Towards a Framework for Empirically Testing Conditions for External Validity Bias in Causal Effect Estimation

Andrew P. Jaciw
Fatih Unlu
Thanh Nguyen

Background

Problems of causal generalizations are becoming increasingly prominent in methodological research in education. This reflects concerns about the narrow applicability of average treatment effect findings, and the policy priority to better understand which programs work better for whom and under which conditions (Bryk, 2014; Tipton and Olsen, 2018).

Several approaches to casual generalization have emerged in education and related fields. They include the "heterogeneity of replication approach" (Shadish, Cook and Campbell 2002; Cook 2002), methods of reweighting (such as by Tipton, 2013) and G theory (Cronbach, Rajaratnam, and Gleser, 1963; Shavelson & Webb, 2008). Moderators of impact play a central role in each of these methods. Analysis-based solutions (as opposed to design-based solutions) to problems of causal generalization require identifying and addressing effects of effect moderators. For example, to satisfy the sampling ignorability assumption, reweighting methods depend on identifying and adjusting for effects of moderators of impact that are imbalanced between study and inference population, (Tipton and Olsen, 2018).

What is currently missing in the field, and what this work addresses, is an empirical framework for testing the accuracy of generalized casual inferences. That is, we require empirical tests of how effective adjusting for moderators is at producing accurate generalizations.

In this work, we empirically evaluate the capacity of moderators of impact to account for effect heterogeneity that places limits on external validity of casual inferences. The method is based on a framework for evaluating replicability of casual effects across sites of a multisite trial. Due to the brevity of this proposal, we append discussion of the framework itself. (We consider it a contribution in its own right that formalizes the connection between effect heterogeneity and generalizability.) We focus instead on empirical tests motivated by the framework: they boil down to assessing the level of variation in program impact across sites of a trial, and potential for its reduction through modeling effects of moderators. The work is consistent with the SREE theme of "Practical Significance" and "Communicating What Matters" as the goal of the work is an empirical solution for advancing accuracy of generalized causal inferences.

Method: Estimation and Research Questions

Raudenbush & Bloom (2015) argue that in the context of making generalizations to a population of sites, a natural cross-site impact heterogeneity parameter of interest is the mean squared difference between the site-specific impacts and the cross-site average mean impact.
Bloom, Raudenbush, Weiss, and Porter (2017), Weiss et al. (2017), and Weiss, Miratrix, and Henderson (2019) discuss that an HLM-based "fixed intercept, random (treatment) coefficient" (FIRC) estimator that produces consistent estimates of this parameter. A second similar approach uses an HL model with random intercepts instead of fixed intercepts; that is, a "random intercept, random (treatment) coefficient" (RIRC) estimator.

In this work we examine levels of impact heterogeneity and trends in its reduction using both approaches to estimation. Due to the word limit we describe the underlying HL models in Appendix B. With both FIRC and RIRC we estimate impact variation across sites net of individual-level sampling variation prior to modeling any moderating effects, and after inclusion of specific sets of site-level moderators.

Our research questions are:

1. In the context of a multisite trial, what is the level of cross-site impact heterogeneity expressed in standardized units ($IH^*_{ES}(FIRC) = \frac{\sqrt{\tau^*_1}}{sd}$) before adjusting for the effect of any moderators? (This quantity is "Impact Heterogeneity" (*IH*) expressed as a standardized effect size (*ES*) based on *FIRC* estimation, where τ^*_1 is impact variation across sites, *sd* is the standard deviation of the outcome variable.)
2. What are the levels after adjusting for the effect of moderators grouped in terms of site averages of student covariates, site averages of teacher covariates, site-characteristics, and combination of these covariate types?
3. Are the reductions in (2) statistically significant when compared to the magnitude of heterogeneity in (1)?
4. Is the level of heterogeneity and trends in its reduction from modeling effects of moderators similar when we estimate *FIRC* and *RIRC*?

Intervention, Sample and Data

Impacts are assessed for 40 sites with over 10,000 students total. The intervention is an inquiry-based program that uses hand-on instruction to increase students' science, math and literacy skills. The program was implemented over the course of one school year in grades 4 – 8. We focus on the reading outcome as assessed through the SAT-10 reading test. Covariates and models assessed are shown in Table 1.

Table 1. Models assessed with sets of moderators explored.

Site-level covariates		Model (Moderator Effects Included)								
		1	2	3	4	5	6	7	8	9
Teacher factors	Math / Science Degree Ranks					X	X			X
	Years Teaching					X	X			X
	Adopt Constructivist Principles Math					X	X			X
	Adopt Constructivist Principles Science					X	X			X
Student factors	Average pretest		X		X		X		X	X
	Proportion Male			X	X		X		X	X
	Proportion Free or Reduced Price Lunch			X	X		X		X	X
	Proportion Minority			X	X		X		X	X
	Proportion English Learner			X	X		X		X	X
	Proportion in grades 4-8			X	X		X		X	X
Site factors	Region							X	X	X
	Locale							X	X	X
	Number of students per site							X	X	X

Results

Findings are summarized in Figure 1 and Table 2.

1. We observed similar levels of impact heterogeneity before adjustment for effects of site-level moderators, with estimates of IH^*_{ES}, of .091 for *FIRC* and .106 for *RIRC*. Corresponding variance components, τ^*_1 , are statistically significant (p<.001).
2. Inclusion of moderating effects of all variables combined (Model 9), or of site or regional characteristics and site averages of student covariates (Model 8) accounted for close to all or all effect heterogeneity. With RIRC in particular, modeling individual sets of covariates (e.g., student-only, or teacher-only) led to smaller reduction in heterogeneity, while combining sets (e.g., student and teacher moderators) was more effective at reducing impact variation.
3. Significance of reductions in heterogeneity compared to the initial magnitude of heterogeneity are shown in Table 2.
4. There are similarities between FIRC and RIRC in heterogeneity and trends in its reduction from modeling effects of moderators; RIRC also seemed to give more finely-graded results.

Figure 1. FIRC and RIRC estimates of levels of effect heterogeneity prior to and following adjustment for effects of moderators of treatment impact

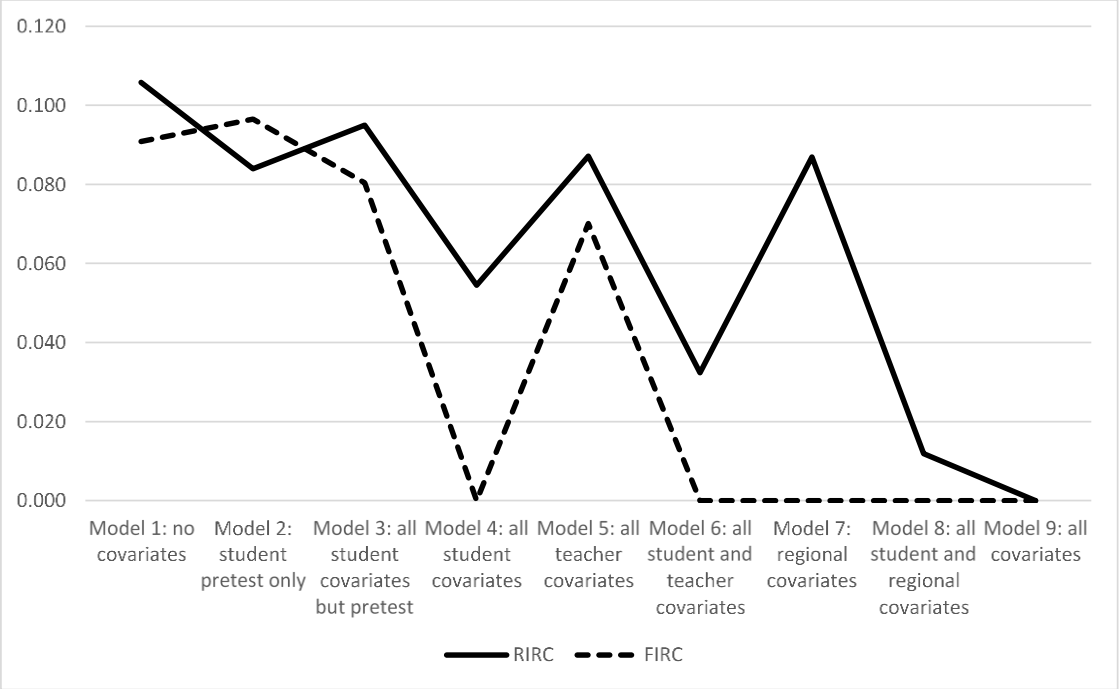

Table 2. FIRC and *RIRC* estimates and tests of hypotheses concerning random effects

	FIRC	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**
1	$IH^*_{ES}(FIRC) = \frac{\sqrt{\tau^*_1}}{sd}$	0.091	0.097	0.080	0.000	0.070	0.000	0.000	0.000	0.000
2	*Deviance (degrees of freedom)*		1.4 (1)	6.1 (8)	6.7 (9)	13.3 (8)	40.2 (17)	5.7 (9)	53.7 (18)	89.4 (26)
3	$H_0: \tau^*_1 < \tau^*_{1(Model\ 1)}$	-				*	****		****	****
4	$H_0: \tau^*_1 = 0$	***	***	***		**				
	RIRC									
5	$IH^*_{ES}(RIRC) = \frac{\sqrt{\tau^*_1}}{sd}$	0.106	0.084	0.095	0.054	0.087	0.032	0.087	0.012	0.000
6	*Deviance (degrees of freedom)*	-	1.1 (1)	5.4 (8)	15.7 (9)	11.1 (8)	38.6 (17)	7.0 (9)	40.8 (18)	67.4 (26)
7	$H_0: \tau^*_1 < \tau^*_{1(Model\ 1)}$	-			*		***		***	****
8	$H_0: \tau^*_1 = 0$	****	***	***	**	***	*	***		

*p<.10, **p<.05, ***p<.01, ****p<.001

Conclusions

In this study external validity bias was .10 SD and statistically significant before any adjustment. This is not a trivial amount, if we assume empirical benchmarks for the importance of effects. However, impact heterogeneity (and, correspondingly, external validity bias) were reduced to zero with inclusion of all available covariates with *FIRC* and *RIRC*. The results also hold promise that generalization tools such as the Generalizer (Tipton and Miller, 2015) can reduce bias, as combinations of pretest with site-based covariates (variables like those from publicly available datasets that the Generalizer uses) led to complete or almost complete reductions in bias.

References

Bloom, H., Raudenbush, S., Weiss, M. & Porter, K. (2017). Using multi-site experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness, 10*(4), 817 – 842, DOI: 10.1080/19345747.2016.1264518. Retrieved from: https://doi.org/10.1080/19345747.2016.1264518

Bryk, A. S. (2014). 2014 AERA Distinguished lecture: Accelerating how we learn to improve. *Educational Researcher, 44*(9), 467–477.

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.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16*, 137-163.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *The Journal of Human Resources, 22*, 194–227.

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

Raudenbush, S. W. & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation, 36*(4), 475-499). Doi: 10.1177/1098214015600515

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

Shavelson, R. J., & Webb, N. M. (2008). *Generalizability theory and its contributions to the discussion of the generalizability of research findings*. In K. Ercikan & W. M. Roth (Eds.), Generalizing from educational research (pp. 13–32). New York, NY: Routledge.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38*, 239-266.

Tipton, E. & Miller, K. (2015). *Generalizer [Web-tool*]. Retrieved at http://www.generalizer.org

Tipton, E. & Olsen, R.B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher, 47*, 516-524.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N.

(2017). How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence from Past Multisite Randomized Trials. *Journal of Research on Educational Effectiveness,10*(4), 843-876, DOI: 10.1080/19345747.2017.1300719. Retrieved from https://doi.org/10.1080/19345747.2017.1300719

Weiss, M. J., Miratrix, L., & Henderson, B. (2019). *An Applied Researcher's Guide to Estimating Effects from Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates*. Manuscript Under Review.

Appendix A: A Framework for Evaluating the Generalizability of Causal Effect Estimates from Experiments

Our framework is a direct extension of Within Study Comparison (WSC) approach traditionally used to empirically test the success of Quasi-Experimental Designs (QEDs) at replicating Average Treatment Effect estimates from experiments (Lalonde, 1985; Fraker and Maynard, 1986). WSC compares analysis-based (QED) solutions against benchmark design-based solutions (from RCTs). The goal is to identify a valid counterfactual – as good as an experimental control – through analytic strategies. Put another way, WSC assesses if the ignorability assumption is met. WSC efforts spanning several decades have evaluated the success of (1) models, (2) covariates, and (3) sampling choices, at increasing accuracy of QED-based impact findings.

Our extension of the approach, which we refer to as e-WSC, evaluates whether adjusting for moderators of impact allows accurate extrapolation of causal impact findings from sample to inference population. The goal of e-WSC is to replicate a known impact quantity at an inference site through adjustment of impact quantities from other study sites.

We adopt a variant of the WSC that uses a multisite trial to evaluate whether impact at a given site can be replicated using impact quantities at other sites. This variant is referred to as a "1 versus *N*-1" or "one-out" strategy because the goal is replication of impact at a given ("one out") site *j* using results for *N*-1 other sites, and this can be done *N* times (i.e., with each site taking a turn being the inference site.)

The Framework

Assume an *N*-site trial. Our goal is to predict impact at site $j=a$, using performance outcomes from *N*-1 other sites. The steps of the process are described in Table A1 below, with corresponding quantitative expressions in Table A2.

Table A1. Steps in "one out" e-WSC

Step	Goal
1	Estimate the unbiased impact quantity at inference site $j = a$: $E(\Delta_i \mid T = 1, j = a)$, where Δ_i is the difference in potential outcomes for individual i: $\Delta_i = y_i(T = 1) - y_i(T = 0)$. (The goal of e-WSC is to replicate this target of inference; that is, impact at site $j=a$.) (Quantity 1 in Table 2.)
2	Assume an experiment has not been conducted at inference site $j=a$; therefore, generalize impact for that site using the average of impacts across the remaining *N*-1 study sites: $E_{j \neq a}\big(E(\Delta_i \mid T = 1, j)\big)$ (quantity 2 in Table 2)
3	Calculate external validity bias (EVB) in the inferred quantity for site $j=a$ as $E_{j \neq a}\big(E(\Delta_i \mid T = 1, j)\big) - E(\Delta_i \mid T = 1, j = a)$ (Quantity 3 in Table 2).
4.	Summarize average absolute EVB as the root squared difference of this quantity (Quantity 4 in Table 2).
5	Calculate the root mean squared difference in expectation over all *N* sites (i.e., where we assume each site takes a turn being an inference site.) (Quantity 5).
6	We include Quantity 6 in Table 2 to show that equal weighting of sites in the grand mean impact is but one option. (A function "g" indicates that other estimands, such as "precision weighted" averages, may also be used as the impact quantity inferred for a given site.)

Table B2. Quantitative components in an e-WSC analysis

	Without covariate adjustments	With covariate adjustments
1.Impact for site $j = a$	$E(\Delta_i \mid T = 1, j = a)$	$E(\Delta_i \mid T = 1, j = a)$
2. Generalized impact quantity for site $j = a$	$E_{j \neq a}(E(\Delta_i \mid T = 1, j))$	$E_{j \neq a}(E(\Delta_i \mid T = 1, X = X_a, j))$
3. EVB in impact quantity inferred for site $j = a$	$E_{j \neq a}\big(E(\Delta_i \mid T = 1, j)\big) - E(\Delta_i \mid T = 1, j = a)$	$E_{j \neq a}(E(\Delta_i \mid T = 1, X = X_a, j)) - E(\Delta_i \mid T = 1, j = a)$
4. Magnitude of EVB for site $j = a$	$\sqrt{[E_{j \neq a}(E(\Delta_i \mid T = 1, j)) - E(\Delta_i \mid T = 1, j = a)]^2}$	$\sqrt{[E_{j \neq a}(E(\Delta_i \mid T = 1, X = X_a, j)) - E(\Delta_i \mid T = 1, j = a)]^2}$
5. Average magnitude of EVB across all sites	$\sqrt{E_j\{[E_{j \neq a}(E(\Delta_i \mid T = 1, j)) - E(\Delta_i \mid T = 1, j = a)]^2\}}$	$\sqrt{E_j\{[E_{j \neq a}(E(\Delta_i \mid T = 1, X = X_a, j)) - E(\Delta_i \mid T = 1, j = a)]^2\}}$
6. Generalized average magnitude of EVB	$\sqrt{E_j\{g[[E_{j \neq a}\big(E(\Delta_i \mid T = 1, j)\big)] - E(\Delta_i \mid T = 1, j = a))]^2]\}}$	$\sqrt{E_j\{[g[E_{j \neq a}\big(E(\Delta_i \mid T = 1, X = X_a, j)\big)] - E(\Delta_i \mid T = 1, j = a)]^2]\}}$

Note: Δ_i is the difference in potential outcomes for individual *i*: $\Delta_i = y_i(T = 1) - y_i(T = 0)$. EVB = "External Validity Bias".

Appendix B: Models used to Estimate Effects

For a given multisite trial, our first goal is to assess levels of impact heterogeneity across sites. Our second goal is to determine the extent to which adjusting for effects of moderators accounts for this variation.

The first step in the estimation of cross-site variation in impacts is to define the estimand of interest. Raudenbush & Bloom (2015) argue that in the context of making generalizations to a population of sites, a natural cross-site impact heterogeneity parameter of interest is the mean squared difference between the site-specific impacts and the cross-site average mean impact.

Bloom, Raudenbush, Weiss, and Porter (2017), Weiss et al. (2017), and Weiss, Miratrix, and Henderson (2019) discuss that an HLM-based "fixed intercept, random (treatment) coefficient" (FIRC) estimator produces consistent estimates of this parameter. They argue that among alternative estimators, including those that estimate this parameter using the distribution of site-specific impact estimates yielded by OLS or empirical Bayes methods, this estimator has the desirable property that it weights the average impact for each site in proportion to the precision with which that impact can be estimated. The distribution of OLS estimates of site-specific effects confounds site-level estimation error with true impact variation and empirical Bayes estimators understate true cross-site impact variation.

HLM Specifications

Stage 1. The "base-model" for estimating impact heterogeneity. Following Weiss, Miratrix & Henderson (2019) and Weiss et al. (2017), we specify a base model for obtaining a FIRC estimate of impact heterogeneity:

Level-1 (students):

$$y_{ij} = \sum_{r=1}^{R} \alpha_r RAB_{rij} + \beta_j T_{ij} + \sum_{r=1}^{L} \gamma_r X_{lij} + \varepsilon_{ij} \quad (3)$$

y_{ij} is the outcome for student *i* at site *j*. RAB_{rij} equals one if individual *i* from site *j* belongs to random assignment block *r*, and 0 otherwise. (This accommodates trials in which students are randomized within blocks within a site. If the site is the block, then *R=j*). T_{ij} takes values 0 or 1, indicating random assignment of individuals to control or treatment, respectively. Site-centered student-level baseline covariates, X_{lij}, are included to improve the precision of parameter estimates.

Level-2 (sites):

$$\beta_j = \gamma_1 + u_{1j} \quad (4)$$

We assume the following:

$$\varepsilon_{ij} \sim N(0, \sigma^2_{|X,RAB})$$

$$u_{1j} \sim N(0, \tau_1^2)$$

$$Cov(\varepsilon_{ij}, u_{1j}) = 0$$

A quantity of main interest is the estimate of impact variation across sites:

$$\widehat{\tau_1^2} = \widehat{Var(u_{1j})} \quad (5)$$

Note that the model above does not control (or adjust) for any moderators. Therefore, we consider this estimate as the unconditional or uncontrolled estimate of the cross-site impact heterogeneity. We can also estimate impact heterogeneity as the standard deviation of the distribution of impact variation and in units of the pooled standard deviation of the outcome; that is, in the metric of a standardized effect size:

$$IH_{ES} = \frac{\tau_1}{sd} \quad (6)$$

IH stands for "Impact Heterogeneity" with subscript *ES* for "Effect Size", and *sd* is the standard deviation of y_{ij}. This metric has the advantage that is allows both a more direct comparison with the mean impact also expressed as an effect size, and a synthesis of such quantities across studies, as in meta-analysis.

Stage 2. Specifying the "moderator-adjusted model" for estimating remaining impact heterogeneity. A second goal is to index impact heterogeneity conditional on interactions between site-level covariates and treatment. To do this, we retain the level-1 model (Equation 3) while introducing site-level covariates (posited moderators) in the equation for the coefficient for treatment. Specifically, we have at Level-2:

$$\beta_j = \gamma_1^* + \sum_{p=1}^{k} \gamma_{1p} \overline{X_{jp}} + \sum_{r=k+1}^{k+1+l} \gamma_{1r} Z_r + u_{1j}^* \quad (7)$$

The terms in Equation 7 are interacted with treatment assignment variable at Level-1 (T_{ij}). The coefficients γ_{1p}, and γ_{1r} represent the moderating effects of site-level attributes on the impact of the program on the outcome. They include site averages of *k* individual-level covariates, $\overline{X_{jp}}$, (e.g., site averages of student pretest) and *l* site-level covariates. We are interested in (a) their effects individually, (b) their effects in specific combinations, and (c) the variance of u_{1j}^* which expresses the remaining heterogeneity in impact conditional on interactions in the model. That is, u_{1j}^* are site-specific deviations in impact after conditioning outcomes on the main effects of site-level covariates and their interactions with treatment. We are interested in the following estimate:

$$\widehat{\tau_1^{*2}} = \widehat{Var(u_{1j}^*)} \quad (8)$$

As above, we are interested in heterogeneity expressed in the metric of a standardized effect size; that is, we will estimate the following quantity:

$$IH_{ES}^* = \frac{\tau_1^*}{sd} \quad (9)$$

In this work we also present results of "random intercept, random (treatment) coefficient" (RIRC) estimation. The HL models are analogous to ones above and the procedure of adding moderator effects is the same. RIRC differs from FIRC in that instead of modeling fixed intercepts, differences across sites in average achievement are captured through a site-level random effect.