of intraclass correlation coefficients in designs with two levels of nesting (students, classrooms, schools) and three levels of nesting (students, classrooms, teachers, schools).

**Significance / Novelty of study:**
*Description of what is missing in previous work and the contribution the study makes.*

Existing practice in the design of field experiments in education research relies heavily on designs that randomize entire schools to treatment. One reason for the reliance on these designs appears to be a fear that randomization at a lower level in the hierarchy will lead to contamination or spillover effects. Rhoads (2011) has shown that even if there are contamination effects they would frequently have to be quite large in order for CR designs to be preferred to RB designs in designs where students are nested within schools. While this work is useful, it is frequently the case that educational researchers need to deal with multiple levels of nesting when they design experiments. By extending the results of Rhoads (2011) to designs with two and three levels of nesting the current paper will help researchers decide when contamination threats should lead researchers to prefer one experimental design to another.

**Statistical, Measurement, or Econometric Model:**
*Description of the proposed new methods or novel applications of existing methods.*

In the interest of brevity, the abstract presents results only for three level models (students nested in classrooms which are nested in schools). Attention is further restricted to the case where designs that randomly assign entire schools are compared to designs that assign classrooms within schools. For clarity of exposition, we assume a fully balanced design, so that a total of \(2m\) schools participate in the experiment, with each school containing \(p\) classrooms and each classroom having \(n\) students. The model and notation are as follows. Let \(Y^E_{ijk}\) represent the outcome of the \(k^{th}\) individual within the \(j^{th}\) classroom within the \(i^{th}\) school who is assigned to the experimental condition and \(Y^C_{ijk}\) represent the outcome of \(k^{th}\) individual within the \(j^{th}\) classroom within the \(i^{th}\) school who is assigned to the control condition. Then

\[
Y^E_{ijk} \sim N(\mu^E_{i}, \sigma^2_{Bl} + \sigma^2_{TS} + \sigma^2_C + \sigma^2) \hspace{2cm} (1)
\]

\[
Y^C_{ijk} \sim N(\mu^C_{i}, \sigma^2_{Bl} + \sigma^2_{TS} + \sigma^2_C + \sigma^2) \hspace{2cm} (2)
\]

\[
\text{Cov}(Y^C_{ijl}, Y^E_{ijk}) = \sigma^2_{Bl} + \sigma^2_C; \hspace{2cm} (3)
\]

\[
\text{Cov}(Y^C_{ijl}, Y^C_{ijk}) = \sigma^2_C + \sigma^2_{Bl} + \sigma^2_{TS}; z = E, C; l \neq k \hspace{2cm} (4)
\]

\[
\text{Cov}(Y^C_{ijl}, Y^C_{ikm}) = \sigma^2_{Bl}; z = E, C; j \neq k. \hspace{2cm} (5)
\]
The model described by (1)-(6) is a standard model to use in the context of hierarchical linear modeling (see, eg. Hedges and Rhoads, 2009; Konstantopoulos, 2011) but with slightly unconventional notation. This notation is adopted to allow for the use of a single notation to describe designs that randomize classrooms within schools and designs that randomize schools.

Notice that the total variance at the school level is decomposed into two parts. $\sigma^2_{Bl}$ represents the variance accounted for when schools are used as blocks and classrooms are randomized within schools. $\sigma^2_{TS}$ represents the variance due to the interaction of treatment with schools. We define $\sigma^2_S = \sigma^2_{Bl} + \sigma^2_{TS}$. The results will depend on the intracluster correlation coefficients (ICCs), defined as $\rho_S = \sigma^2_S / (\sigma^2_S + \sigma^2 + \sigma^2_T)$ and $\rho_C = \sigma^2_C / (\sigma^2_S + \sigma^2 + \sigma^2_C)$. Results will also depend crucially on $\omega_S = \sigma^2_{TS} / \sigma^2_S$, which represents the proportion of between school variation that cannot be removed by blocking. By definition, $0 \leq \omega_S \leq 1$. We use the symbol “RB” to represent the blocked design that randomizes classrooms within schools and the symbol “CR” to represent the cluster randomized design that randomizes entire schools.

It is also necessary to formalize the concept of contamination. The paper restricts attention to contamination processes that will tend to shrink the observable effect size. We assume that contamination will lessen an otherwise positive average treatment effect but cannot make an otherwise positive treatment effect negative. This allows us to formalize contamination as follows.

Let the two treatments under investigation be labeled E and C (for “experimental” and “control”). Let $\mu_E$ and $\mu_C$ be population mean outcomes under treatments E and C in the absence of contamination and let $d = \mu_E - \mu_C$. The “•” notation is used to denote averaging across a given subscript. Thus, the expected value of $Y_{\cdots}^E - Y_{\cdots}^C$ with contamination is given by $E_{RB}(Y_{\cdots}^E - Y_{\cdots}^C) = d(1 - c_T)$, where $0 \leq c_T \leq 1$. We interpret $c_T$ as the proportion of the average treatment effect that is removed by contamination. We assume that randomizing school successfully removes all contamination so that $E_{CR}(Y_{\cdots}^E - Y_{\cdots}^C) = d$.

**Usefulness / Applicability of Method:**

*Demonstration of the usefulness of the proposed methods using hypothetical or real data.*

The usefulness of the method is evident from the description in the Findings/Results section.

**Findings / Results:**

*Description of the main findings with specific details.*

(May not be applicable for Methods submissions)

Results are presented for the evaluation of the two designs with respect to the statistical power criterion. In the interest of brevity, results for the mean squared error criterion are not presented.
Under the assumptions delineated above,
\[
\text{var}_{CR}(Y_{m}^{E} - Y_{m}^{C}) = \frac{2\sigma^2}{mpn} \left(1 + (n - 1)\rho + (pn - 1)\rho_s\right) \frac{1}{1 - \rho_s - \rho_c} \quad \text{and}
\]
\[
\text{var}_{RB}(Y_{m}^{E} - Y_{m}^{C}) = \frac{2\sigma^2}{mpn} \left(1 + (n - 1)\rho + (pn \omega_s / 2 - 1)\rho_s\right) \frac{1}{1 - \rho_c - \rho_s}. \]

The number of schools required to achieve fixed type I and type II error rates under a given design is proportional to \(t^{-2}\), where
\[
t_{des}^2 = \frac{\text{var}_{des}(Y_{m}^{E} - Y_{m}^{C})}{\text{E}_{des}^2(Y_{m}^{E} - Y_{m}^{C})}. \]

The RB design is preferred provided that contamination is small enough that fewer clusters are required to achieve a fixed type II error rate under the RB. It is useful to define the maximum allowable contamination (MAC) which is the largest contamination rate possible such that the RB design will still be preferred. In the current situation MAC is given as
\[
\text{MAC}_{p,het}(n, p, \rho_c, \rho_s, \omega_s) = 1 - \frac{1 + (n - 1)\rho + (pn \omega_s / 2 - 1)\rho_s}{1 + (n - 1)\rho + (pn - 1)\rho_s}. \]

In order to understand the practical implications of the MAC equation given above it is necessary to have some idea of what sorts of ICCs can be expected in educational data. Very little information about ICCs for three level models. The best information available to date comes from Xu and Nichols (2010). The ICCs reported depend on grade level, subject matter and test type. However, the school level ICCs generally range from about .1 to about .3. The classroom level ICCs range from as low as .05 to as high as .40, but are generally also in the .1 to .3 range.

Figure 1 in the Appendix graphs MAC as a function of \(\rho_s\) for a very small classroom ICC (\(\rho_c = 0.05\)) and Figure 2 shows the identical graph for a very large classroom ICC. Results are presented for small (\(n=5\)) and moderate (\(n=25\)) classroom sizes and for a small (\(p=2\)) and a large (\(p=10\)) number of classrooms per school. A horizontal line is added to Figure 1 at \(\rho_s = 0.10\) on the assumption that this is a likely value of the school ICC when \(\rho_c = 0.05\). The same logic accounts for adding a horizontal line at \(\rho_s = 0.20\) to Figure 2. The results obtained are consistent with those found in Rhoads (2011) for two level models. That is, for likely values of design parameters it appears that contamination can remove between 20-60% of the treatment effect before the design that randomizes schools is preferred to the design that randomizes classrooms within schools.

**Conclusions:**

*Description of conclusions, recommendations, and limitations based on findings.*

The current paper extends results in Rhoads (2011) to designs with multiple levels of nesting. The conclusions mirror those in Rhoads (2011). That is, contamination should not lead us to randomize at a higher level unless it removes at least 20% of the treatment effect.
The practical utility of this methodological investigation is limited by the lack of knowledge about how strong contamination processes are. Future empirical work on the potential for contamination in educational experiments is warranted.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


Appendix B. Tables and Figures

Not included in page count.

Figure 1: MAC as a function of the school level ICC when the classroom ICC is small.
Figure 2: MAC as a function of the school level ICC when the classroom ICC is large.