Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 14, Number 2, January 2009 ISSN 1531-7714

Evaluating the Collaborative Strategic Reading Intervention: An Overview of Randomized Controlled Trial Options

John H. Hitchcock, *Ohio University*
Anja Kurki, *American Institutes for Research*
Chuck Wilkins, *Edvance Research, Inc.*
Joseph Dimino, *Instructional Research Group*
Russell Gersten, *Instructional Research Group*

When attempting to determine if an intervention has a causal impact, the 'gold standard' of program evaluation is the randomized controlled trial (RCT). In education studies random assignment is rarely feasible at the student level, making RCTs harder to conduct. School-level assignment is more common but this often requires considerable resources compared to designs where classrooms can be assigned within a school. This article describes the costs and benefits of testing the effects of a classroom based instructional intervention using the multi-site cluster RCT. Topics covered include a discussion of various design options, statistical power, contamination, prior evidence, generalizability of results, ease of recruitment and need for data collection. The purpose of the article is to inform practice by providing program evaluators with an in-depth look at various RCT design options that were considered when searching for a way to efficiently evaluate a school-based intervention.

This article discusses the challenges and advantages of different types of randomized controlled trials (RCTs) when a classroom level instructional intervention is being evaluated. Although many of the take-home points we attempt to convey are independent of the actual intervention described here, it may help to have some background understanding of the trial. The intervention under review for this study is *Collaborative Strategic Reading* (CSR), and the focus is on whether it can impact achievement in 5th grade classrooms with high numbers of English language learners (ELLs). Addressing reading comprehension for these types of classrooms represents a critical need since this is the grade at which students begin to focus on content as opposed to reading skills. Meanwhile this is also the grade at which ELL students are typically transitioned to full English immersion with less support, which of course creates a series of new learning challenges. Yet there are few if any evidence-based programs designed for these students. Indeed, Gersten, Hitchcock, Harps, & Santoro (2008) conducted a *What Works Clearinghouse Review*[1] of related interventions and found only a handful of empirically-supported approaches, and none of them focused on the 5th grade (see also Gersten & Hitchcock, 2008). Other reviews on this topic have, meanwhile, also found a dearth of empirical evidence (e.g., Cheung & Slavin, 2005; Goldenberg, 2008; Slavin & Cheung, 2005).

RCTs have recently received greater prominence in evaluation work given pushes by funding agencies to establish strong causal evidence pertaining to intervention impacts in education (National Research Council, 2002; Raudenbush, 2005; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). This is not to suggest that RCTs should be the only design to consider, particularly because random assignment is not always feasible and there

[1] Details on the What Works Clearinghouse are available at: http://ies.ed.gov/ncee/wwc/. It is noteworthy that these systematic literature reviews engage in extensive searches of the literature to identify causal evidence pertaining to interventions.

are options, such as the Regression Discontinuity design, which can potentially yield unbiased estimates of program impacts (Shadish, Cook, & Campbell, 2001). In addition, the use of propensity score matching (PSM) can add considerable rigor to quasi-experiments assuming there is sufficient data to apply the technique (Rosenbaum & Rubin, 1983; Rubin, 1997). Furthermore, PSM can be used in innovative ways to draw causal data from highly diverse programs such as the effects of special education (Morgan, Frisco, Farkas, & Hibel, 2008), Kindergarten retention (Hong & Raudenbush, 2005), and the impacts of test timing and accommodation in the context of college entrance exams (Rudner & Peyton, 2006). In each example, randomization was infeasible yet researchers were able to draw important causal inference pertaining to program/policy impacts. With that said, when randomization *can* be accomplished there is considerable statistical and theoretical evidence to suggest it is the best overall approach for establishing causal evidence (Holland, 1986; Shadish, Cook, & Campbell, 2001). So with a nod to other techniques that should be in an evaluator's toolkit, we offer a "hands on" overview of RCT design choices we faced in the hopes of helping others to design and evaluate RCTs.

Although the RCT described here attempts to address a knowledge gap pertaining to ELL instruction, the purpose of this article is *not* to describe the outcomes of the evaluation. Instead, this article is predicated on the notion that (a) there are few cases in the RCT literature where authors present various design options they might have used but chose not to, and (b) descriptions of roads not taken can provide useful background for readers interested in designing their own trials, critiquing evidence generated from RCTs, and thinking through practical matters such as sample recruitment. Indeed, the topic of recruitment has not been well covered within the education research literature so far as we know. It is our hope that the article will thus promote a practical overview of cluster RCT design choices, outline their various strengths and weaknesses relative to this type of program evaluation, and increase readers' capacity to understand and critique RCTs.

There is, of course, no one correct RCT design for evaluating the impacts of a classroom level instructional intervention. Rather, design choices should be informed by both practical and analytical considerations. These include the specific features of the intervention (e.g., whether it is delivered at the school, teacher or student level), likelihood of contamination between treatment and control conditions, statistical power, and anticipated difficulty of sample recruitment. Hence, the following section provides an overview of CSR and description of outcome and implementation measures. Once this background is set, we provide a description of various design choices that could have been used for the evaluation and close with a description of the options we finally chose. Along the way, we offer discussion about matters of design, statistical power, and sample recruitment.

AN OVERVIEW OF CSR AND STUDY BACKGROUND

Over a decade of research has examined the processes and efficacy of CSR in heterogeneous classes, which included students with learning disabilities and students acquiring English as a second language (Klingner & Vaughn, 1996, 1998, 1999, 2000; Klingner, Vaughn, & Schumm, 1998; Vaughn, Hughes, Schumm, & Klingner, 1998). With that said, all previous studies have been quasi-experimental or qualitative in nature. CSR incorporates reading comprehension strategies (see Palincsar & Brown, 1984; Rosenshine & Meister, 1994) and cooperative learning techniques (see Johnson & Johnson, 1989; Kagan, 1990). Once teachers are fully trained to implement the intervention, students are taught strategies in a whole class setting. After students are able to implement the strategies independently, they are placed in small learning groups and presented with a reading passage. Prior to reading the passage, students preview the text to determine what they know and what they think they are going to learn. They are also trained to recognize when they comprehend material and when they do not. In CSR parlance, students are told that understanding material means the concepts are "clicking." If students are experiencing difficulties comprehending material, they are "clunking." Students use the "fix-up" strategies, which are based on context and structural analysis to determine meaning. After reading a portion of the text, students use the "Get the Gist" strategy to ensure they grasp the main idea of a given section. Finally, during wrap up, students ask and answer recall questions, as well as questions that require them to use their background knowledge to move "beyond the text." They also write a summary of the selection.

During a CSR session, students of varying reading abilities work in small groups to help each other apply the above strategies. Each child is given a role to help the group members in implementing the process. Teachers remain active participants in the process because they must monitor groups so as to promote conditions that should maximize students' comprehension of expository text. Although CSR applies aspects of cooperative learning, the teacher is primarily responsible for its implementation. An important implication then is that CSR is a teacher-level intervention and this characteristic is oftentimes revisited below.

The primary research question of interest for the study is whether CSR students outperform control students on the proposed outcome measure, the Group Reading Assessment and Diagnostic Evaluation (GRADE;

Williams, 2001), a test of reading comprehension. A second research question is whether the CSR intervention will have an effect on ELL and non-ELL students. Other questions will investigate issues such as whether the level of implementation of CSR is related to the student outcomes. These are however more ad hoc analyses that are not as tightly connected to the design issues discussed below.

To effectively use CSR, students must consistently implement a series of critical, strategic behaviors in a sequential fashion. There are also key teacher behaviors that, when implemented, will facilitate students' comprehension of the selection. Fidelity of CSR implementation is measured using an adapted version of the CSR Implementation Validity Checklist developed by Vaughn et al. (1998). Student behaviors, teacher behaviors and aspects of the classroom environment are rated as either present or absent on the checklist. There is a column for field notes where observers write comments and document any modifications to CSR that occurred during the observation. Field notes help observers respond to a series of questions at the end of the observation. The questions address how well the groups functioned, variability of CSR implementation across groups, evidence of adaptations and modifications and whether any of the strategies were implemented as a whole group.

Incidentally, a second observation tool is used to determine if any of the CSR techniques are adopted in other instructional contexts as well as to develop a better understanding of control group instruction. Hence, an observation measure was developed to gather data on specific conditions and practices under which comprehension and vocabulary instruction is delivered in study classrooms.

DESIGN OPTIONS

Now that CSR has been described along with the context and goals of the study, we cover the factors that influenced our design choices. Such factors included:

- *Potential for contamination between treatment and control conditions*, by contamination we refer to situations where members of the control condition gain access to the treatment and apply it, which would of course obfuscate findings;
- *The 'business as usual' condition*, which will affect the level of statistical power required for the design depending on the degree to which the control group is already incorporating similar reading strategies and/or cooperative learning techniques;
- *Need for generalizability of the results*, which will affect the size and statistical power of the design as well as how schools in the study would be recruited (random sample vs. non-random sample of schools with specific characteristics);
- *Ease of recruitment*, which can affect the level of randomization, statistical power and size of the study;
- *Information from prior studies regarding the effectiveness of the intervention,* affects the required statistical power and consequently the size of the study;
- *Need for original data collection* (student testing, classroom observations, fidelity of implementation check lists, etc.), will together with the budget affect the size and statistical power of the design;
- *Assessment of presumed practicality of findings,* will cover our assessment of the degree to which teachers and administrators will consider the end report of the study to be useful;
- *Spatial Organization and nature of the intervention* (see Bloom, 2005), which deals with the best fit between the unit of treatment delivery and level of random assignment (be it students, teachers, schools, etc.). That is, the unit of treatment delivery for CSR is at the teacher level, which as noted below, opens up the possibility of using classroom-level random assignment.

At the outset, we considered four design options based on these factors: (1) student-level randomization, (2) school-level randomization, (3) classroom-level randomization, and (4) the multi-site cluster trial where classrooms are assigned to conditions *within* each school. It should be noted here that, regardless of the design option, our intent was to compare CSR and CSR only against a strong counterfactual. Hence, we chose to apply a balanced, two-armed trial (i.e., equal sample allocation to one of two treatment conditions). Note that some choices were dismissed more quickly than others. For reasons noted below, option three was dismissed out of hand but others required careful investigation. Following is our assessment of each design option while keeping these factors in mind.

Student Level Random Assignment

It is a matter of course that randomizing students in lieu of classrooms or schools can be quite cost effective, assuming that statistical power assumptions are comparable (e.g., the minimum detectable effects size are the same). But CSR would be difficult to evaluate using this option. Because CSR is a teacher delivered intervention, the natural level of randomization is at the teacher/classroom level. That is,

the intervention targets all children in the eligible classrooms and is typically delivered through regular instruction. Students within a school could possibly be randomly assigned to receive the additional CSR while developing a scheme for keeping instructional time constant across comparison groups. But teachers would then be asked to differentiate their instruction within a classroom while withholding a novel approach from select students. Should teachers view CSR as beneficial, it would be problematic for them to refrain from using it. As we discuss below, there is no way we can conveniently utilize student-level randomization to form treatment and control groups but we did consider assigning students randomly to classrooms over and above randomly assigning teachers/classrooms to treatment and control conditions. The reason for this is we might have improved statistical power by reducing estimates of intra-class correlations (ICCs) in our a priori statistical power analyses. Briefly, the ICC is a ratio of within cluster (e.g., school, classroom, etc.) variance to the total population variance (Bloom, 2005). If an ICC is zero, this allows one to assume the error structure of observations is independent, making standard "single-level" analyses (e.g., independent t-tests, ANOVA) appropriate. In a trial where there is some type of clustering (e.g., students in schools), it is almost always the case that the evaluator must assume there is a non-zero ICC (see Murray, 1998). From a design perspective, note that statistical power decreases as the ICC increases (Raudenbush, 1997; Raudenbush, Martínez, & Spybrook, 2007; Schochet, 2005, 2008).

Statistical Power

The statistical assumptions applied to power analyses described below are listed in Table 1. Also note that the four power equations presented here are pulled from Schochet's (2005) work and (2008) article. Alternatively, a multi-level power analysis software package, *Optimal Design*, is available from the William T. Grant Foundation (Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2008) and can be used to derive similar estimates.

If student level randomization were feasible, fewer schools would be needed for the study. For example, student level random assignment could be conceptualized as random assignment of students to treatment and control conditions within schools, where schools are considered as random effects.[2] Power calculations for this type of a design must acknowledge the possible variation in the treatment effects between schools. Accordingly, the power for this design would be calculated by using the below equation (Schochet, 2005, p. 17).

$$\text{MDES} = \text{Factor}(\alpha, \beta, \text{df}) * \sqrt{\frac{2\rho_1(1-c_1)}{s} + \frac{2(1-\rho_1)}{s(m)}}$$

Where

- MDES refers to the minimum detectable effect size
- s is the number of schools;
- m is the average number of treatment or control group students in each school;
- ρ_1 is the ICC at the school level;
- c_1 is the correlation between the treatment and control group means within school;

If schools on average would include four classrooms with 25 students, and all students could be randomly assigned to treatment or control groups, a total of 6-7 schools would be required to attain a minimal detectable effect size of 0.20, assuming .80 power and the other assumptions listed in Table 1. As the below sections will demonstrate, this is an enviable sample size (putting aside any desire to generalize findings to a broader context).

Recruitment

The required sample being significantly smaller, recruitment could be an easier task. However, "selling" student level randomization may turn out to be a difficult task that may require an already established research partnership with school districts. In particular, depending on the relationship with the school district, securing parental consent for randomizing students as well as for data collection may present a challenge. Therefore student level random assignment, although a statistically 'powerful' alternative may not be the most feasible one.

$$\text{MDES} = \text{Factor}(\alpha, \beta, \text{df}) * \sqrt{\frac{2\rho_1(1-c_1)}{s} + \frac{2\rho_2(1-c_2)}{sk} + \frac{2(1-\rho_1-\rho_2)}{sk(.5n)}}$$

Where

- s = number of schools;
- k = number of classrooms;
- n = number of students in classroom (average)

[2] Alternatively it would be possible that students are randomly assigned to treatment and control conditions within classrooms. In this case, there would be a need to acknowledge not only the extent to which treatment effects vary across schools but also the extent to which treatment effects vary across classrooms within schools. Then the following equation could be used to calculate power

Table 1: Assumptions used in the statistical power analysis calculations.

Level of randomization	Minimal detectable effect size	Intraclass correlation	Significance level , two-tail	Explained Variance[1]	Teachers[2]	Students[3]
School-level randomization	0.20	0.15	0.05	0.5	4	25
Student-level randomization	0.20	0.15 between sites	0.05	0.5	4	25
Classroom level randomization within schools (multi-site cluster RCT)	0.20	0.15 between classrooms 0.10 between classrooms with students randomly assigned to classrooms	0.05	0.5	4	25

[1] Variance of the outcome explained by baseline covariate
[2] Average number of teachers per grade level
[3] Average number of children per classroom

Establishing a Control or 'Business as Usual' Condition

The fact that students are randomly assigned within a school (or classroom) to either receive the intervention or regular instruction should yield a very clear control or "business as usual" condition where many qualities for the treatment and control groups are the same (school environment, classroom environment, classroom teacher, other instruction).

Generalizability of Results

Usually schools recruited for a student level randomized trial are from one district/city due to the small number of schools required. Hence the results will tend to be less generalizable beyond a specific location, compared to other options (all things being equal). Moreover, if the participating schools are purposively selected results cannot be generalized to all schools even within the district.

Data Collection Costs

Student level randomized trials, requiring the fewest schools, have the lowest costs of data collection. As a result, a student level randomized trial (acknowledging parental consent issues) offers an opportunity for more original data collection (student testing, observations) and qualitative data collection.

Connection between Unit of Treatment Delivery and Randomization

A design based on student level randomization is not an ideal alternative for a teacher-level intervention such as CSR, as it is delivered through teachers during regular instructional time. Thus, the appropriate level of treatment delivery and randomization is a classroom/teacher. If CSR could be implemented as a supplemental program, student level randomization would become a feasible option. However, the treatment would then not be delivered in a regular classroom environment, but perhaps through instruction given by paraprofessionals or instructional specialists.

School-level Random Assignment

School-level random assignment is an appropriate design in a situation where the classroom/teacher level intervention is difficult to deliver within schools without worry of contamination, or where proper implementation requires that all teachers (in the school, at the grade level, in certain subject area) use the intervention. Moreover, this design helps when it is plausible that providing the intervention to a random sample of teachers from a school may negatively affect the existing school culture and atmosphere. For example, consider a design that endeavors to assess the impact of a core curricula package. Proper implementation of core curricula will generally

require participation from all teachers within a school (or grade) because it can be too much to ask teachers to vary the packages if they are expected to coordinate plans throughout the year and help align curricula across grades. Hence, any related impact study would probably make school level assignment the design of choice. But even if working with an intervention that could conceivably be implemented with a sub sample of teachers, there are multiple reasons why implementation of a specific intervention would be prone to contamination:

- The intervention is desirable from the teachers' point of view (it may carry some prestige; training may yield more marketable skills, etc.).
- The intervention is believed to provide substantial benefits for students.
- The intervention targets high-need children.
- The intervention is easy to learn and materials used in implementation are widely available.
- The school's community is collaborative, encouraging sharing of materials, ideas and approaches to teaching.

Keeping these caveats in mind, we still felt that contamination in the CSR study was a remote possibility because successful implementation of the intervention requires a mix of initial training, follow-up coaching, and familiarity with materials that are not widely available. Put another way, if casual contact between treatment and control teachers would diminish observed treatment impacts to an appreciable degree, then the training milieu would be unnecessary and our understanding of CSR would be quite poor. At any rate, although school level randomization is the most appropriate design for whole school interventions and it addresses contamination in classroom/teacher level interventions, it has its drawbacks. The most notable of them is the design requires a larger sample of schools.

Statistical Power

Designs relying on school level random assignment require larger samples of schools and classrooms. In addition, depending on the planned analyses (whether students are modeled as nested within classrooms/teachers, and classrooms/teachers nested within schools); a minimal number of classrooms/teachers per grade may be required.

MDESs were calculated using the equation:

$$\text{MDES} = M_{J-k} * \sqrt{\frac{\rho(1-R^2)}{P(1-P)J} + \frac{(1-\rho)}{P(1-P)nJ}},$$

where

$M = (t_\alpha + t_\beta)$, and is the multiplier that translates the standard error into a minimum detectable effect estimate. It is equal to the *t* critical value for α, the significance level of the intended statistical test, plus the *t* critical value for β, the likelihood of detecting significant effects given a true effect of a particular size (i.e., the power of the test);

ρ= intra-class correlation between schools, assumed to be 0.15 (see Schochet, 2005, 2008);

P= the proportion of treatment schools;

J = the total number of schools in the analysis;

R^2=the amount of variance in the outcome that school-level pretest explains, assumed to be 0.5.

n = the number of students within each school.

For example, a paired school level randomization that would match similar control and treatment schools before randomization would require 66 schools for minimal detectable size of 0.20 (other assumptions being as defined above). This is close to ten times the number of schools needed for the student-level assignment design, and as noted below, close to double required of the classroom/teacher level multi-site clustered trial.

Recruitment Costs

The best possible recruitment involves long-term relationship building which begins well before a specific research project is launched. However, study timelines often require rapid recruitment, making studies with school level random assignment less plausible. Oftentimes it is not plausible and/or desirable to recruit a large enough number of schools from one school district (for reasons such as face validity; generalizability of results, etc.). Recruitment from multiple sites (school districts) requires a careful approach in which criteria for recruitment that are appropriate for the specific study are decided well in advance.

Establishing Control or 'Business as Usual' Condition

Due to the larger sample requirement and potential problems related to recruitment, the final sample may include schools from multiple districts that may be quite different regarding school level demographic characteristics, such as school size, percent of students receiving reduced/free lunch, percent of minority students, etc. Establishing appropriate recruitment criteria will alleviate this problem. However, the randomization process used to create treatment and control groups has to be carefully thought out in order to incorporate potential preferences by participating school districts (such as "each geographical sub-district should have one school in treatment condition") and potential blocking required for

establishing equivalent treatment and control groups. In other words, face validity of random assignment can often be promoted by appropriate blocking procedures to ensure comparable numbers of school-types are in each treatment arm (Raudenbush, Martinez, & Spybrook, 2007). Moreover, although randomization of schools may create equal treatment and control groups at the school level, it is possible that statistically significant differences exist at teacher or student levels. For example, some schools may experience problems retaining teachers, making the experience level of teachers in schools very different.

Generalizability of Results

Usually schools recruited for a school-level randomized study are not a random selection of eligible schools from the participating school districts, hence limiting the generalizability of the results. Moreover, school districts participating in the study often form a convenience sample, mostly driven by the recruitment efforts. As a result, the criteria used for recruitment have doubled importance: the criteria will not only identify schools that are appropriate for the study (such as high poverty or high ELL percentage), but will also affect the face validity and generalizability of results.

Data Collection Costs

The sample of teachers/classrooms in a school-level RCT is larger, thus increasing the cost of teacher-level data collection. In addition, student data collection is likely to be more expensive and complex to implement, even if a random sample per classroom is tested. Although these are concerns that exist in other designs, sampling of students in school level randomized studies intensify questions related to attrition and management of missing data. Furthermore, if students are sampled during the fall, how should students who attrite during the year be treated in the analytical models?

Connection between Unit of Treatment Delivery and Randomization

If the intervention is delivered at the classroom level, it often does not require whole school implementation to be successful. If the risk of contamination between treatment and control classrooms is relatively low, a design that includes randomization at the classroom level is likely to be a more cost effective alternative, often requiring less than half of the schools and classrooms than a design applying school level random assignment. For these reasons, we were not satisfied with school-level assignment for the CSR study.

Classroom Level Randomization (Ignoring School Clustering)

Although one could use pure classroom level random assignment in which classroom are randomly assigned to treatment and control conditions without considering school-level clustering, the study team decided at the outset that this approach is not appropriate for the CSR study. It could be argued that recruiting schools with similar background characteristics and students would alleviate the problem caused by the clustering of classrooms within schools; we believe that classrooms that are clustered within schools should be explicitly accounted for in the study design. In our experience, there are too many school-based factors, such as differences among core curricula, that might undermine causal inference. Thus, the team ended up randomly assigning classrooms/teachers to treatment and control groups within schools (classroom level multi-site trial) which is discussed in detail below.

The Multi-site Cluster RCT

The final design that was considered for this study was a multi-site cluster RCT. This design is in essence a series of mini-experiments, conducted across a number of sites/schools, where classrooms are randomly assigned to either treatment or control. In a multi-site cluster RCT, each control classroom is compared to treatment classrooms within the same school, ensuring no school-based differences in those comparisons, as well as providing an opportunity to replicate program impacts across multiple sites (see Schafer, 2001).

Of course, when comparing treatment and control groups in an RCT, it is always important that the groups be as similar as possible on all factors except the intervention. Randomization is the primary tool used to achieve this end, but while the process eliminates group differences *on average*, in any particular instantiation there will still be differences between the treatment and control groups on unknown variables to an unknown degree (this is the basic problem raised in the previous section). While no design can guarantee complete equivalence on all non-experimental factors between the control and treatment groups, there are ways to improve the results over and above randomization, such as stratification. Blocking is of little use if there is little variation between the blocks created for randomization purposes. However, in this design a block is a school, and it is often safe to assume reading instruction techniques and core curricula will likely vary across sites. Designing the study to minimize such differences improves power and quality of the results. In a design in which schools contain only treatment or control classrooms, the school-based differences provide an additional source of variation that is eliminated in the multi-site design, thus improving power and reducing the necessary sample size. See Kalaian (2003) for additional details on design logic and Raudenbush & Bryk (2002) for a

detailed review of HLM models that can be used to test intervention impacts.

Recruitment

All other things being equal, using a multi-site design can make recruiting easier. First of all, the number of schools required by this approach is (typically) nearly half as many as in a school-based design. Schools might also be reluctant to participate in an RCT unless they believe that the intervention is likely to be beneficial. In a school-based design, there is a 50%[3] chance that the school will end up as a control school and none of the students will benefit from perceived benefits of the intervention, at least as part of the experiment. On the other hand, in a multi-site design, every school will be fully participating in the study, with at least a portion of classrooms getting the intervention. There may still be some concern among control teachers, but it is usually easier to explain the need for some control classrooms, than it is to explain that an entire school may need to serve as a control. Of course, there may be motivation issues with control teachers, but that is no different than in other designs, and is not a recruiting issue per se.

In our own recruitment efforts for the CSR study, we have thus far been successful with arguing that control teachers may actually be better off than the treatment teachers because they can make decisions about using the intervention with the benefit of hindsight. That is, they don't need to put the effort in learning how to use it until after confirmatory evidence becomes available and after seeing it in action within their schools. Furthermore, CSR teachers can and are being trained to disseminate what they learn to colleagues, so control teachers will have the option to learn the approach once the study is over.

Data Collection

There are always complex logistical features associated with collecting data in a large RCT. While this design does not eliminate these, it can ameliorate the process. In school-based designs, it is often easier to get data from treatment schools (who are more invested in the study) than from control schools. In the multi-site design, while control teachers may be less invested in the study, the logistical issues at the school-level may be easier to deal with, since the school officials (e.g., the principals) tend to be heavily invested in the study.

Connection between Unit of Treatment Delivery and Randomization

In the multi-site design, the classroom serves as both the level of random assignment and as the level at which the CSR intervention is being delivered. When evaluating a teacher-led intervention such as CSR, this can promote the use of parsimonious analytic models and promote inference. This congruence is critical for the analysis and interpretation of the results. As mentioned earlier, some other designs, particularly student-level assignment, do not have this feature.

Statistical Power

For the multi-site design, the MDES can be calculated by the following equation (assuming schools as fixed effects):

$$\text{MDES} = \text{Factor}(\alpha, \beta, \text{df}) * \sqrt{\frac{2\rho_2}{s(.5k)} + \frac{2(1-\rho_2)}{s(.5k)n}}$$

Where

s is the total number of schools in the sample

k is the number of classrooms per school[4]

n is the average number of students per classroom

ρ_2 is the between-classroom variance (ICC)

In most situations, the MDES is reduced (i.e., power is increased) by adding additional classrooms than by adding students within a classroom. In order to see this, note that when n is equal to $(1- \rho_2)/ \rho_2$, both terms in the square root are equal. For larger n, the first term is larger. For a typical ICC=.20, this means that, if there are at least four students per classroom, then the first term is larger, and this term is reduced only when classrooms are added, not students. The take home point is that K, the number of classrooms, is far more important for determining statistical power than n, the number of students per classroom. This can have practical implications for design and execution of a trial. For example, student-level attrition leads to far less precision loss than classroom level attrition.

It should also be noted that the above equation does not take into account baseline covariates, which can greatly increase statistical power. For purposes of comparing the relative power of the multi-site RCT with the school-based design discussed earlier, if we use the same assumptions that yielded a required sample size of 66 schools, we find that the multi-site design requires only 19 schools.

Statistical Analysis Issues

[3] This could of course be higher or lower in an unbalanced design (i.e., designs with unequal sample sizes in treatment conditions).

[4] This calculation assumes that there is a balanced allocation of classrooms between treatment and control. In practice, there may be an odd number of classrooms available in some schools, so this balance will only be achieved overall, not within each school.

In addition to power issues, there are other aspects of statistical analysis tied to the choice of design. Of particular importance in a multi-site design, the variation of the treatment effect can be looked at across schools if one uses a random effects model (which would require a slightly larger sample size). Then, by considering ways in which the schools are different from one another (e.g., enrollment, curriculum, %ELL, etc.), exploratory hypotheses can be formed to account for differential treatment effects across schools as a function of various school factors. Consider the CSR study again. It may be the case that some schools are more adept at utilizing cooperative learning techniques simply as a function of experience. Should the data analysis suggest intervention impacts vary across schools, we might explore whether such experience explains the variation. There may be a host of additional factors that explain effect size variation (if it does indeed vary) and such analyses can do much in terms of advancing program knowledge.

Adding Student Level Random Assignment to the Design

Initially for this study, we had considered randomly assigning students to classrooms, in addition to randomly assigning the classrooms to treatment conditions, in the hopes of improving power. Again, the mechanism through which we might increase power would be to reduce the ICC. There were two reasons why we decided not to do so in the end. First there were the logistical difficulties. While schools often do not have a problem with the random assignment of classrooms to condition, they are much more hesitant about having the student randomly assigned to classrooms. Although many students may be assigned by a pseudo-random process, there are certainly many exceptions where students are purposely placed in one class or another. And even if the school's process is essentially random, getting them to let an experimenter do the random assignment and coordinate the effort is difficult.

The second reason was that the power benefits turned out to be rather modest. Randomly assigning students to the classrooms does not change the fact that the intervention is being administered at the classroom level, so clustering effects must still be taken into account. Student level random assignment eliminates (on average) any classroom level effects at the beginning of the year, but by the time of posttest, students have been clustered in those classrooms for an entire academic year. It was determined that, for our study, assigning the students to classrooms randomly would lower our end of year ICC from .15 to .10 due to the baseline reduction in ICC, but this did not yield a large enough increase in power to justify the logistical difficulties alluded to earlier.

Contamination

The only real drawback to the multi-site design, at least compared to the school-level design, is the issue of contamination, in which control students become exposed, to one degree or another, to the treatment. In this case, the concern would be if control teachers were to implement CSR due to finding out about it from treatment teachers. As mentioned earlier, there are many reasons why such contamination might occur.

If contamination occurs, the control condition is no longer an appropriate counterfactual, and the ability of the study to find a statistically significant difference is compromised. Essentially the probability of a type II error is inflated (i.e., power is reduced). If contamination is determined to be likely, and steps cannot be implemented to prevent it, then the multi-site design should not be used. Therefore, it is important to (a) evaluate the likelihood of contamination, (b) implement processes to ensure contamination does not occur and (c) track the occurrence of contamination so that, if it does occur, appropriate logistical steps can be taken to stop it and appropriate statistical modifications can be taken during the analysis phase.

With that said, our evaluation of this threat for CSR determined that contamination was minimal. Treatment teachers require both a two-day training on how to use the intervention as well as follow-up coaching throughout the academic year. Coaching is offered because it is probable that the two-day training alone will not lead to adequate use of CSR and we would not expect to find a significant treatment impact without this effort. CSR is not something that can be easily transmitted via casual contact between teachers; serious contamination would require control teachers to gain access to techniques best disseminated via coaching.

Although we felt confident that contamination would not be a problem, we nevertheless took steps to ensure that it would not occur. It would take a concerted effort on the part of both a CSR teacher and a corresponding control teacher for contamination to take place, and it was felt that educating teachers on the importance of obtaining scientifically valid results was the most important and best way to prevent contamination. Hence, we emphasized to both treatment and control teachers the importance of maintaining the integrity of treatment conditions. We also will use the aforementioned observations to check for the use of CSR in control classrooms. If minor contamination is observed, we can intervene to try to prevent it from becoming serious. Specific classrooms in which serious contamination is observed can be noted, so that sensitivity analysis can be done to see if the treatment impacts are different in those classrooms.

CONCLUSION

RCTs have seen limited use in education because of concerns surrounding intervention fidelity, politics in assignment, and concerns that study outcomes had limited utility (Cook, 1999; Gueron, 2002). This trend does appear to be changing given federal initiatives that promote the use of randomization in study designs. Again, we do not wish to lose sight of the fact that other designs allow one to draw causal inferences about program impacts, but we do argue that when it is possible to use a RCT then one should do so. Hence, the aim of this paper was to describe different RCT design choices when evaluating a classroom-level intervention, in the hopes of providing readers with a practical overview of various options. There is of course no single correct RCT design for a classroom level instructional intervention. Different approaches are more or less appropriate, depending on the practical and analytical circumstances of the study. We believe that, given the examples provided here, the multi-site cluster RCT yielded the best overall option. But the point of the paper is not to advocate for this design, but instead to walk the reader through the various options we considered. The paper also endeavored to demonstrate the dynamic nature of designing RCTs – the final design was developed and modified over time to better match the analytic and pragmatic constraints. It is our hope that readers will benefit from our presentation of the choices we considered, relative to the characteristics of the CSR intervention, and apply some of the above ideas when reading about RCTs or even when designing one.

References

Bloom, H.S. (2005). *Learning more from social experiments: Evolving analytic approaches.* New York: Russell Sage Foundation.

Cheung, A., & Slavin, R. (2005). Effective reading programs for English language learners and other language-minority students. *Bilingual Research Journal*, *29*(2), 241-267.

Cook, T.D. (1999, March). *Considering the major arguments against random assignment: An analysis of the educational culture surrounding evaluation in American schools of education.* Paper presented at the Harvard Faculty Seminar on Experiments in Education, Cambridge, MA.

Gersten, R., & Hitchcock, J.H. (2008). What is credible evidence in education? The role of the What Works Clearinghouse in informing the process. In S.I. Donaldson, C.A. Christie, & M.A. Mark (Eds.). *What counts as credible evidence in applied research and evaluation practice?* (pp. 78-95). Thousand Oaks, CA: Sage Publications, Inc.

Gersten, R., Hitchcock, J.H., Harps S., & Edwards-Santoro, L. (2008, March). *The What Works Clearinghouse Review of academic interventions for English-language learners: Implications for the instruction of Spanish-speaking students and future research.* Paper presented at the 2008 Annual Meeting for the American Education Research Association, New York.

Goldenberg, C. (2008). Teaching English language learners: What the research does – and does not – say. *American Educator*, *32*(2), 1-44.

Gueron, J.M. (2002). The politics of random assignment: Implementing studies and affecting policy. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 15-49). Washington, D.C.: Brookings Institution.

Holland, P.W. (1986). Statistics & causal inference. *Journal of the American Statistical Association, 81,* 945-970.

Hong, G., & Raudenbush, S.W. (2005). Effects of Kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation & Policy Analysis, 27*(3), 205-224.

Johnson, D.W., & Johnson, R.T. (1989). Cooperative learning: What special education teachers need to know. *The Pointer*, *33*, 5-10.

Kagan, S. (1990). *Cooperative learning resources for teachers.* San Juan Capistrano, CA: Resources for Teachers.

Kalaian, S. A. (2003). Meta-analysis methods for synthesizing treatment effects in multisite studies: hierarchical linear modeling (hlm) perspective. *Practical Assessment, Research & Evaluation*, *8*(15). Retrieved January 14, 2009, from http://pareonline.net/getvn.asp?v=8&n=15.

Klingner, J.K., & Vaughn, S. (1996). Reciprocal teaching of reading comprehension strategies for students with learning disabilities who use English as a second language. *The Elementary School Journal, 96,* 275-293.

Klingner, J.K., & Vaughn, S. (1998). Using collaborative strategic reading. *Exceptional Children*, *30*, 32-37.

Klingner, J.K., & Vaughn, S. (1999). Promoting reading comprehension, content learning, and English acquisition through collaborative strategic reading (csr). *The Reading Teacher*, *52*, 738-747.

Klingner, J.K., & Vaughn, S. (2000). The helping behaviors of fifth graders while using collaborative strategic reading during ESL content classes. *TESOL Quarterly*, *34*, 69-98.

Klingner, J.K., Vaughn, S., & Schumm, J.S. (1998). Collaborative strategic reading during social studies in heterogeneous fourth grade classrooms. *The Elementary School Journal, 99,* 1-22.

Morgan, P.L., Frisco, M.L., Farkas, G., & Hibel, J. (2008). Effects of special education services on children's learning and behavior: A propensity score matching

analysis. *The Journal of Special Education*. Retrieved January 14, 2009, from http://sed.sagepub.com/pap.dtl.

Murray, D. M. (1998). *Design and analysis of group randomized trials*. Oxford: Oxford University Press.

National Research Council. (2002). Scientific research in education. In R. J. Shavelson & L. Towne (Eds.), *Committee on Scientific Principles for Educational Research*. Washington, D.C.: National Academy Press.

Palincsar, A.S., & Brown, A.L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*, 117-175.

Raudenbush, S. W. (1997). Statistical analysis and Optimal Design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185.

Raudenbush, S.W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34(5)*, 25-31.

Raudenbush, S.W., & Bryk, A.S. (2002*). Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5-29.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*, 479-530.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rubin, D.B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*(8 Suppl. 2), 757–763.

Rudner, L M., & Peyton, J. (2006). Consider propensity scores to compare treatments. *Practical Assessment Research & Evaluation*, *11*(9). Retrieved January 14, 2009, from http://pareonline.net/pdf/v11n9.pdf.

Schafer, W.D. (2001). Replication: A design principle for field research. *Practical Assessment, Research & Evaluation*, *7*(15). Retrieved December 2, 2008, from http://PAREonline.net/getvn.asp?v=7&n=15.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W.H., & Shavelson, R.J. (2007). *Estimating casual effects using experimental and nonexperimental designs*. Washington, D.C.: American Education Research Association.

Schochet, P.Z. (2005). *Statistical power for random assignment evaluations of education programs*. Princeton, NJ: Mathematica Policy Research.

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

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Slavin, R.E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Education Research, 75*(2), 247-284.

Spybrook, J., Raudenbush, S.W., Liu, X., Congdon, R., & Martínez, A. (2008). *Optimal Design for longitudinal and multilevel research: Documentation for the "Optimal Design" software*. New York: William T. Grant Foundation. Retrieved January 14, 2009, from http://sitemaker.umich.edu/group-based/files/od-manual-20080312-v176.pdf.

Vaughn, S., Hughes, M.T., Schumm, J.S., & Klingner, J.K. (1998). A collaborative effort to enhance reading and writing instruction in inclusion classrooms. *Learning Disability Quarterly*, *21*, 57-74.

Williams, K.T. (2001). *Group Reading Assessment and Diagnostic Evaluation*. Circle Pines, MN: American Guidance Service.

Citation

Hitchcock, J.H., Kurki, A., Wilkins, C., Dimino, J., and Gersten, R. (2009). Evaluating the collaborative strategic reading intervention: An overview of randomized controlled trial options. *Practical Assessment Research & Evaluation*, *14*(2). Available online: http://pareonline.net/getvn.asp?v=14&n=2.

Author

Address Correspondence to:

John Hitchcock
Ohio University
College of Education
McCracken Hall, 305(B)
Athens, OH 45701

hitchcoc [at] ohio.edu.