Abstract Title Page

Title: The Use of Moderator Effects for Drawing Generalized Causal Inferences

Author: Andrew Jaciw

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 (HR) 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 (CE) approach (Cronbach, 1975; Cronbach, 198; Cronbach et al., 1980) focuses on 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:

This work has a three-fold purpose. The first is to set out the rationale for an alternative approach for assessing the generalizability of research findings. (It is related to but not equivalent to the CE method.) The second is to formalize this approach through a quantitative model. The third is to apply the approach to the results of the Tennessee STAR multisite trial of the effects of class size reduction to investigate (1) the generalizability of the average effect estimate across sites, and (2) the power of 'convenience variables' (basic demographic variable) to strengthen the external validity of the findings.

Significance / Novelty of study:

Prior efforts to establish systematic approaches to addressing the problem of generalizability in educational research have separated into two camps: CE puts external validity on par with internal validity and sees grand average impact estimates from experiments as having little relevance for informing the general picture. HR considers the grand average from meta-analysis to be the generalized outcome; however, heterogeneity in the effect does not allow combining impact estimates and so stands in the way of generalizability. The approach we propose exploits

[*] We regroup some of the headings for the structured abstract because not all are applicable and it makes sense to combine some of them given that this work has both theoretical and empirical results.

the benefits of each of these approaches: Like HR in continues to regard results of experiments as a basis for internal and external validity; but like CE it regards interactions of context with treatment as informing the general picture.

Statistical, Measurement, or Econometric Model <u>also</u> Findings / Results::
Results 1: The rationale for an approach for assessing the generalizability of results of an RCT

Our approach has as its starting point the idea that knowledge of impacts for subgroups can help to generalize the findings of an average program effect measured at one set of sites (where the program has been used) to another set of site (where the program has not been used.) If certain covariates are found to moderate the impact at the sites where the program is being implemented, then knowledge of these effects can increase the accuracy of the prediction of the impact at other sites (where administrators may want to know what would happen if the program had been (or is) introduced.)

If a program has a constant treatment effect across sites then the impact at one set of sites likely generalizes to another set of sites. However, if a program effect varies across sites then this variation indicates that the program effect does not generalize across sites – for example, information about the effects of class size reduction 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. In that case, estimates of moderating effects of covariates can inform the general picture by identifying the conditions under which we can expect impacts to differ. For example, knowing that benefits of small classes are greatest in small schools with English proficient populations of students gives us some basis for predicting whether a new school is likely to experience a benefit (i.e., we would be interested in whether the school is small and has many English proficient students[†].) Our main thesis is that moderator effects give us a basis for drawing generalized inferences.

The model that we describe below operationalizes this approach. The empirical analysis gives us an application. The first goal is to establish whether there is heterogeneity in the impact of a program across sites (in our empirical work, we examine whether the impact of reduced class size on student performance varies across sites) If so, then the average impact at one set of sites or across all sites does not necessarily generalize to other sites. The question then is whether site characteristics interact with treatment, and whether modeling these interactions accounts for the between-site variation in the impact; if yes, then the moderator effects serve as a basis for drawing more accurate generalizations, since we can account for features of context that intensify or suppress the effect of the treatment.

There are three points to note about the model and the analysis.

- The approach allows us to determine whether moderator analyses can *in principle* increase the generalizability of findings. We use the results of a multisite trial where we have an unbiased impact estimate for each site to measure how accurate estimates of impact from one set of sites would be if used to infer impact at other sites. We use the unbiased impact estimates available for each site as the benchmark to assess the accuracy of our between-site comparisons. (The method presented here is an extension of the

[†] Such information is obviously undergirded by causal explanatory theory, and choice of moderators may be critical, a point we will discuss in the conclusion in light of the results of the empirical study given below.

approach used by Lalonde (1986) and Bloom et al., (2005) in econometrics and used by Wilde and Hollister (2002) and Agodini and Dynarski (2004) in education. Their chief concern was with quantifying levels of selection bias in program effect estimates from comparison group studies.)

- Related to the point above, our analysis also allows us to address the question that Lalonde (1986) and colleagues have investigated: they considered the extent to which average performance varies across sites in the absence of treatment and the extent to which regression adjustments reduce this variation. This variance is an indicator of effect of selection into sites – it is due to systematic differences between sites leading to variation in average performance. Estimates of program effects from comparison group studies can have both types of inaccuracy: (1) due to a difference between the treatment and comparison group in average performance that is not attributable to treatment, and (2) due to a difference between them in the actual effect of the program (what we would measure if we could run an RCT for both the program and comparison groups.) We will express both forms of inaccuracy in the model below.
- We are interested not only in whether this variance is present but whether our covariates account for it; and if they do not, why not. That is, our work leads to the question of whether covariates that have a theoretical basis are better at accounting for the variation than all-purpose demographic variables that are routinely available.

Results 2: Model

We consider the impact at a specific site, q. We assume that a randomized trial has *not* been carried out at this site. We will use information about performance of students at other sites to infer what the impact is at q.

We start with the quantity $\hat{} \frac{\sum_{p \neq q} \frac{y_{**p,t}}{nJ}}{N-1} - \frac{y_{**q,c}}{nJ}$ as an estimate the impact at q. It consists of the difference between the cross-site average of performance under treatment at sites *p* other than *q*, and average performance at *q* in the absence of treatment (*n* is the number of students per teacher, *J* is the number of teachers per school, *N* is the number of schools, and *y* is student performance measured after the program has run its course, i.e., the posttest.) We show in the full work that the Mean Squared Error for this estimate averaged across all sites is:

$$\text{MSE}= \frac{N}{N-1}\tau_0^2 + \frac{N}{N-1}\tau_1^2 + \frac{2N}{N-1}\tau_{10}^2 + (\frac{\upsilon^2}{J} + \frac{\sigma^2}{Jn} + \frac{\upsilon^2}{JN} + \frac{\sigma^2}{nJN})$$

τ_0^2 is the between-site variance in the site-average performance in the absence of treatment.

τ_1^2 is the between-site variance in the site-average treatment effect.

τ_{10}^2 is the school-level covariance between site-average performance in the absence of treatment and site-average treatment effect.

υ^2 is the within-school teacher-level sampling variation.

σ^2 is the within-teacher student-level sampling variation.

The magnitudes of these parameters set limits to the extent to which, in expectation, results from all but one site are useful for predicting what the impact will be at the one site. In this study we

are chiefly interested with τ_1^2 as a source of inaccuracy, because it is this quantity that can potentially be reduced through modeling moderating effects of school-averages of certain covariates. (Prior studies, like Lalonde's (1986), are interested in estimating τ_0^2 and studying whether it can be reduced.)

Results 3: Empirical Analysis*[‡]*:

We use results from the Tennessee Class Size reduction experiment (Project STAR) (Finn and Achilles, 1990; and Mosteller, 1995) - a multisite trial - to apply our model and illustrate our approach to generalizability. Students were randomized in kindergarten to small classes, regular classes, or regular classes with an aide. The experiment lasted 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 outperformed the controls by .20 standard deviation units.

The STAR experiment showed variation in impact across sites. As described above, we can regard this variation as indicating that the average effect does not generalize 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.

We used HLM (Raudenbush and Bryk, 2002) to estimate the variance components described above and to produce the empirical results described here.

<u>*Empirical result 1:*</u> Figure 1 (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 average performance and in the treatment effect for the reading outcome. The points compare the proportion of variance in (1) school average performance and (2) school departures from the grand average of the treatment effect that remain after modeling the covariates and their interactions with the treatment indicator: $(1-\frac{\hat{\tau}_0^2-\hat{\tau}_0^{2*}}{\hat{\tau}_0^2}, 1-\frac{\hat{\tau}_1^2-\hat{\tau}_1^{2*}}{\hat{\tau}_1^2})$. Here, τ_0^{2*} and τ_1^{2*} are the variance components from the conditional models (i.e., after including main and interaction effects in the model.) (Covariates are identified in the figure, we do not present a separate list or exact definitions due to the limited space of this abstract.) 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 the case of this multi-site trial, the covariates do not account for systematic differences across schools in the impact, and therefore, are not useful for establishing generalizations about the effects of small classes on reading achievement.*

<u>*Empirical result 2*</u>: In Table 1 we display estimates of τ_1 / SD and τ_1^* / SD (for the model that includes all covariates and their interactions with treatment), where *SD* is the standard deviation in the posttest. The purpose is to show how much uncertainty results from using the average of *experimental* impact estimates from other sites to estimate impact at a given site. We express this

[‡] This section includes a short description of **Setting, Population / Participants / Subjects, Intervention / Program / Practice, Research Design** for the empirical component of this work**.**

in *SD* units, which is a familiar scale. We see that these effects are not small (.23) (if we consider effect sizes as small as .20 as educationally important) and accounting for cross-site difference in the impact by modeling effects of moderators does not reduce this uncertainty by much (the term drops to .19). (We include estimates of the other variance components for reference.) In the case of this multi-site trial, experimental effect estimates do not generalize across sites (Here we report outcomes for reading only; in the full paper we show results for math – which are similar.)

Usefulness / Applicability of Method:
This work provides a methodology for assessing the generalizability of program effect estimates. It is an alternative to existing approaches (HR and CE) and has the benefit of allowing us to quantify the extent to which impacts vary (and hence don't generalize across situations) as well as the extent to which moderator effects increase generalizability by accounting for heterogeneity. We determined that, at least in the case of the intervention we explored, experimental estimates do not generalize well (relying on *experimental* estimates of impacts from other sites on average results in bias of .23 sd before adjustment for covariates, and .19 after adjustment: *what is an unbiased effect estimate at one site, may be biased for another site*. This serves as a cautionary message. The finding that all-purpose demographic variables don't do much to improve the accuracy of the effect estimates through modeling their interactions with treatment leads to a discussion of what types of information we should be collecting to better generalize results from randomized trials.

Conclusions:

In the empirical analysis we examined whether experimental impact estimates generalize, and we found that for the intervention considered, they do not: The estimate of the grand average impact is sufficiently different from the impact at a given site (in expectation) that it cannot be considered generalizable to, or representative of, the unbiased impact at the site. Modeling moderating effects does not improve the situation. (The accuracy of estimates from comparison group studies (as opposed to from experiments done at other sites) is even worse, with average bias being .53 sd units prior to covariate adjustment and .38 sd units after regression adjustment for the effects of all 'convenience covariates' combined.) This delivers a stark message: effect estimates gathered from existing sites cannot be trusted for inferring impacts at new sites. The question is, why? One possibility is that the covariates used to account for differences in average performance across sites, or in the average impact of a program across sites, are uninformative in the sense that they don't address the selection mechanism of individuals into sites or tap into the mechanisms through which characteristics that are imbalanced across sites interact with the program. Our recommendation is for programs to invest in articulating theory of what moderates program impacts and developing reliable measures of these factors. As the program moves from development though scale-up theory-based moderators can help establish external validity by accounting for variations in impact across conditions, something that all-purpose demographics do not do. The result would be much greater informational yield from our experiments and a bigger return on our investment in research efforts.

Appendices

Appendix A. References

Agodini, R., & Dynarski, M., (2004). Are experiments the only option? A Look at dropout prevention programs. *The Review of Economics and Statistics, 86,* 180-194.

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

Bloom, H. S., et al. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effect. 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.

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.

Lalonde, R., (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76, 604-620.

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.

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.

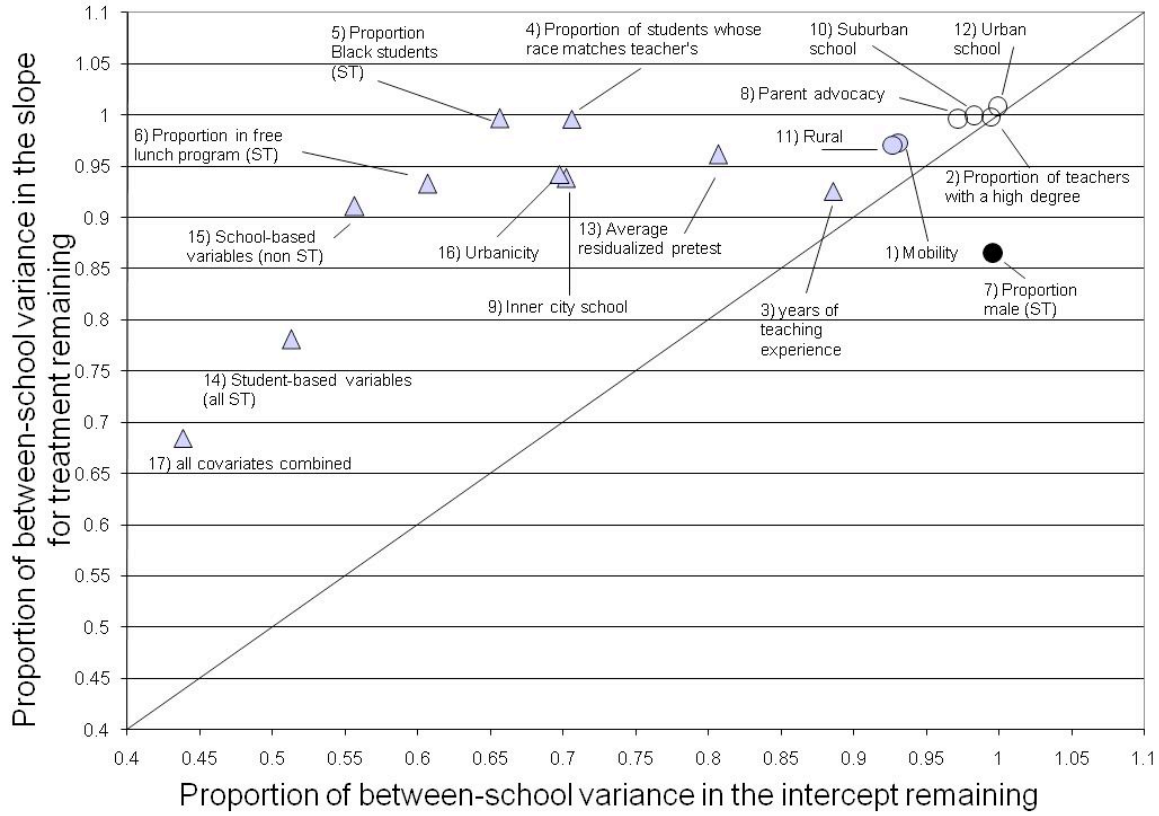
Riecken, H. W., & Boruch, R. F., (1974). Why and when to experiment. In H. W. Riecken, R. F. Boruch, D. T. Campbell, et al. (Eds.), *Social experimentation: A method for planning and evaluating social interventions.* New York: Academic Press.

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

Wilde, E. T., & Hollister, R., (2002). How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes. Institute for Research on Poverty Discussion paper no. 1242-02.

Appendix B. Tables and Figures

Figure 1: Reading outcome: Proportion of squared bias remaining (Estimates of $(1-\frac{\hat{\tau}_0^2-\hat{\tau}_0^{2*}}{\hat{\tau}_0^2}, 1-\frac{\hat{\tau}_1^2-\hat{\tau}_1^{2*}}{\hat{\tau}_1^2})$ *)*



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.

Table 1: Summary of Bias

	Average bias due to variation in the baseline effect (in standard deviation units of the posttest)	Average bias due to variation in the treatment effect (in standard deviation units of the posttest)	Average bias due to variation in the baseline effect plus the variation in the treatment effect (in standard deviation units of the posttest)
Without adjustment (i.e., using the results from the model with no covariates)	$\frac{\hat{\tau}_0}{SD} = .38$	$\frac{\hat{\tau}_1}{SD} = .23$	$\frac{\sqrt{\hat{\tau}_0^2 + \hat{\tau}_1^2 + \hat{\tau}_{10}^2}}{SD} = .53$
With adjustment (i.e., using the results of the model that includes all covariates.)	$\frac{\hat{\tau}_0^*}{SD} = .25$	$\frac{\hat{\tau}_1^*}{SD} = .19$	$\frac{\sqrt{\hat{\tau}_0^{*2} + \hat{\tau}_1^{*2} + \hat{\tau}_{10}^{*2}}}{SD} = .38$

Note: all effects estimates are significantly different from zero.