Abstract Title Page

Title: External validity in the context of RCTs: Lessons from the causal explanatory tradition

Author(s): Andrew Jaciw, Denis Newman

Abstract Body

Background / Context:
Randomized trials give us a powerful methodology for assessing the efficacy and effectiveness of programs in education. Randomizing cases to conditions results in statistically equivalent groups and, barring undesirable effects of attrition, yields unbiased impact estimates (Boruch et al., 2002; Cook, 2002; Cook and Payne, 2002; NRC, 2002; Riecken and Boruch, 1974; Shadish, Cook, and Campbell, 2002).

RCTs are considered to be the gold standard method for establishing the internal validity of causal inferences; however, their use has been criticized on several grounds (Berliner, 2002; Cronbach, 1982; Cronbach et al., 1980; Erickson and Gutierrez, 2002; Phillips, 2006). An often mentioned critique has to do with the role of context and limitations in the generalizability of the results. The RCT is optimal for establishing internal validity, but this does not assure other kinds of validity. We seldom limit our interest to an internally valid result as it held in some setting in the past, rather, we want to know if the result will replicate in new and different settings. While much work has been done establishing standards for assuring internal validity the same has not been done for external validity.

There are several approaches to establishing external validity. Formal random sampling followed by random assignment is a preferred approach but is seldom feasible (Cook, 2002). Purposive sampling of heterogeneous instances underlies the heterogeneity for replication method for establishing generalized causal inferences (Shadish, Cook and Campbell, 2002). Here a generalized effect is established by showing that is robust across many settings and conditions. The causal explanatory approach (Cronbach, 1975; Cronbach, 1982; Cronbach et al., 1980) focuses the mechanisms underlying the interactions of treatment with contextual variables. The goal is to comprehensively account for the conditions under which effects of a program are likely to replicate thereby informing the general picture of a program and its effects.

Purpose / Objective / Research Question / Focus of Study:
The purpose of the current work is to apply several main principles of the causal explanatory approach for establishing external validity to the experimental arena. By spanning the paradigm of the experimental approach and the school of program evaluation founded by Lee Cronbach and colleagues, we address the question of how research programs that involve experiments can be expanded to make external validity more of a priority. We bring to bear three central concerns of the causal explanatory approach on the activity of conducting randomized trials with a view to establishing external validity: (1) the role of interactions, (2) the need for ecologically relevant generalizations (3) the time-dependency of generalized causal inferences[*].

This work is written with the SREE conference theme in mind: *Building an Education Science: Investigating Mechanisms*. The conference provides an opportunity to re-visit the tradition of program evaluation founded on the ideas of Lee Cronbach. It is described as 'causal explanatory' because it focuses on explanations through mechanisms as a basis for understanding not just 'what works' but for whom and in which settings - that is, to establish the reach of causal effects.

[*] Given the brevity of this structured abstract we will consider here only the first two concerns, and use a case study to illustrate the first only.

Significance / Novelty of study[†]:
There is little work that we know of that considers how external validity should be established with randomized trials or the implications of the method for external validity. The exception is the work by Shadish, Cook and Campbell (2002) who advocate an approach that utilizes meta-analysis to combine effects estimates into global causal propositions (i.e., the heterogeneity of replication approach to establishing generalized causal inferences.) We do not disagree with this approach but we believe that the causal explanatory approach to program evaluation can give new insight into what it means to establish external validity through experiments, and investigation of the basic premises of the causal explanatory approach may serve to eventually increase the informational yield from RCTs in a way that strengthens the external validity of results.

A basic practical interest undergirds this work: how to shore up research programs, from development to scale up, which utilize RCTs, to ensure that the results inform us not just about average effects under experimental conditions, but about the variation and reach of effects in application. (A primary purpose of this work is to stimulate discussion among SREE members and conference attendees about these issues at a time when the focus on federally-funded research is on producing results that are not only internally valid but also timely and locally relevant. We see the use of an electronic poster medium with two way dialogue with members of the research community as a first step towards a possible symposium involving this topic at a later SREE conference.)

Description of Method[‡]:
This work consists of a novel application of the core ideas of the causal explanatory approach to program evaluation to the problem of establishing generalized casual inferences from RCTs.

This work is structured as follows: We consider three ideas critical to the causal explanatory approach to establishing generalized causal inferences (for the structured abstract we consider only the first two.) Each is followed by a brief commentary discussing the implications for establishing external validity through experiments, and a case study for illustration (for the structured abstract we present only a single case study, in brief.) We conclude the work by summarizing the applications of the causal explanatory approach to the problem of drawing generalized causal inferences through the conduct of randomized trials.

Lessons from the Causal Explanatory Approach to Establishing External Validity

The Importance of Interactions:

The causal explanatory perspective: Interactions are critical for understanding the generalizability of findings. Cronbach questions that 'gross experimental comparison can produce useful rules for schooling, where a treatment is multifaceted, cannot be standardized, and interacts with pupil background' (1975, 122). Under these circumstances, differences among

[†] According to the SREE structured abstract for a Methods submission, the following subheadings may not be applicable: Setting, Population / Participants / Subjects, Intervention / Program / Practice, Research Design, Data Collection and Analysis. We do not include these hradings. The current work is an analytic essay where we use several case studies and a survey of critical points from a particular tradition in program evaluation to discuss challenges to establishing external validity using RCTs.

[‡] This is an analytic essay supported by case studies, therefore we do not present a statistical, measurement or econometric model.

schools in how they implement treatment and contextual differences that lead to variations in treatment 'swamp out' a generalized effect such as the average effect of treatment. Accordingly, "a general statement can be highly accurate only if it specifies interactive effects that it takes a large amount of data to pin down" (Cronbach, 1975, p. 126). External validity is an exercise in accounting for the effects of interactions of contextual factors with treatment (though Cronbach would regard an experimental effect not as the result of a manipulation of 'the treatment'; rather, any impact is the joint influence of multiple treatments, only some of which we have control over (through randomly assignment, for instance.))

With the causal explanatory approach one can imagine a record that contains detailed information about how a certain kind of program interacted with various local factors across different settings in the past. This diversified account would serve generalization by listing the things that have mattered in the past, and identifying and unifying them through explanatory theory, to form a basis for making predictions for new settings in the future.

Commentary: The role of moderator effects in RCTs

In the case of RCTs there is limited opportunity to explore interactions (Cook, 2002)[§]. We assert that the quality of moderators matters: moderators that support causal explanation will either go further in accounting for variations in the impact that are observed than if we merely examine the interactions of treatment with 'convenience variables' – demographics that we analyze in terms of their interactions with treatment only because they are available – or, if they do not, this provides a basis for challenging the theory that is the basis for selecting the moderators. (In the case study below, we show that 'convenience variables' have little power to predict variations in impact in a multisite trial.) Starting at the program development stage, by considering the mechanism by which intervention effects are either intensified or suppressed through their interactions with context, we can increase the informational yield of subsequent experiments and lead to refinement of theory by putting to the test moderators theorized to account for variations in impact. Consistent with the causal explanatory approach, we assert that by examining moderating effects of variables that have a theoretical foundation, the general picture will emerge as we substantiate the theory on which the interaction with treatment are predicated.

Case Study 1:

We use results from the Tennessee Class Size reduction experiment (Project STAR) to illustrate our point that 'convenience covariates' do little to elaborate the general picture, and why there is a need to investigate theory-based interactions if the goal is to establish the full reach of the findings. (The details of this experiment are provided in Finn and Achilles (1990) and Mosteller (1995). Students were randomized in kindergarten to small classes, regular classes, or regular classes with an aide. The experiment lasted for four years. Teachers were also randomized to classes. The outcome measures were scale scores in reading and math. A main finding is that by the end of the second year, students in small classes had close to a .20 standard deviation advantage in achievement over students in regular classes or regular classes with an aide.)

The STAR experiment showed variation in impact across sites. We can regard this variation as indicating that the finding of an average effect does not generalize – information about the

[§] Work by Bloom (2005) indicates that the opportunity may be greater than we think, at least in one sense: power for detecting moderator effects may be greater than for average effects of the same size in group randomized trials in cases where the moderator is identified below the level of randomizations (for example, if we are interested in differences in impact betweens subgroups of students where schools have been randomized.)

effects of small classes at one set of sites, or averaged across all sites, does not allow us to reliably predict what the impact would be at a different set of sites. Moderators can inform the general picture by identifying the conditions under which we can expect impacts to differ. We analyzed the STAR data to see if moderator effects account for cross-site variation in impact. STAR data have available only 'convenience variables' – simple demographics not theoretically tied to the intervention – that can serve as moderators. Figure 1 (see Appendix B) displays the effect of modeling the main and interactive effects of one or more moderators with treatment on the between-school variation in the treatment effect for the reading outcome. We see that the basic demographics account for between-school differences in the average effect, but not in the treatment effect (modeling the covariates shifts the points leftward, but not downward.) In other words, they do little to inform us about the general picture concerning the effects of the intervention and why they vary across schools.

The STAR experiment is an instance where moderators grounded in causal explanatory theory could have gone much further in establishing the external validity of the findings. The incapacity of the all-purpose covariates to account for variance in the impact sets a limit to conclusions about generalizability. Cohen, Raudenbush and Ball (2002) stress the importance of having valid measures of resources for understanding the effects of programs such as class size reduction. Similarly, we believe that generalizability would be served by obtaining measures of school capacities and practices that, based on causal explanatory theory, provide the rationale for *why* a certain program has desired effects in some but not all situations.

The need for ecological relevance:
The causal explanatory perspective: Establishing generalizations in the social sciences is difficult because we deal with open systems (as opposed to isolated systems where the sufficient conditions for a phenomenon may be few) (Cronbach, 1982). RCTs are normally conducted within open systems. It is misleading to think that controls over the design can help us work up to a closed-like system: a finding observed under highly controlled circumstances may have greater reproducibility but the contrived arrangement may give results that have little relevance to the real world. A feature of the openness of systems is that time interacts with treatment. Generalizations in the social sciences are empirical in nature, they don't always hold true, and can be considered working hypotheses that are subject to different degrees of challenge as exceptions to the rule are found (Cronbach, 1982). Generalizations decay. 'The half-life of an empirical proposition may be great or small. The more open a system, the shorter the half-life of relations within it are likely to be' (Cronbach,1975, p. 123).

Commentary: The scalability and timeliness of experimental findings
Results that are non-timely or that are limited by the study context have limited relevance. With a randomized trial we can ask: which universe are we generalizing to: a universe of studies or situations of actual implementation? The limited context of a randomized trial or even a set of such trials may not address the full consequences that happen when the program being tested is introduced at scale. A limited trial such as the Tennessee STAR class size reduction experiment may only scratch the surface of what the effects of this initiative are, once scaled-up and introduced at the levels of, for instance, a state. (In California, implementation of reduced class size in K-3 across the state had unintended consequences (e.g., certain recruitment practices) that may have counteracted the desired impact (Cook, 2002). Here no sum of small scale trials can

give the full picture as concerns the generalizability of effects when the intervention is introduced 'in vivo'.)

Conversely, certain RCTs may be conducted on large scales over relatively long periods (consider some of the RCTs conducted through NCEE which involve randomization of as many as 40 – 80 schools and thousands of students) so that the trials cannot be seen merely as test runs. In this context generalizability takes on a different interpretation – it is not about gathering many pieces of evidence to inform the long-term picture. The intervention is sufficiently in place *now* and over a broad enough sample that the question about the variation and sustainability of effects concerns the study sample. This is especially pertinent if not all subjects are receiving the same benefit: monitoring the impact through intermediate analyses and identifying stopping rules (approaches borrowed from medicine) in the case of negative impacts is critical if the trial is large and long-term. Whether an intervention works, for whom it works, and the boundary conditions for its effects – all factors that have a bearing on external validity – are not a postscript to the intervention but are immediate occurrences with ongoing consequences.

(In the case study addressing this point we will consider a specific program of short-term small-scale trials intended to provide timely results to inform local program adoptions (Newman, 2008). With quick-turnaround being essential, and with programs expected to interact with local conditions and characteristics of subjects, external validity becomes about accounting for a range of effects in the present, rather than accumulating and compiling results over time[**].)

Usefulness / Applicability of Method and Conclusion:
In this work we applied some of the tenets of the causal explanatory tradition of program evaluation to address the question: what does it mean to establish external validity in the context of RCTs? The causal explanatory research tradition gives valuable insight into how research programs (from development to scale up) that include RCTs can be supplemented to strengthen external validity. Specifically we have argued that we cannot separate the issue of external validity as pertaining to RCTs from the nature of the design, the scale of the intervention and the timeliness of the result. External validity as something to be achieved in the future once a critical mass of internally valid results are accumulated is only one rendering of this form of validity and it may not be appropriate in all cases. We have also argued for investing in the collection of measures of contextual factors that are theorized to interact with the treatment process and account for variation in the effect under study. Accounting for treatment heterogeneity establishes external validity by addressing the conditions under which a treatment works better or worse. Not doing so seems is a lost opportunity. This calls for having a working theory of how and for whom a program works, which should be addressed early in the program's development. The current work (summarized briefly here) is intended to inform the standards for establishing external validity of program effects in education that are evaluated through randomized trials.

[**] One can also imagine a way of determining the bounds for an effect that combines cumulative evidence from past studies with current and local results, where each piece of evidence is weighted in some optimal way.

Appendices

Appendix A. References

Berliner, D. C., (2002). Educational research: The hardest science of all. *Educational Researcher, 31,* 18-20.

Bloom, H. S., (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning More From Social Experiments*. New York: Russell Sage Foundation.

Boruch, R., de Moya, D., & Snyder, B., (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in educational research.* Washington, DC: Brookings Institution Press.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L., (2002). Resources, instruction and research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in educational research.* Washington, DC: Brookings Institution Press.

Cook, T. D., (2002). Randomized experiments in educational policy research: A critical examination of the reasons the education evaluation community has offered for not doing them, *Educational Evaluation and Policy Analysis, 24,* 175-199.Cook and Payne, 2002;

Cook, T. D., & Payne, M. R., (2002). Objecting to the objections to using randomized assignment in educational research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in educational research.* Washington, DC: Brookings Institution Press.

Cronbach, L. J., (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 116-127.

Cronbach, L. J., (1982). *Designing evaluations of educational and social programs.* San Francisco, CA: Jossey-Bass.

Cronbach, L. J., et al., (1980). *Towards reform of program evaluation.* San Francisco, CA: Jossey-Bass.

Erickson, F., & Gutierrez, K., (2002). Culture rigor and science in educational research. *Educational Researcher, 31,* 21-24.

Finn, J. D., & Achilles, C. M., (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557-577.

Mosteller, F., (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*, 113-127.

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

Newman, D. (2008). Toward school districts conducting their own rigorous program evaluations: Final report on the "Low Cost Experiments to Support Local School District Decisions" project. Palo Alto, CA: Empirical Education Inc.

Phillips, D. C., (2006). A guide for the perplexed. Scientific educational research, methodolatry, and the gold versus platinum standards. *Educational Research Review (1), 15-26.*

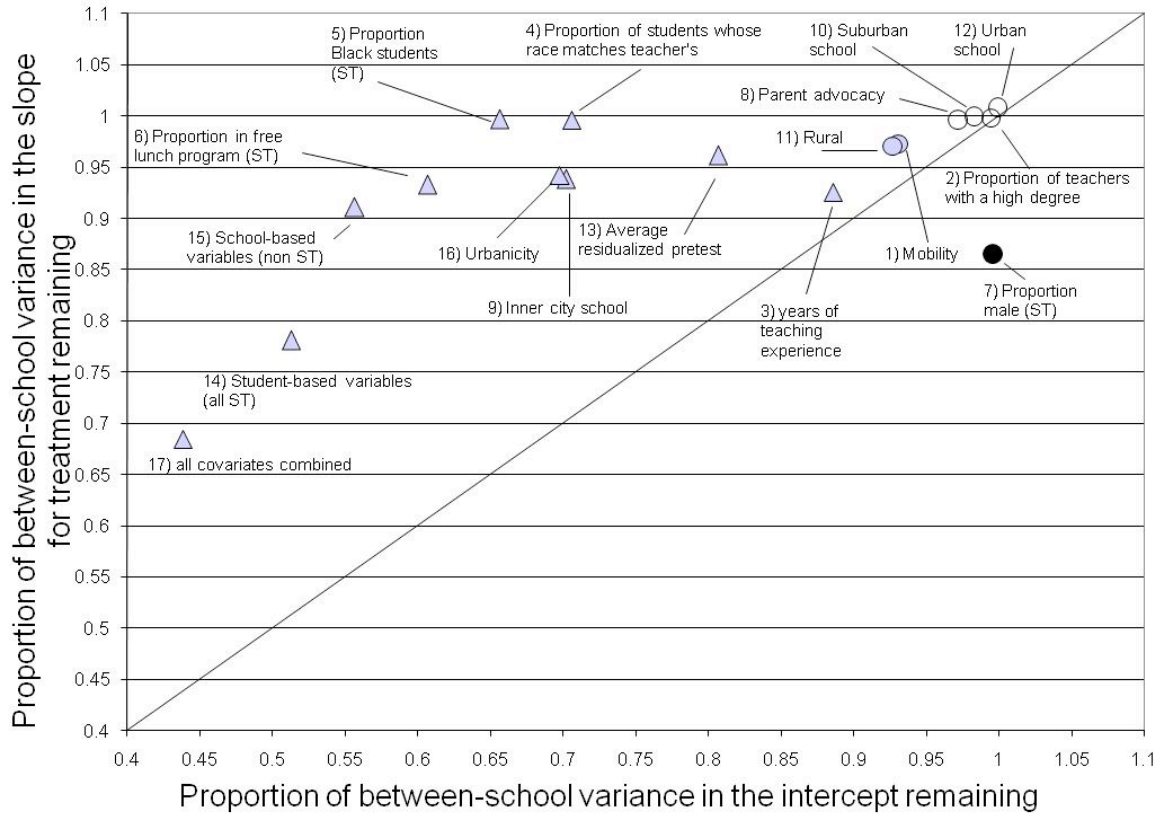
Raudenbush, S. W., & Bryk, A. S., (2002). *Hierarchical Linear Models (2nd ed).*. Thousand Oaks, CA: Sage.

Riecken, H. W., & Boruch, R. F., (1974). Why and when to experiment. In H. W. Riecken, R. F. Boruch, D. T. Campbell, et al. (Eds.), *Social experimentation: A method for planning and evaluating social interventions.* New York: Academic Press.

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

Appendix B. Tables and Figures

Figure 1: Reading outcome: Proportion of between-school variance remaining after modeling main effects of covariates and their interactions with treatment



Note: A gray marker indicates that the main effect(s) of the covariate(s) is/are statistically significant ($p<.05$); A black marker indicates that the interaction(s) between the covariate(s) and treatment is/are significant ($p<.05$). An empty marker indicates that neither of these conditions hold. A triangle indicates that the model that includes both the main and interactions effect(s) *results in a better fit* than the reference model (i.e., the model without any school-level main or interaction effects.) A circle indicates that the model that includes both the main and interactions effect(s) *does not result in a better fit* than the reference model.