Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 15, December 2007 ISSN 1531-7714

Randomized Field Trials and Internal Validity: Not So Fast My Friend

James H. McMillan, *Virginia Commonwealth University*

The purpose of this article is to summarize eight potential threats to internal validity that occur with randomized field trial (RFT) studies. Depending on specific contextual factors, RFTs do not necessarily result in strong internal validity. Of particular concern is whether the unit of random assignment is the same as the number of replications of the intervention, threats as a result of local history, and subject effects. The eight threats are described, with suggestions for providing adequate monitoring to know if they rise to the level of likely or plausible threat to internal validity.

Educational reform, programs and practice are now being evaluated with scientifically based research. The "gold standard" for generating rigorous evidence is the randomized (true) experiment, namely Randomized Control Trials (RCT) or Randomized Field Trials (RFT), or, at the very least, on quasi-experiments in which there is "equating" of pretest differences (National Research Council, 2004). The emphasis is on determining the causal link between interventions, such as programs, curricula, or materials, and student performance. It is the criteria used for determining whether studies evaluated by the What Works Clearinghouse meet evidence standards, and how the research designs of federally funded programs are evaluated. However, this emphasis on doing randomized experiments may be misleading unless attention is paid to three important conditions. The first is being sure that the design actually accomplishes the reason for using random assignment – to achieve statistical equivalence of the experimental and control group prior to, during, and after the intervention is implemented. The second is the need to evaluate internal validity on the basis of many factors that are common in field studies. Third, determining causality, which is why experiments are conducted, is heavily dependent on contextual factors peculiar to each study.

It is a tribute to Don Campbell and Julian Stanley that their seminal publication *Experimental and Quasi-Experimental Designs* (1963) has had such staying power. In particular, their eight internal threats to validity, along with their labels, continue to be the ones still emphasized in educational research textbooks (some now list a few more, such as experimenter effect or diffusion of treatment). In addition, most texts describe randomized designs or true experimental designs as ones that "control" for these threats to validity, implying that if they are controlled they are no longer threats to internal validity. However, consistent with Cook and Campbell (1979), Schneider, Canroy, Kilpatrick, Schmidt, and Shavelson (2007), Shadish, Cook, and Campbell (2002), this is clearly *not* the case in field studies. Consequently, it is important for researchers to understand that causality in randomized experimental studies in the field is often difficult to determine. Certainly, when random assignment is used to place participants into interventions, resulting in a "randomized trial" it does not absolve the researcher of the responsibility to consider appropriate threats to internal validity, including selection bias if the randomization is not adequate to statistically "equate" the intervention and control group. Indeed, it can be argued that RFT's have many more potential threats to internal validity than would highly controlled quasi-experiments.

This article focuses on the so called "gold" standard, RFTs, with implications for quasi-experiments. It will be demonstrated that simply calling a study "randomized" does not mean that there are likely to be few, if any, plausible threats to internal validity. Quite the contrary, field experiments, whether randomized or not, have a multitude of possible threats to internal validity. Random assignment helps in arguing that some threats are controlled, but depending on the nature of the experiment, many possible threats remain. Eight possible threats are considered here – there are more that could be included (McMillan, 2000). These are not controlled in RFTs, and as such, may constitute plausible rival hypotheses that explain outcomes.

Unit of Randomization and Local History

A study can be labeled as a "randomized experiment" if there is *random assignment* of subjects to intervention and control groups (and/or different interventions). The reason that random assignment is so important is that, if carried out properly, it results in comparison groups that are statistically equivalent in every way possible except for the intervention. The contention is to assure that observed differences on the dependent variable are not due to differences between the groups, most importantly ruling out the threat of selection bias. It helps ensure that confounding variables are not systematically related to an intervention or control group, making alternative explanations unlikely.

What needs special attention is the phrase *if carried out properly*. Random assignment is a means to an end – the ability to assume statistical equivalence of the groups prior to the pretest or intervention so that any potentially confounding or extraneous variables are the same for each group. There must be a sufficient number of units to be randomized to achieve this end, as well as procedures that replicate the intervention for each subject independent from other subjects. Randomly assigning four intact classes to two interventions will not achieve this goal. As obvious as this seems, there are many instances when this procedure is used with a claim that there has been random assignment, and that selection threats are controlled. On the other hand, if randomizing a group of 40 homogeneous fifth graders to two interventions, statistical and confounding variable equivalence would probably be achieved, as long as the interventions were administered individually for each subject. Most actual field research situations are somewhere between these extremes. At best, RFTs control many possible threats, but not all. At worst, relying too much on RFTs without appropriate consideration of all possible threats to internal validity will result in misleading conclusions about program effectiveness (Chatterji, 2007).

A further word is needed about unit of analysis and how interventions are administered, and unit of analysis. Ideally, there is an independent replication of the treatment for each subject if individuals are used to determine total *n*. Take as an example the effect of viewing a video tape on student attitudes. The preferred procedure would be having each student view the videotape alone, so that the intervention is essentially replicated over and over. This procedure is what helps establish the high probability that possible confounding variables are not plausible explanations of the results. Contrast this procedure with a more typical approach – random assignment of students to two groups, with the videotape played for students as they sit together. In the latter method "random assignment" of subjects is used but each intervention is replicated only once. This is problematic because of the strong probability that confounding variables associated with one of the classes would affect the results (e.g., teacher, group dynamics, participant dependencies, unforeseen events, disruptions, etc.). That is, students within each group are exposed to common influences in addition to the intervention. There is simply no way to control confounding variables of this nature in field settings (e.g., students getting sick, disruptions, teacher fatigue, emergencies, etc.). It is essential to monitor implementation of the intervention to rule out such threats. The What Works Clearinghouse has used this principle in classifying many studies as "does not meet evidence screens:"

> *There was only one intervention and/or comparison unit, so the analysis could not separate the effects of the intervention from other factors.*

The issues concerning the appropriate statistical unit of analysis have been discussed for years (Shadish, et al.). The strongest design, from an interval validity perspective, is achieved when the unit of analysis is equivalent with the number of independent replications of the intervention. This suggests that researchers should use intervention delivery modes that are consistent with what is used in practice, and then use the number of treatment replications of delivery as the unit of analysis (McMillan, 1999). If the intervention is delivered by the teacher to the class as a whole, then classroom would be the appropriate unit of analysis. If the intervention is done individually with students, such as testing a computer simulation, the unit would be determined by the number of students in the study. If the intervention is at the school level, such as a study of the effect of a new procedure to discipline

students, then school would be the appropriate unit of analysis.

There are statistical tools to address the unit of analysis problem, such as hierarchical linear modeling (HLM), but a key component of the success of these procedures is having a sufficient number of "higher" or "cluster" units. At issue is whether unique *random* effects for each unit, those incorporated in HLM, control *nonrandom* confounding variables associated with particular units. Obtaining enough higher units, such as schools or classrooms, has obvious drawbacks related to the scope and expense of the study. When sufficient resources are not available, researchers would be well advised to treat the study as quasi-experimental, using techniques to help control the effects of confounding variables. Shadish et al., for example, suggest switching replications or using multiple pretests.

Intervention (Treatment) Fidelity

In an ideal experiment within-intervention variation is minimal. One of the most troublesome difficulties in field studies, however, is that invariably each replication of the intervention is not exactly like other replications. It is simply not realistic to assume that interventions are standardized, even if there is a detailed protocol and experimenters do not make mistakes in the intervention. As pointed out by Shadish et al., fidelity of the intervention is often compromised for several reasons: 1) when intervention specifics do not correctly reflect theory, 2) when there is inadequate check on the implementation of the intervention, and 3) when there is no indication of between-group differences in what is implemented.

Essentially, in field experiments, the independent variable is the intervention-*as-implemented*. The actual nature of the intervention needs to be monitored and documented to obtain accurate causal conclusions. This can be accomplished through interviews or self-reports of subjects, observations, and third party reports about what occurred. Consider testing the efficacy of using targeted formative assessment strategies, such as giving students specific and individualized feedback, on student motivation. Two groups of teachers are utilized – one group attends workshops and receives materials about providing feedback, the other group acts as a control. We can be sure that each experimental teacher will not come up with the same feedback or give it to students in the same way, even if they attended the same workshops and received the same materials. To fully understand what was responsible for causing change in student motivation, then, it is necessary to know what differences occurred in the implementations of the intervention. If there is no evidence about intervention fidelity we are less sure that the differences observed are consistent with theory and operational definitions.

As noted above, intervention fidelity is also important in determining whether there were any additional events or occurrences during the study confounded with treatment. This is why it is important for experimenters to become, in the words of Tom Cook (2006), "anthropologists of their study." There is a need to monitor very carefully and have a full understanding of intervention fidelity.

Differential Attrition (Mortality)

When subjects in the intervention group drop out of a study after random assignment at rates that are different from subjects in a control or comparison group, it is likely that such treatment-correlated attrition will be confounded in unknown ways (Shadish et al.; West & Sagarin, 2000). This is a problem when subjects literally leave an intervention, fail to participate in some intervention activities, or fail to complete dependent variable measures. In essence, substantial differential attrition results in a quasi rather than true experiment because the groups become unequal, even if randomly assigned in the beginning. If there is such attrition, it is important to explore the reason for the loss of subjects and analyze how that affects the results. Tracking participants can and should be used in field experiments to minimize the threat of differential attrition by determining if the attrition is random or systematic. If it seems that there may be bias associated with differential attrition, characteristics of subjects who have dropped out of both intervention and control groups can be compared. It is also helpful to determine the reason why subjects have not completed the intervention and/or taken the posttest.

Instrumentation

There are many ways in which weaknesses in how data are collected can adversely affect the internal validity of an RFT, none of which are necessarily controlled by random assignment. The concern is whether something in the way data are gathered differentially impacts the results in either the experimental and control groups. This could occur with observer, rater, or recorder error or bias, ceiling and floor effects, and in changing measures with single group longitudinal studies. Essentially, there is measurement bias when subject responses are influenced or determined by variations in the instrument(s) and/or procedures for gathering data. An obvious example is if the experimental group has one observer and the control group a different observer, or when the experimental group presentations are rated by one person and the control group rated by a

different person. In these instances, the unique effect of the observer or rater is problematic. Strong measures of evidence for reliability based on agreement of scorers prior to implementing the experiment does not rule out this threat. These reliability coefficients underestimate the degree of difference. Of course, it is better to have such evidence of reliability in the scores than not have it.

Another kind of threat occurs if the observer, rater, or recorder knows which individuals are the treatment subjects and which are in the control group. It is essential to keep observers "blind" with respect to knowledge of assigned group. If that condition is not maintained, scoring is likely to be affected by observer or rater expectations. A similar threat is if the scorer or rater knows which measures are the pretest and which ones are the posttest. If the design is pretest-posttest, whether or not subjects are randomly assigned, the measures should be coded for whether they are pre or post measures and for experimental or control subjects, but knowledge of the code or group assignment must not be known. The rater or scorer should simply be given a "stack" of tests or reports.

Diffusion of Intervention (Treatment)

An important principle of good randomized studies is that the intervention and control groups are completely independent, without any effect on each other. This condition is often problematic in field research. When the effect of the intervention spreads to the control group, or when the control group knows about the intervention, behavior and responses can be initiated that otherwise would not have occurred. Sometimes subjects affected by an intervention interact with control subjects because they are in close proximity, such as friends in treatment and control groups, by being in the same school or neighborhood, or by being in the same class. In this circumstance the changes caused by the intervention are diffused to the control subjects through their interaction with each other. Trochim (2005) refers to these as social interaction

When control subjects realize that intervention subjects have received something "special" they may react by initiating behavior to obtain the same outcomes as the intervention group (compensatory rivalry) or may be resentful and be less motivated (resentful demoralization). In some circumstances teachers or parents of control subjects may provide additional activities to match the affect of the intervention (compensatory equalization).

Diffusion can be prevented by isolating intervention and control subjects, but when this is done in education (e.g., intervention in one school; control group in another school), other threats to internal validity increase in plausibility and would often prevent randomization. Tracking of subjects and asking appropriate questions about the possible influence due to diffusion helps limit the plausibility of this treat internal validity.

Subject Effects

There are a variety of behaviors that can be attributed to the subjects because of the sampling and procedures used in an experiment. These include compensatory rivalry and equalization, resentful demoralization, Hawthorne effect, demand characteristics, social desirability, and subjects wanting to please the experimenter. Most of these factors dilute the effect of the treatment and make it more difficult to show differences. Others, like resentful demoralization, inflate true differences. These threats occur quite independently from whether or not there is random assignment of units, and are especially troublesome when it is clear to subjects that a specific outcome is desired. This tends to occur in experiments in which the goal of the study is to demonstrate that an intervention is effective, such as what occurs in some research conducted to determine the efficacy of specific curricula or program. Compensatory rivalry and equalization, which masks the effect of the treatment, is problematic if the control group has information about or observes the treatment, or thinks of itself as an "underdog." The subjects will be motivated to perform like the treatment group. This often occurs when the experimental unit of interest is departments and groups. In a school, for instance, compensatory rivalry could be a factor in a study that assigns treatment and control groups by class within the same school.

Experimenter Effects

All too often researchers try to prove a point rather than taking the perspective of an unbiased, objective investigator. This leads to experimenter effects that are both deliberate (bias) and unintentional. Researchers have attitudes, values, biases and needs that may contaminate the study. Obviously, if researchers have a vested interest in the study there is motivation to find results that will enhance their position. This is a common circumstance in instances where there is a need to show that specific interventions have a positive impact on student learning. Since experimenters typically have research hypotheses about the outcomes, it is important to include procedures that minimize the plausibility that these effects could constitute rival hypotheses. Unintended and unconscious effects, as well as those that are obvious, may affect the results by giving preferential treatment to the intervention that is being tested, even when the researcher is careful. This could occur in the nature of experimenter-subject

interaction, such as the use of a more positive voice tone, by displaying different attitudes, or by being more reassuring with the experimental group.

Experimenter effects are most probable in studies where the researcher is the one gathering data, administering the instrument, and/or carrying out the intervention. In these circumstances specific procedures should be included in the study to limit the plausibility of experimenter effects. For example, the researcher can develop a specific protocol for gathering data and administering the intervention.

Novelty Effect

When a new intervention is introduced or there is some change to normal or established routines that are new or novel, subjects can be motivated or respond differently simply because of the change. For example, if students are accustom to direct instruction with worksheets, an intervention using small groups may motivate positive behavior and desirable responses because it is novel and different. This threat is most plausible when the change results in students being more engaged, social, and stimulated. A change to a more negative practice may have the opposite effect.

Summary

The intent of this article is to show that there are *always* some possible internal validity threats to randomized field experiments, and often some of these are plausible. Sometimes the threats are fatal flaws, rendering the findings useless. While randomization is effective in controlling many threats to internal validity, such as selection (keeping unit of randomization and unit of analysis the same), other possible threats need to be considered to determine if they rise to the level of plausibility. If plausible, these threats compromise causal interpretations. In some cases results are simply uninterruptible, as is the case with many studies reviewed by the What Works Clearinghouse. It is the responsibility of researchers to identify possible threats and then include design features that will gather information to lessen the probability that the threat is plausible (see Reichardt, (2000), for a typology of strategies for ruling out threats to experimental validity, and Chatterji (2007), for a review of "grades of evidence to use with RFTs). It is also the responsibility of consumers of research to know how to look for plausible threats to determine for themselves the credibility of the findings. For the most part, when experiments are conducted in naturally occurring places like schools, looking for threats to internal validity takes on special requirements, whether or not the study used random assignment.

References

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research on teaching.* Chicago: Rand McNally.

Chatterji, M. (2007). Grades of evidence: Variability in quality of findings in effectiveness studies of complex field interventions. *American Journal of Evaluation, 28*(3), 239-255.

Cook, T. D. (2006, June). Workshop on Quasi-Experimental Design and Analysis in Education. Evanston, IL: Northwestern University.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand-McNally.

McMillan, J.H. (1999). *Unit of analysis in field experiments: Some design considerations for educational researchers.* ERIC Document Reproduction Service No. ED428135.

McMillan, J.H. (2000. April). *Examining categories of rival hypotheses for educational research.* Paper presented at the annual meeting American Educational Research Association, New Orleans.

National Research Council (2004). Implementing randomized field trials in education: Report of a workshop. Committee on Research in Education. L. Towne and M. Hilton (Eds). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Reichardt, C.S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.). *Research design: Don Campbell's Legacy.* Thousand Oaks, CA: Sage Publications, Inc.

Schneider, B., Carnoy, M., Kiopatrick, J., Schmidt, W.H., & Shavelson, R.J. (2007). *Estimating causal effects using experimental and observational designs: A think tank white paper.* Washington, D.C.: American Educational Research Association.

Shadish, W. R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

West, S. G., & Sagarin, B. J. (2000). Participant selection and loss in randomized experiments. In L. Bickman (Ed.), *Research design: Don Campbell's legacy.* Thousand Oaks, CA: Sage Publications, Inc.

Citation

McMillan, James H. (2007). Randomized Field Trials and Internal Validity: Not So Fast My Friend. *Practical Assessment Research & Evaluation*, 12(15). Available online: http://pareonline.net/getvn.asp?v=12&n=15

Author

James H. McMillan
Virginia Commonwealth University
Jhmcmill [at] vcu.edu