Causal Inference and Assessing Learning Outcomes Across Multiple Methods: A Primer for International Evaluators in Education

John H. Hitchcock, Anthony J. Onwuegbuzie, Shannon David, and Anne-Maree Ruddy

ABSTRACT
A key consideration when assessing learning outcomes is generating and assessing evidence for whether such outcomes are influenced by the introduction of a new program or policy. That is, policymakers and researchers must evaluate causal evidence to differentiate those procedures that yield desired learning from ones that do not. Studies evaluating causal evidence have been utilized in the United States over the last two decades; yet, international education policymakers have only recently begun to recognize the benefits of such studies to inform policies in education. This article provides a primer on causal validity, and reviews several of the design threats (e.g., regression-to-the-mean, maturation, history effects, sample loss, selection) that undermine the capacity to make internally valid statements. The article also provides a series of randomized controlled trial design considerations in the context of thinking about this approach to conducting research within a broader program of inquiry.

KEYWORDS
Causal validity; program evaluation; randomized controlled designs; education policy

A common step in education research and evaluation is generating evidence about whether student outcomes (e.g., learning outcomes) are caused by some change in a program or policy (e.g., Bloom, 2005; Chen, 2005; Gersten et al., 2005; Nastasi & Hitchcock, 2008, 2016; Reynolds, 1998; Shadish, Cook, & Campbell, 2002; Slavin, 2002). In these situations, researchers must evaluate causal evidence to differentiate procedures that yield desired outcomes from ones that do not and, in turn, make this information available to decision makers. Evaluating causal evidence is particularly relevant today, at a time when several governments are investing in public education while attending to policies and practices that are deemed to be effective (e.g., Flannagan, 2016; Stoker & John, 2009). Yet, causal attribution should be made with some care. After all, if basic alternative (i.e., rival) explanations for some observed improvement cannot be ruled out, expansive changes to education practice might be taken that could turn out to be fruitless, and this can come at enormous public expense. Therefore, it is critical to promote the use of causal evidence when crafting education policy.

The initial focus of this article, therefore, is on offering a primer on causality and how causality can be considered across a systematic program of education research. This can entail use of multiple methods. For example, researchers might start with pilot studies that might include small-n studies such as single-case designs (SCDs) (e.g., Kratochwill et al., 2013), and then testing interventions under ideal circumstances (e.g., efficacy randomized controlled trials [RCTs]), to real-world applications (e.g., effectiveness RCTs), and eventually to synthesizing causal evidence across a variety of studies (e.g., meta-analyses, research syntheses, meta-syntheses, meta-summaries). Because RCTs in education almost always deal with causal inference in the context of clustered units (e.g., classrooms, schools), part of this article offers a description of how multilevel modeling (see Raudenbush & Bryk, 2002) is used in these designs to account for contextual effects. From there, a brief review of how study designers might integrate observations, interviews, and so on to generate a
stronger understanding of treatment contrasts (e.g., Grissmer, 2016; Nastasi & Hitchcock, 2016) is offered. Overall, we contend that causal inference is best supported in the context of a coordinated program of research, which will typically entail use of multiple methods.

A Primer on Causality in the Context of Education Research

To infer causation, one is making a guess that some event (e.g., introduction of a new learning program or way of teaching) yields a change in some variable, such as a learning outcome (cf. Rubin, 1974; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007; Shadish et al., 2002). Use of the words infer and guess is purposeful. The reason for this is the so-called fundamental problem of causal inference (see Schneider et al., 2007). With respect to this problem, the issue at hand is that once a unit (e.g., person, student, or groups such as a classroom) is exposed to some change (e.g., a new instructional program), one cannot observe how the unit would have performed had the change not occurred. That is, unless there is reason to believe that a unit can fully revert back to baseline (i.e., the condition before treatment), with no carry over effects, one cannot directly observe a counterfactual condition. This idea of observing a unit in two conditions without worry of carry over effects might work in some engineering examples—particularly when inanimate objects are involved—such as when seeing how equipment performs in one condition, ensuring there was no change to the equipment after data collection (e.g., any wear and tear is judged to be negligible), and then observing this same unit in another condition. With this assurance that the equipment has not been changed, the fundamental problem of causal inference can be solved. In the social sciences, however, such an assurance is rarely plausible; we almost always assume that people will change in some manner after exposure to some new condition. Consider education programming with students who mature over time and who are continuously exposed to new circumstances; as we might hope, any operating assumption is that they are always changing. Furthermore, students cannot, of course, be in two conditions at once (e.g., one cannot simultaneously teach them mathematics using a new curricula and a standard curricula), yielding a fundamental problem when making causal attributions about education packages (e.g., programs, interventions).

Within experimental design, causal inference can be viewed via a lens of internal validity, which deals with the degree to which an observed change in a dependent variable is likely a function of the introduction of a treatment, or an independent variable (see Shadish et al., 2002). If this inference is easily defensible within the confines of a study, then the findings/interpretations have high internal validity. If there are multiple plausible explanations for an observed change other than the treatment, then the findings/interpretations have low internal validity. Ideally, policymakers will make decisions about education programs on the basis that they are supported by studies that yield findings/interpretations with strong internal validity (that the program, or intervention, in fact, yields theorized outcomes). This also holds for intervention developers, who should work towards generating strong causal evidence. It is, however, the case that many (not all) intervention developers have conducted multiple studies to establish that a given teaching approach is supported, but do not adequately address plausible alternative explanations for a treatment effect (cf. Brass, Nunez-Neto, & Williams, 2006; Cook, 2002; Cook, Scriven, Coryn, & Evergreen, 2010; Shadish et al., 2002). An intervention developer might, for example, design a pre-post study without a comparison group—known as a one-group pretest-posttest design. For example, a new computerized mathematics package is theorized to yield improved learning outcomes in third-grade mathematics. The design strategy used is to pretest a large group of third graders, use the new package (i.e., a treatment) for an academic year, and then posttest the students in the study. Now, assume that students demonstrated tremendous growth in measured mathematics skills, using a psychometrically sound instrument. Assume further that this type of study has been replicated several times. Given this description, is there sufficient evidence for a policymaker to recommend that schools adopt the mathematics package? Skepticism is warranted.

The reason for such skepticism is not because the observed growth is doubted, but rather the reason for it. Recall, the fundamental problem of causal inference. These students could not have been in both the counterfactual (i.e., not exposed to the computerized mathematics package) and in the treatment; therefore, we cannot easily rule out plausible alternative explanations for the observed effect. More to the point, one cannot know whether the observed growth is due to the treatment, or whether some other factor drove the improvement because a pre-post design without a control group does not yield evidence where alternative plausible explanations are easily ruled out. Indeed, the design in this example is commonly referred to as a pre-experimental design (cf. Gay, Mills, & Airasian, 2011) because it is considered to be an extremely weak insofar as controlling for many threats to internal validity (e.g., history, maturation, testing, instrumentation, regression) and external validity (e.g., pretest-treatment interaction) (Gay et al., 2011). In fact, as contended by Gay,
Mills, and Airasian (2008), pre-experimental designs are so weak and potentially misleading that they should be avoided; as such, these designs are not useful for causal purposes. This use of a one-group pretest-posttest design leaves key, unanswered questions. One can assume that there was indeed a change in test performance, but what are the potential reasons for it? Is exposure to the new computerized mathematics package the single best explanation for the improved scores? Consideration of these questions are informed by the so-called threats to internal validity (Benge, Onwuegbuzie, & Robbins, 2012; Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979; Onwuegbuzie, 2003; Shadish et al., 2002). In other words, the internal validity for any given study is potentially threatened by a number of alternative explanations, unless a design renders them implausible. In the scenario described earlier, there is no control group; therefore, a simple alternative explanation for the growth in learning outcomes is student maturation. Recall, that in every study conducted by the hypothetical developer, students were in the third grade, pre-tested at the beginning of the year, and posttested at the end; and, consequently, any observed growth simply can be attributed to the fact that study participants were a year older and exposed to a full year of instruction. In other words, observed improvement might have occurred independently, or even in spite of, the computerized mathematics package.

There are a number of other potential threats to internal validity when a control group is not used. For example, regression-to-the-mean (Campbell, 1957) also can occur; this is a phenomenon wherein extreme scores tend not to be replicated. This threat is especially prevalent in studies where the lowest performing students are selected for treatment and then posttested. Any observed improvement may be attributed, at least to a degree, to this threat. To explain, Sir Francis Galton famously observed that very tall fathers tend to not have children as tall as themselves, and vice versa (Salsburg, 2001). Applying this idea to the world of education program evaluation, students who score very low on a measure might well show some improvement on a later test, not because they necessarily have improved skill but because of the regression-to-the-mean phenomenon, and one might erroneously attribute the improvement to a treatment effect.

Yet another worry in an individual study can be the history threat, and this is the case where another event occurs that might explain an observed treatment effect. For example, a group of students might be exposed to the mathematics package the same year that a new principal started at the school and this person’s management style yields a better learning environment. If so, this could have caused the improved scores. A tricky element to this threat is that researchers might not have the contextual knowledge necessary to account for such event that could impact all students in a study and that might also influence performance on the outcome measures. To make matters worse, it is plausible that several of these threats (e.g., maturation, regression-to-the-mean, history) can come into play and, therefore, it can be quite difficult to be assured that any observed improvement is a function of the mathematics package, even if allowing for solid outcome measures of mathematics skill, large sample sizes, and replication. Hence, skepticism around causal claims is warranted when designs of these types (i.e., ones without a comparison or control condition) are used, and decision makers, therefore, might not be in a position to recommend wide adoption of a new program based on these research designs.

Fortunately, several of these threats can be reasonably accounted for by simply adding a comparison condition to any given study. The logic at work is that we cannot simultaneously observe one student in both conditions relevant to a study; thus, one option is to work with two groups of students, both of which are, on average, very similar to each other. In the parlance of experimental design, the two groups are, on average, equated. That is, if there are no systematic differences between the treated group of students (i.e., the group exposed to the mathematics package) and the control group (presumably receiving standard instruction), then the counterfactual can be observed. Importantly, adding a control group renders implausible threats to internal validity. Consider maturation; if the observed improvement in scores is driven by the fact that students had one year of instruction, then we should see similar growth in learning outcomes across both study conditions. If there is such a similarity, we might infer that the mathematics package did not yield a meaningful instructional effect. If, however, there is much greater growth among treated students compared to those in the other group, then maturation is no longer a plausible explanation for an observed effect, yielding findings with much better internal validity. This also holds for the aforementioned regression-to-the-mean and history threats because of the same logic, which is that these phenomena should occur in both conditions. In sum, inferring a causal relationship between the mathematics package and desired learning outcomes becomes much more defensible when a comparison group is used, assuming the comparison group did not perform as well as the treated group on outcome measures.

When adding a comparison group, it does become necessary to attend to the process by which students are placed (or selected) into each study condition. Consider an approach where it is expedient to allow students to volunteer to attend classes with the new mathematics package or to remain in the standard instructional condition. With this approach, there is limited assurance of there being no systematic differences between
groups. This is because it might well be that more motivated students self-select into the treatment condition. And if these same students outperform their control counterparts on a posttest, then we cannot remove motivation as a plausible explanation for the improved learning outcomes. There might be other reasons that drive group selection, some of which we might think of and others that we might not, but collectively, the addition of a control group adds a new threat to internal validity called selection. A common way to control for this threat is to randomly assign students to treatment or control conditions, yielding the RCT design. The use of random assignment can render the selection threat implausible and greatly improve causal inference because it can be argued that the two study groups are, on average, equivalent.  

When random assignment is used, the design can be viewed as a true experiment on the basis that any estimate of the magnitude of the treatment effect will not be biased by other factors. However, random assignment is no panacea for establishing causal inference, and evaluators must attend to a number of additional design considerations. One general consideration is that the two conditions should be treated as similarly as possible, in ways that can isolate a treatment effect. For example, if the group is exposed not only to the new mathematics package, but also to several other new interventions during the course of the study, then a meaningful treatment effect solely for the mathematics package cannot be estimated. This threat is known as multiple treatment interference.

Another consideration is that the results of the random assignment must be carefully protected, otherwise a number of additional internal validity threats ensue. Random assignment is often undermined by units changing condition (i.e., crossing over from one condition to another); when large numbers of crossovers occur, it becomes impossible to have assurance that the two study groups are equated. Another concern is sample loss, also called study attrition or mortality. The concern here is that sample loss undermines equivalence if a large number of students leave, or especially if there is large, differential sample loss from one study group as opposed to another. For example, if a RCT begins with 200 students across two study conditions (say 100 students per condition), and 50 leave one group but not the other, then any resulting causal inference is undermined because it is difficult to know whether such loss is responsible for any posttest differences observed in the study. Fortunately, the What Works Clearinghouse (WWC) has identified sample loss thresholds that inform whether attrition is likely to bias treatment impact estimates badly (WWC, 2014), and there are statistical corrections (e.g., multiple imputation and maximum likelihood estimates) that can be used to estimate the likely treatment impact had all data been available (see Enders, 2010; Puma, Olsen, Bell, & Price, 2009). But a general stance to adopt as the best way to address attrition, and more broadly missing data, is to not have any in the first place (cf. Allison, 2002; Boruch, 1997). To that end, careful planning that focuses on the prevention of missing data can be conceived by program evaluators and it also helps to document carefully exactly when sample loss occurs, from which group, and at what stage of the study (Schulz et al., 2010).

Furthermore, collecting follow-up quantitative data (e.g., demographic information or other statistical data representing the cognitive, affective, and/or psychomotor domain) and qualitative data (e.g., interviews, observations) with attriters can help researchers understand their characteristics and the reason for why people left the study, respectively, and this, in turn, can inform thinking about the reason why data are missing and provide overall information that can be used when making causal inferences (e.g., Nastasi & Hitchcock, 2016). Ultimately, it is assumed that these causal inferences have the potential to be the basis of informed decision making for researchers and policymakers alike.

Planning RCTs gets at the heart of a key aspect of education program evaluation, and several nations are at a point where these designs should be considered to develop an empirical understanding of whether desired learning outcomes are being achieved via their changes in programming. For instance, the United Kingdom has instigated a number of RCTs in medicine, public health, and now in more recent times, education. What began as a considered approach in the British Prime Minister’s office in 2010 has morphed into RCT utilization by multiple government units and the development of quasi-government organizations, resulting in approximately 300 studies. Increasingly, the U.K. government and quasi-government organizations have commissioned an ever-growing number of RCTs over the last decade in public policy and specifically in education (Pearce & Rama, 2014). Goldacre (2013) is a prominent public voice for the increase in RCTs in education in the United Kingdom, expounding on the fledgling approach from the sciences. Australia, on the other hand, is new to conducting RCTs in the world; yet, the focus has been predominantly on health or public health with a significantly lower albeit stagnant percentage of RCTs in education over the last two decades (Ames & Wilson, 2016).
Program Evaluation and the Ethics of Random Assignment

Questioning the ethics of withholding a promising treatment from students should always be encouraged. To weigh some of the issues, suppose that policymakers are thinking of compelling schools to adopt a series of cooperative learning techniques, such that students are expected to help each other learn. Embedded in this idea is the expectation that cooperative learning causes improvements in academic achievement. So, what is the basis for this expectation? Is it based on theory? Is it because the approach is popular in other countries? Could there have been promising evidence from earlier studies? Whatever the case, it is clear that someone has conjectured that the cooperative learning approach will yield better learning outcomes; therefore, readers should wonder about the ethics of withholding it from students. A fundamental question of ethics follows: Is it fair to withhold a treatment? To begin to answer this question in the context of education RCTs, another question can be posed: Are we sure that the treatment works? Begin with an assumption that the answer is actually, yes; we can be sure that the treatment works for the types of students and contexts to be included in the study. If so, then the ethical issues are clear; the study should not be conducted on the basis that a treatment that is known to work is being withheld, and, furthermore, there is little to be learned from conducting the study. Yet, careful thought here is warranted. In many cases, we have precious little causal evidence to guide education programming. First, many new interventions make sense, but one often learns that, in fact, they do not yield the hoped for outcomes that developers had in mind. Consider that in most education studies, use of a true control group that actually receives no instruction is rare. In medical investigations, one might introduce some pharmaceutical intervention in the form of a pill, and a control group takes a placebo pill that has no active ingredients. However, in education settings, children are already receiving instruction. The challenge then is not to deliver an education in the first place, but to figure out which instructional alternatives are best. For this reason, the use of the phrase comparison group is often used in lieu of a control group to provide a reminder that most education RCTs do not compare treatment and no treatment conditions, but rather some variant of what we might call Treatment A versus Treatment B.

In a case where a computerized mathematics package is being tested, students in the comparison condition would need to receive standard mathematics instruction in order to yield a meaningful, policy-relevant question, which focuses on whether the new approach yields better learning outcomes compared to status quo teaching, and status quo teaching might be quite effective. Hence, any ethical objection to random assignment in an education RCT should be considered in light of whether we truly know whether a new instructional approach is an unambiguous improvement over and above existing efforts. In many cases, policymakers will be hard pressed to state that they are sure some new approach is better, and this is often the basis for justifying random assignment because the point of the study is to yield the sort of evidence that allows for strong causal inference.

This line of thinking applies even when dealing with well-known instructional ideas. For example, there is much empirical support for cooperative learning at both the secondary (e.g., Ginsburg-Block, Rohrbeck, & Fantuzzo, 2006; Johnson & Johnson, 2009; Rohrbeck, Ginsburg-Block, Fantuzzo, & Miller, 2003) and post-secondary (e.g., Onwuegbuzie, Collins, & Elbedour, 2003; Onwuegbuzie, Collins, & Jiao, 2009; Onwuegbuzie & DaRos-Voseles, 2001) levels; however, it appears from a quick review of search engines that little is known about how cooperative learning works in different international settings compared to status quo instruction. Therefore, it is plausible that cooperative learning is not better compared to what is already being undertaken. Therefore, even if dealing with a well-known instructional approach, it might be argued that we do not have empirical evidence that it makes a difference in a specific setting.

In other words, there are threats to the external validity of findings pertaining to cooperative learning, including population validity (i.e., extent to which findings pertaining to cooperative learning are generalizable from the sample of individuals on which a study was conducted to the population from which the sample was drawn), ecological validity (i.e., extent to which findings pertaining to cooperative learning from a study can be generalized across settings, conditions, variables, and contexts—thereby representing the extent to which these findings are independent of the setting or location in which the study took place), temporal validity (i.e., extent to which research findings pertaining to cooperative learning can be generalized across time—or the extent that these findings are invariant across time), and specificity of variables (i.e., occurs when one or more of the following seven types of variables used in the cooperative learning study are so unique to the study that the findings are not generalizable: type of participants, time, location, circumstance, operational definition of the independent variables, operational definition of the dependent variables, and types of instruments used) (cf. Benge et al., 2012; Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979; Onwuegbuzie,
Another ethical point is that one can design a study so that comparison children are exposed to the treatment if it is found that there is a benefit at the end of the study. With this feature in place, children would be treated with some empirical evidence supporting the notion that the new approach yields an instructional benefit. For yet another point, consider the possibility that implementing new approaches requires intense resources, and it is often impossible to deliver such approaches at full scale. Hence, if not all students can be treated at once, randomly choosing a group is arguably a fair method to use to determine who will be treated first, unless it is necessary to select students on the basis of triage assessment where the most needed are identified. One last point is that it is generally a requirement that people should agree (consent) to be in a RCT, with the understanding that they might or might not be in a new instructional condition, and that it might or might not yield intended learning outcomes. Should potential study participants demand to be in one study condition or another, then they can often be accommodated with the understanding that they will not be in the study.

Implementation Studies, Efficacy, and Effectiveness Trials

RCTs tend to yield intensive studies that require a lot of planning, and so the developmental status of an intervention should be considered before mounting a trial. A common program (or pattern) of studies that one might follow would be to pilot a procedure, and figure out what seems to work well, what teachers tend to find acceptable, and so on. Earlier, the idea of conducting studies without a comparison group has been criticized, but this is only if the primary intent at hand is to obtain causal evidence. However, if the focus is on intervention implementation logistics, teacher preferences, obtaining a sense of delivery cost, and so on, then a design without randomization or even a comparison group might be reasonable, supporting the idea of pursuing a program of study with multiple methods. Furthermore, there is nothing inherently inappropriate about thinking through the causal evidence that such designs yield, as long as there is a healthy appreciation for the limitations borne out of threats to internal validity and the focus remains on implementation. That is, quasi-experimental designs (e.g., designs with a comparison group formed without randomization; time series designs) and pre-experimental designs can play a role. As Shadish et al. (2002) note, RCTs happen to be an important element in an effort to understand treatment impacts but these designs cannot be the only options available, especially because it might be premature for an evaluator to focus on causal questions. Indeed, because most RCTs are expensive and require considerable planning, it is often advisable to use these designs only after an intervention has been supported by earlier, promising evidence.

Once implementation needs are better understood, a common technique might be to use a pilot study where the best possible implementation is attempted. This is the so-called efficacy study (Flay et al., 2005). The goal here is to determine whether an intervention works under ideal circumstances. If conducted in the context of randomization, one has an efficacy trial (or efficacy RCT). If positive results are found, this might not yield the information that policymakers need to inform decision making because implementation was optimal. Therefore, a follow-up study would be to determine whether a treatment worked under real-world implementation. This is generally referred to as an effectiveness study; so, if random assignment is used, one might think of this as an effectiveness trial (or effectiveness RCT) (Flay et al., 2005). Once an effectiveness trial is conducted, researchers are in a better position to comment on whether the instructional approach should be further disseminated and the degree to which findings hold across other settings, students, measures, and so on, or what Campbell (1957) originally coined as external validity (p. 297). To be clear, external validity always should be evaluated, whether dealing with pilot studies, efficacy trials, effectiveness trials, or repeated investigations, but such validity questions can take on greater prominence when there is evidence that real-world implementation of a new instructional approach yields intended learning outcomes.

A General Design Shell for a School-Level RCT

So, how does one conduct an RCT in education settings? Following is a general design shell. This means it lacks many details that one would normally need to propose a real study. For example, specific research questions are not posed; the purpose of the design shell is to promote thinking about how RCTs might in general be pursued in an education setting. Note that the following design shell accounts for some common real-world problems that undermine the capacity simply to assign students to study conditions and to examine treatment effects with simple independent samples t tests and effect size estimates.
Some initial changes in education might come at the school level. That is, whole schools will be asked to change procedures. If so, whole school random assignment should be considered. There are both logical and statistical issues that support this statement (Bloom, 2005). Logically, if one wishes to change whole schools, then it makes no sense to randomly assign classrooms or students to conditions, although assigning whole groups to study conditions raises some complexities with respect to standard statistical assumptions. As an aside, data should be normally distributed (i.e., normality assumption) and the variance in scores should be approximately the same in both treatment and control groups (i.e., homogeneity of variance assumption). However, statisticians have performed data simulations and found that commonly used models (e.g., independent t tests, analysis of variance [ANOVA]) are generally robust to violations of the assumptions that data are normally distributed and variances across groups are homogenous and [approximately] equally sized (Glass, Peckham, & Sanders, 1972; Maxwell & Delaney, 2004). Notwithstanding, there is an assumption that cannot be violated: observations must be independent (e.g., Maxwell & Delaney, 2004; Stevens, 1996).

Perhaps the most straightforward explanation of the independence assumption is that knowing the score (i.e., observation) of one individual in a study should not provide information about the score of another individual. That is, there is independence between and among observations (e.g., test scores) within a group. Suppose that one conducts an RCT to determine whether daily aspirin use (i.e., the treatment) yields better heart outcomes. Now, imagine two treated patients sitting next to each other. Is it reasonable to assume that the impact of aspirin, if any, on heart health for one person will not influence the next person? This should be the case. If one person experiences a great benefit, then there is no reason to think that this will influence the next person. That is, the scores should be independent. Consider different education scenarios, where the assumption of independence is not a reasonable one to make. In general, should one assume that a student’s academic achievement is totally unrelated to the achievement of a peer in the same classroom? It seems likely that independence should not be assumed if the students receive the same instruction from the same teacher(s) and interact with each other while learning. Although it is to be expected that two observations of their learning are different and separate, it is also to be expected that their scores will be related and otherwise not independent—hence, violating the independence-of-errors assumption wherein errors are correlated.

As another example, consider the assumption in schools. If one can think of some schools as being average in student performance levels, whereas others have a historically poor record, and so on, then a researcher can make an informed guess as to whether a student will perform well on an examination by knowing his or her school membership. If so, then it does not follow that the achievement observations to be made in the school are independent. As yet another example, suppose one is conducting a study of cooperative learning. Because this means that students would be expected to teach each other, then the researcher has a strong compulsion to believe that the scores are interconnected because students who understand material well should teach it better than should those who do not. In sum, in the case of education RCTs, the assumption that individual scores are independent is generally not tenable. McMillan (2000) refers to this assumption violation as treatment replication error, which prevails when researchers collect data that are not consistent with the unit of analysis that is of interest. The most common form of treatment replication error occurs when an intervention is administered once to each group of participants or to a few classes or other units, yet only individual outcome data are collected (McMillan, 2000). As described by McMillan (2000), this practice seriously violates the assumption that each replication of the intervention for each and every participant (e.g., student) in the study is independent of the replications of the intervention for all other participants. Moreover, in the case where there is one administration of the intervention to a group, whatever idiosyncrasies that emerge from this administration are confounded with the intervention, thereby likely yielding systematic errors (McMillan, 2000). Additionally, individuals within a group likely influence other group members with respect to the outcome measure(s). Subsequently, this confounding yields alternative explanations to any observed findings (Onwuegbuzie, 2003). And this confounding is even more problematic when the intervention is administered to groups over a long period of time because the number of confounding variables increases as a function of time (McMillan, 2000). Consequently, as noted by Onwuegbuzie et al. (2003), the internal validity of the findings stemming from the majority of the intervention research—including RCTs—in the area of cooperative learning is seriously threatened because of this treatment replication error.

Hierarchical Linear Modeling in School-Level RCTs

If a researcher has a study where whole schools are to be given a treatment, or not, but cares about student-level performance, then a statistical, and design approach, that can help deal with the independence assumption is Hierarchical Linear Modeling (HLM; see Raudenbush & Bryk, 2002). Think of linear modeling as regres-
sion (also note that curvilinear models can be handled). Recall that a standard statistical model for comparing two group means is the independent samples t test. Also, recall that a t test can be converted to a regression model, where a score on a learning outcome is dependent variable (y), there will be an intercept (a), and one dummy (i.e., dichotomous) predictor variable (X, whether a unit was treated or not) and an error term (e), such that: \( y = a + X + e \). This would yield a single-level model. By contrast, a hierarchical model can be conceptualized as many single-level models that are in essence simultaneously estimated and combined into higher aggregated units. In the context of a RCT, where researchers assign schools to a new treatment or not, one can examine the degree of correlation between student scores within schools and statistically account for violations of the independence assumption at the student level (we still assume that school-level scores are independent and more complex models can handle situations where they might not be).

As an example, suppose that evaluators identify a whole school reform approach and wish to assess empirically whether it causes improvements on some learning outcome. If such evidence is found, then policymakers can have reason to promote use of the approach across many schools, especially if the study results have been replicated. The study typically would begin with a proposal. Study proposals should include a rationale and a review of pertinent literature to justify the research questions. Key constructs of interest should be described, the theory for why the intervention should have an impact on the learning outcomes (dependent variables) should be explained, and the psychometric properties of outcome measures should be reviewed. Assume that these features are in place and a basic question is developed: To what extent do schools exposed to the whole school reform model outperform comparison schools on “X” measure? Notice that the question is causal in nature. So, one would want to consider an RCT design, where schools are randomly assigned to have access to the reform model, or not.

An independent samples t test would be justifiable here except that students are grouped by schools. In HLM parlance, students are nested, or clustered, within schools and the assumption of independent scores is untenable for reasons described earlier. As an important side point, some analysts might see the fact that one should not use an independent samples t test as an annoyance. But HLM analyses can and do offer opportunities for analyses of interesting research questions and making inferences about whole groups (e.g., schools). It is hoped that readers ultimately consider several of the advantages of HLM analyses that will not be made readily apparent here because of the simplicity of the research question, and the fact that this article should be considered a primer. Now, consider the following model, drawn from Raudenbush and Bryk (2002):

Level 1 (Student): \( Y_{ij} = \beta_0 + \epsilon_{ij} \)

Level 2 (School): \( \beta_0 = \gamma_0 + \gamma_0 W_j + \gamma_0 X_j + \ldots + u_{0j} \),

where Postscripts i and j, respectively, index the student and school (so, it is appropriate to consider student i from the jth school);

\( Y_{ij} \) is the outcome of student i from the jth school;

\( \beta_0 \) is the intercept for school j;

“…” is the location where covariates and their coefficients can be entered if baseline equivalence is not established for them (which can happen when random assignment is used);

\( \gamma_0 \) is the estimated grand mean adjusted for the proportion of treatment and comparison schools;

\( \gamma_0 \) is the mean difference between the treatment and comparison group;

\( W_j \) is the treatment indicator (a dichotomous variable), 0.5 for treatment and -0.5 for control;

\( X_j \) is the pretest covariate (i.e., grand-mean centered) for the jth school;

\( e_{ij} \) is the student-specific error (for student i from the jth school), which is assumed to be normally and independently distributed; and

\( u_{0j} \) is the school-specific error, which is assumed to be normally and independently distributed.

This model is used simultaneously to estimate relationships at two different levels. The first is the student level, where: Level 1 (Student): \( Y_{ij} = \beta_0 + \epsilon_{ij} \). Y ij is the learning outcome score (Y) for student “i” in school “j.” If there are 20 schools in a study, then j goes from 1 to 20. If there are, on average, 200 students per school, then there are 4,000 students; one can think of i ranging from 1 to 4,000. \( \beta_0 \) is the intercept for a given school. The way that the model is set up, this value would be the average school score. Thus, the term: \( Y_{ij} = \beta_0 \) indicates that one would guess any given student’s score by knowing the average for that school. Therefore, if a school had an average score of say 500 on some measure, \( Y_{ij} = \beta_0 \) means that the researcher would guess the score for any student “i” in school “j” can be estimated by knowing the school’s average score, \( \beta_0 \), and allowing for some variation.
of individual scores around this mean (eij). The “...” reference allows for other predictors in the model, such as indicators of student socioeconomic status.

Now consider Level 2 of the model: $\beta_{0j} = \gamma_{00} + \gamma_{01}Wj + \gamma_{02}Xj + \ldots + uoj$. It is critical to notice that the $\beta_{0j}$ term, the same one from Level 1, is now a dependent variable. In this way, the Level 1 and Level 2 models are linked together. Therefore, whereas the school average is a key term used to guess the particular score of a student, the level 2 model deals with estimating the school average. The first term here is the grand mean, denoted as $\gamma_{00}$. The term $\beta_{0j} = \gamma_{00} + uoj$ means that the average score of a school is estimated by knowing the total grand mean, plus allowing for error. In fact, a diagnostic step typically followed when conducting these analyses is to conduct a null model. This model allows for estimating score variance for students within schools, and variance between schools. Briefly, the procedure will yield a statistical estimate of the degree of score dependence of students within schools, called the intraclass correlation (see Raudenbush & Bryk, 2002), or ICC. If the ICC is zero, the independence assumption is not violated (but do not expect this to be the case). Also, the more important model term now is: $\gamma_{01}Wj$. This term is at the heart of the research question: To what extent do schools exposed to the whole school reform model outperform comparison schools on “X” measure? Recall that $\gamma_{01}$ is the mean difference between the treatment and comparison group. A zero value indicates no difference (meaning no treatment effect, or the whole school reform did not impact learning outcomes). Any non-zero value must be considered in a descriptive sense. $Wj$ is the treatment indicator (a dichotomous variable), typically 0.5 for treatment and -0.5 for control. If the value, which again is the observed mean difference between treatment and control schools, is statistically significant, then one can reject a hypothesis of no difference. In other words, the null hypothesis that the treatment and control schools being equal in the population of schools, is untenable. Any decision about the null hypothesis also should be considered in light of confidence intervals and effect sizes, which are typically the observed mean difference between schools. A mean difference can be standardized so as to aid interpretation (Grissom & Kim, 2005). Typically, one divides a mean difference by a pooled standard deviation (i.e., the standard deviation reflecting variance in the treatment and control groups; also known as Cohen’s (1988) $d$) or by the comparison group standard deviation only (i.e., known as Glass’s [1976] $d$). The trade-offs here entail working with a variance estimate that is informed by a larger sample size, which is obtained by using data from both groups, or working with an estimate that only uses comparison group data, and, thus, variance that would not have been altered by treatment exposure. In principle, both approaches can be used if doing so yields important information, although analysts should be careful to report the estimate that they believe to be most accurate when communicating with policymakers.

It is common in school-level trials to attempt to conduct well-powered studies with as few schools as possible. One technique is to include the pretest covariate in analyses (e.g., how schools performed on a measure at baseline) because this can yield a more powerful analysis, in the sense that there is greater statistical power. It is, however, the case that the most important assumption that must be met for analysis of covariance (ANCOVA) and multiple analysis of covariance (MANCOVA)—namely, homogeneity of regression slopes—implies that the covariate must be highly correlated with the dependent variable(s) but not related to the independent variable (in this case, treatment assignment). However, as noted by Henson (1998), few covariates exist that meet these criteria—especially when study participants are not randomly assigned to groups (i.e., in quasi-experimental designs), which are endemic to educational research. Unfortunately, if an appreciable correlation exists between the covariate and the independent variable, as is often the case, then the covariate also can reduce the variance in the independent variable—culminating in reduced power and effect size. Thus, the homogeneity of regression assumption means that the regression slopes of the covariate and the dependent variable in each group must be identical, or at least similar, if the single pooled regression slope can be used accurately with all groups. To the extent that the individual regression slopes are different, the part correlation of the covariate-adjusted dependent variable with the independent variable will more closely resemble a partial correlation, and the pooled regression slope will not provide an adequate representation of some or all of the groups. In this case, the ANCOVA/ MANCOVA will introduce bias into the data instead of providing a correction for the confounding variable (Henson, 1998).

Ironically, as noted by Henson (1998), ANCOVA/ MANCOVA typically is appropriate when used with randomly assigned groups; however, it is typically not justified when participants are not randomly assigned to groups—as is the case when schools are assigned to (intervention) groups. Two general considerations in the context of school-level trials are then to determine whether a school-level variable, such as average achievement from a prior year, is used, and to ascertain whether baseline covariates yield a clearer estimate of a treatment effect by using sensitivity analyses (i.e., conduct modeling with and without baseline covariates to determine whether findings hold regardless of the analytic approach used). In the event that findings do not converge across
methods, it will be important to report the discrepancy and, in many cases, the more conservative of the two sets of results should be highlighted.

For now, the key conceptual point to grasp is that the model accounts for clustering of students within schools, and the treatment indicator, which is at the school level, allows for an estimate of a treatment effect. Another issue to consider is that the model itself does not account for how schools were assigned to treatment and control conditions. For example, it could be that principals volunteered to go through the school reform process. However, if schools were randomly assigned to conditions, we can think of the term \( y_01W_j \) as offering an unbiased estimate of the treatment impact. This sets the stage for policy discussions. Appendix A offers a brief overview of some random assignment options to consider. More broadly, HLM can be deployed to handle a number of clustering scenarios in education research.

**Statistical power.** A priori statistical power analyses are needed when planning RCTs. Readers should reference Optimal Design Software (Spybrook et al., 2011) or PowerUP! (Dong & Maynard, 2013), both of which are free and on-line. For purposes of this work, a multi-level power analysis was conducted assuming the aforementioned statistical model. In addition, the following additional assumptions were made:

- A two-tailed statistical test is to be conducted with a nominal \( p \) value of .05 as the cutoff criterion.
- There are at least 200 students per school (on average).
- The minimum detectable effect size is 0.25. Note that considerable theory, practical knowledge and a good sense of the outcome measure is needed to know whether this is reasonable or not.
- There is some pretest or other Level 2 covariate that explains 64% of the variance in the dependent variable (i.e., there is a .80 correlation between the baseline and posttest scores). Note that this corresponds with the term \( \gamma_{0Xj} \) above.
- The degree of correlation between students in schools, called the intra-class correlation, is .20. This is a commonly used value based on some prior empirical work, but it is easily altered.
- The desired statistical power is .80 (i.e., the probably of a Type 2 error, or failing to reject the null hypothesis when one should have is .20). This is a nominal value. Sometimes less power is acceptable, such as when conducting a pilot efficacy trial.

Given all of these features (i.e., desired design where schools are randomly assigned, the planned statistical model and aforementioned assumptions), approximately 40 schools are needed for the study; that would be 20 schools assigned to the treatment, and 20 comparison schools.

**Other design features.** With a basic design in place, many considerations remain. None of these are trivial and should be thought through carefully and with consultation. Issues to consider are:

- How will the sample of schools be selected?
- How representative is the sample of the population of interest?
- Sample loss, or attrition, can undermine the study’s validity; so, how can this be prevented?
- How expensive is the proposed treatment relative to current educational approaches? Leech and Onwegbuzie (2004) refer to addressing this question as determining the economic significance of the proposed treatment.
- How else can missing data be prevented and handled when it occurs?
- Is this an efficacy or effectiveness trial?
- How will the treatment be implemented?
- What training is needed to deliver a school-based intervention?
- Who will undertake the random assignment and how?
- What are the measures?
- How will data be collected?
- How will resulting information be used?
- How will it be reported?
- How do stakeholders perceive the new program?

There are other issues as well, but these are not described here because the focus of this work is to offer an introductory primer. Furthermore, these questions raise the need for use of multiple methods and programmatic research. This primer promotes the idea of conducting RCTs with the intent of generating solid, causal evidence that informs policymakers and decision-making across multiple and different studies. Note also that RCTs themselves can be enhanced by applying qualitative research techniques to address several of the questions listed above, such as developing a sense of how a new program is perceived by stakeholders. There are a number of additional ways in which qualitative research techniques might be added to the RCT design (see Grissmer, 2016; Nastasi & Hitchcock, 2016; Shadish et al., 2002).
Programmatic Research Using Multiple Studies and Methods to Build Causal Evidence

Consider a loose analogy: think of the evidence from any given study as a plant, or a flower. Then, think of most large-scale education decisions requiring a whole garden of evidence. Large, well-conducted RCTs can go a long way toward identifying effective programs, but singular studies warrant replication and it is often the case that addressing generalization concerns is undertaken by examining learning outcomes in a variety of settings. For these reasons, programmatic evaluation research to inform education policy is warranted. As noted earlier, there are different types of RCTs. Some can focus on piloting a program, others can represent more fully powered efficacy studies conducted to ascertain whether an intervention can work, given ideal implementation circumstances, and still others can be conceptualized as large-scale effectiveness trials conducted to see whether an intervention can work in real-world (i.e., less than ideal) circumstances. Any one of these styles of RCTs can be replicated and extended depending on the research questions at hand.

This raises two key points. One is that the RCT design has been described in the United States and other areas as the gold-standard of research. This moniker is used because the design has an almost unique capacity to generate what is thought of as an unbiased estimate of a treatment impact (Boruch, 1997). This idea that the RCT is a gold standard does, however, apply only when a causal question is at hand. The RCT itself, if defined by the use of randomization techniques to allocate units to study conditions, has no other special properties. As noted previously, there are any number of useful non-causal questions in program evaluation work that pertain to, for example, program implementation and feasibility, cost, and acceptability among stakeholders. It is also true that there are other designs that can yield findings/interpretations with reasonable internal validity with respect to causal inference (e.g., regression discontinuity, single-case, various quasi-experiments that establish baseline equating between study groups). Although RCT designs hold a special place with respect to the pursuit of causal questions, it is important to keep in mind that a program of inquiry almost always entails combining this form of inquiry with other forms because comprehensive program evaluation often entails multiple types of questions. Indeed, it is normal to combine a number of other design elements into a RCT; one might, for example, interview program implementers, people who have left the study to understand why, conduct observations of control instruction, perform mediating and moderating variable analyses, and so on—all of which typically are considered to be extensions of the key causal question at hand (e.g., Grissmer, 2016; Nastasi & Hitchcock, 2016; Shadish et al., 2002). The overall point here is that RCTs rarely should be thought about in isolation given the wide variety of evaluation questions.

A second point is that RCTs will normally be thought of as an important tool in a systematic and planned body of research. It is certainly possible to examine whether causal properties hold over individual students by combining single-case studies with RCTs (e.g., Hitchcock et al., 2014). One approach that can be helpful when positing broader causal questions can be to statistically synthesize the observed effect sizes from multiple RCTs in the form of a meta-analysis (e.g., Hedges & Olkin, 1985; Lipsey & Wilson, 2001). Broader types of research syntheses can potentially be conducted that includes both meta-analytic findings as well as any information that comes from repeated studies that deal with program implementation and acceptability via qualitative inquiry, mixed methods approaches, and survey work (Newman, Hitchcock, & Newman, 2015) in a way that would yield not only understanding of replicated causal evidence, but also other aspects of a program that policymakers and decision makers at the country, regional, or local levels will want to know.

Conclusion

This article has provided a rationale and generic methodological framework for RCTs that can be applied in a scenario where students are clustered by classrooms and within schools, and that random assignment is conducted at a school level. The hope has been to provide just enough technical information to help education leaders begin to imagine how school-level RCTs can be conducted in a variety of educational contexts in different nations to assess the degree to which changes in policy, instructional practices, or programs influence learning outcomes. An important clarification, however, is that RCTs should be considered within programmatic research that relies on different types of studies both to draw out causal evidence and to understand surrounding context. Governments and quasi-government organizations increasingly need to consider these approaches as one of the many evidence-based policy methods that can be utilized to inform policy decision making. Some of the best causal evidence can be drawn from RCTs, and, hopefully, this document provides an accessible technical overview of how these designs work.
Notes

1. The term counterfactual refers to what happens in absence of a treatment (Shadish et al., 2002). When applied to program evaluation in education settings, a counterfactual could refer to ascertaining what would have happened to a group of students had a program not been introduced.

2. Groups are necessary because if there were a single student (or other unit such as a classroom or school) in one condition, and/or a single student in another condition, any observed difference cannot be confidently attributed to a treatment effect; this is because individual differentiation remains a plausible explanation. One school might, for example, simply have a better set of circumstances than another. However, if groups are used and mean differences serve as the basis for making comparisons (i.e., the statistical solution is used), then individual differences can be assumed to be evened out across groups so long as there is reason to believe that the groups are similar to each other. The basis for arguing that both groups should be similar is reviewed below.

3. It is possible that a history threat occurs wherein a new event happened to only one group. A solid understanding of context and careful observation (e.g., use of qualitative methods) can limit this worry, however.

4. There is a caveat here; typically such equivalence is followed by the phrase in expectation (Shadish et al., 2002). The issue at hand is that any instance of random assignment could yield, by chance, study groups that are not similar. The law of large numbers reduces this possibility because, as sample sizes increase, it becomes less likely that there will be large differences between groups. Furthermore, replication of studies helps remove this concern, and this is a point to be considered later.

5. One can also place confidence intervals around effect sizes (Thompson, 2007).

References


It is often ideal to conduct double-blind RCTs, where neither the study participants nor the investigators are privy to assignment conditions. This condition minimizes the chances of bias from entering the investigation. But in the context of education trials, where some new policy or teaching approach is to be applied, this is generally not possible because teachers and students will necessarily know whether new procedures are being followed. It is, however, simple enough to use independent program evaluators who have no particular fealty to a new intervention, and who are compensated on the basis of developing a scientifically defensible study, regardless of the outcomes. From this perspective, it is important to remember that a study that does not lead to the null hypothesis being rejected might serve as a protection from a push to make large scale changes to an education system in the absence of strong evidence. Under such a scheme, it is even possible for an external evaluator to work with program developers. The former could handle all assignment decisions, analyses, and reporting, and the latter can deal with implementation concerns and articulating the theory for why the new approach should yield desirable learning outcomes.

With such a model, the evaluator could at least conduct blind random assignment, whereby schools are randomly assigned a number, and random assignment is conducted with a list. It is the case that simple random assignment yields the best overall opportunity to equate groups (Schulz & Grimes, 2002), but, many times, it will be desirable to push for balanced assignment (i.e., equal sample sizes in study conditions) for political and or planning purposes. This can be undertaken by using a random allocation rule (Schulz & Grimes, 2002) or matching/stratification techniques (e.g., Boruch, 1997; Shadish et al., 2002). The former technique can be undertaken, in a semi-blind manner, via some variant of the following steps (assuming a simple, two-group RCT):
• Each participating school is labeled with a confidential identification number that will be used in the random assignment process. If there are 40 schools in the study, then simply assign these schools a number from 1 to 40. A list can be developed by a lead Principal Investigator.

• Once the list of schools and their associated identification number is confirmed, only the ID numbers will be forwarded to staff who will carry out randomization. This is undertaken so that the person who conducts the actual random assignment will not know school names when carrying out the procedure. That is, the person who conducts the random assignment will be blind with respect to school names.

• The Rand () function in Excel or a free service, such as Random.org (https://www.random.org/) can be applied. The specific service in Random.org that can be used is the “List Randomizer” (see: https://www.random.org/lists/). List randomizer allows a user to arrange a list (in this case, the school ID numbers) in random order. An analyst can enter the school ID numbers in the available field, with each ID number on a separate line. Once this is completed, a “randomize” button is to be clicked. The program will return the numerical list in random order.

• The top 50% of the schools from the list (e.g., 20 schools) will be assigned to “Group 0.” Should an odd number of schools be recruited, then one can take the 50% of schools and add one more to be included in Group 0. The remaining schools will be designated as belonging to Group 1. At this stage, all schools will have been randomly assigned to one of these two groups.

• The two groups, 0 and 1, will be entered into a new field in the list randomizer. The two groups will then be randomized, and the top group will be designated at the treatment group.

• Therefore, all schools will be randomly assigned to either Group 0 or Group 1, and one of these groups will be assigned to the treatment condition using a random process. This latter step is superfluous if we have an even number of schools, but in the event that there is an odd number of schools in the total sample, then the larger group of the two groups will be assigned to conditions using a randomization procedure.

• Note that this process should always be piloted before the real list of school IDs is formed. Unanticipated problems can and do occur and there might be concerns with clarity. Errors discovered during piloting can be remedied in advance.

• Once the actual randomization process is completed, the randomly assigned school numbers will then be shared with the lead Principal Investigator. This will complete the random assignment procedures. Oftentimes, researchers may want to force balanced allocation. This can be undertaken with standard software packages. An option is to use the list of school IDs and enter them into SPSS, a widely used statistical software package (other packages such as SAS and Stata offer similar routines).

Enter each School ID as a variable, and then run the following syntax:

```
USE ALL.
COMPUTE filter_$(uniform(1)<=.50).
VARIABLE LABELS filter_$( ’Approximately 50% of the cases (SAMPLE).’
FORMATS filter_$( f1.0).
FILTER BY filter_$. EXECUTE.
USE ALL.
COMPUTE filter_$(uniform(1)<=.50).
VARIABLE LABELS filter_$( ’Approximately 50% of the cases (SAMPLE).’
FORMATS filter_$( f1.0).
FILTER BY filter_$. EXECUTE.
```

Using the windows procedure: Click on Data → “Select Cases → Highlight the School ID Variable → Random Sample of Cases (Approximately 50% of cases) → OK.

This will create a new filter variable, approximately one half of the cases will be assigned a 1, and the other one half a 0. These groups then can be randomized to one of two study conditions.

Finally, schools could be paired on some matching variable that is correlated with the learning outcome variable (Shadish et al., 2002). This might be especially helpful if dealing with a relatively small number of units and if there is considerable heterogeneity among them. Although simple random assignment would still technically yield a fair trial, there would in these circumstances be a decent chance that analysts would have to handle baseline differences between study groups and evaluators might deal with face validity concerns if, for example, it were known that there were more high-achieving schools in one study than another study (which can happen in spite of random assignment). In the cases of small sample sizes and sample heterogeneity, then
one might be better off with a simple matching scheme. For example, suppose that one can rank schools on some indicator of aggregate student achievement. If so, a paired matching scheme could look like:

- Top Achieving school (1)
- Second top achieving school (2)
- Randomly assign one of these two schools to the treatment, one to the control ...
- Third top achieving school (3)
- Fourth top achieving school (4)
- Randomize this pair ...
- Repeat ...
- School (39)
- School (40)

Randomize this final pair. This scheme would yield balanced in and prevent any baseline differences between schools with respect to average achievement. This would simplify later aspects of the RCT.