The Use of Moderator Effects for Drawing Generalized Causal Inferences

Andrew Jaciw
Empirical Education Inc.

Purpose of this Study.

1) To provide a rationale for an alternative approach for assessing the generalizability of results from experiments and comparison group studies
2) To formalize the approach in (1) through a quantitative model
3) To apply the approach to results from the Tennessee Class Size reduction experiment (Project STAR)—a multi-site trial

Assertion.

A rationale for an approach to assessing the generalizability of results from RCTs

If a treatment effect is constant across a set of sites, then we can say that the average effect generalizes across these sites. *If a treatment effect varies* across the set of sites, then the impact for one subset of sites in the sample does not generalize to a different subset. *Treatment heterogeneity implies that a simple average impact estimate is inadequate as a generalization, but it presents an opportunity for establishing generalizability by accounting for this heterogeneity through moderator effects.*

Derived Results.

Expressions for uncertainty in local impact estimates

Approaches to estimating impact at site q	Estimates of the impact at site q*	Average mean squared error for estimates of impact at q**
We compare performance at N-1 sites p (other than q) that receive treatment, to performance at the single site of interest, q, where the intervention has not been used.	$\hat{\theta}_1 = \frac{\sum_{p \neq q} \frac{y_{**p,t}}{nJ}}{N-1} - \frac{y_{**q,c}}{nJ}$	$\widehat{\overline{MSE(\hat{\theta}_1)}} = \frac{1}{N}\sum_{q=1}^{N} MSE(\hat{\theta}_1)$ $= \frac{N}{N-1}\hat{\tau}_0^2 + \frac{N}{N-1}\hat{\tau}_1^2 + \frac{2N}{N-1}\hat{\tau}_{10}^2 + \{\frac{N}{N-1}(\frac{\hat{\upsilon}^2}{J} + \frac{\hat{\sigma}^2}{nJ})\}$
We assume that a randomized trial has *not* been carried out at q, but RCTs have been carried out at each of the N-1 other sites. We will use the impact estimates from the other sites to infer what the impact is at q.	$\hat{\theta}_2 = \frac{\sum_{p \neq q} (\frac{y_{**p,t}}{nJ/2} - \frac{y_{**p,c}}{nJ/2})}{N-1}$	$\widehat{\overline{MSE(\hat{\theta}_2)}} = \frac{1}{N}\sum_{q=1}^{N} MSE(\hat{\theta}_2)$ $= \hat{\tau}_1^2 + \frac{4}{N-1}(\frac{\hat{\sigma}^2}{nJ})$

* 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 assume a balanced design.

** $\hat{\tau}_0^2$ is the estimated between-site variance in site-average performance in the absence of treatment.

$\hat{\tau}_1^2$ is the estimated between-site variance in the site-average treatment effect.

$\hat{\tau}_{10}^2$ is the estimated school-level covariance between site-average performance in the absence of treatment and the site-average treatment effect.

$\hat{\upsilon}^2$ is the estimated within-school teacher-level sampling variation.

$\hat{\sigma}^2$ is the estimated within-teacher student-level sampling variation.

Method and Empirical Results.

Background

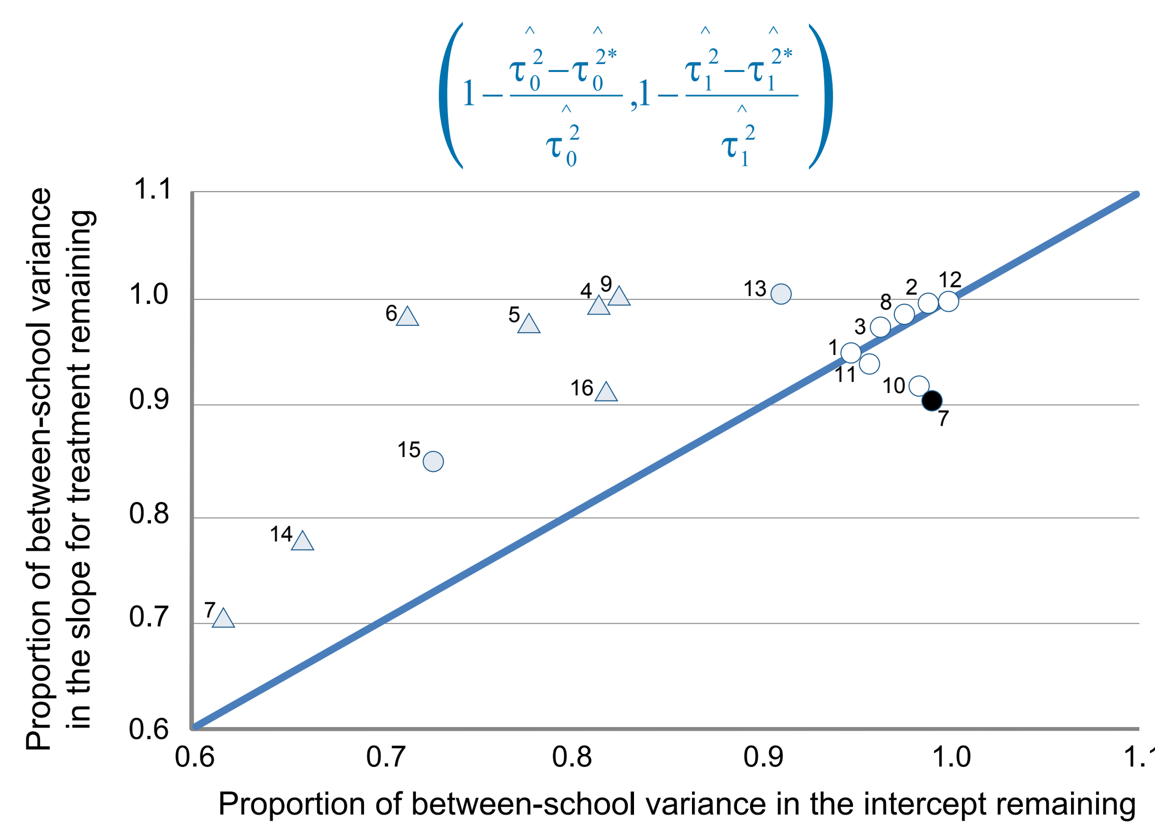
We used results from the Project STAR study to apply our model and illustrate our approach to generalizability.* Students were randomized in kindergarten to small classes, regular classes, or regular classes with an aide. Teachers were also randomized to classes. Randomization was conducted within each of 79 schools. The outcome measures were scale scores in reading and math. The average effect of small classes was significant and positive in both math and reading at every grade level (ranging between .15 and .30 sd units).

*We used SAS PROC MIXED and an HLM approach to estimate the variance components and mean squared errors needed to assess the generalizability of the findings. (We focus on math and reading outcomes at the end of grade 3 for students who persist in the same school and condition over the course of the trial.)

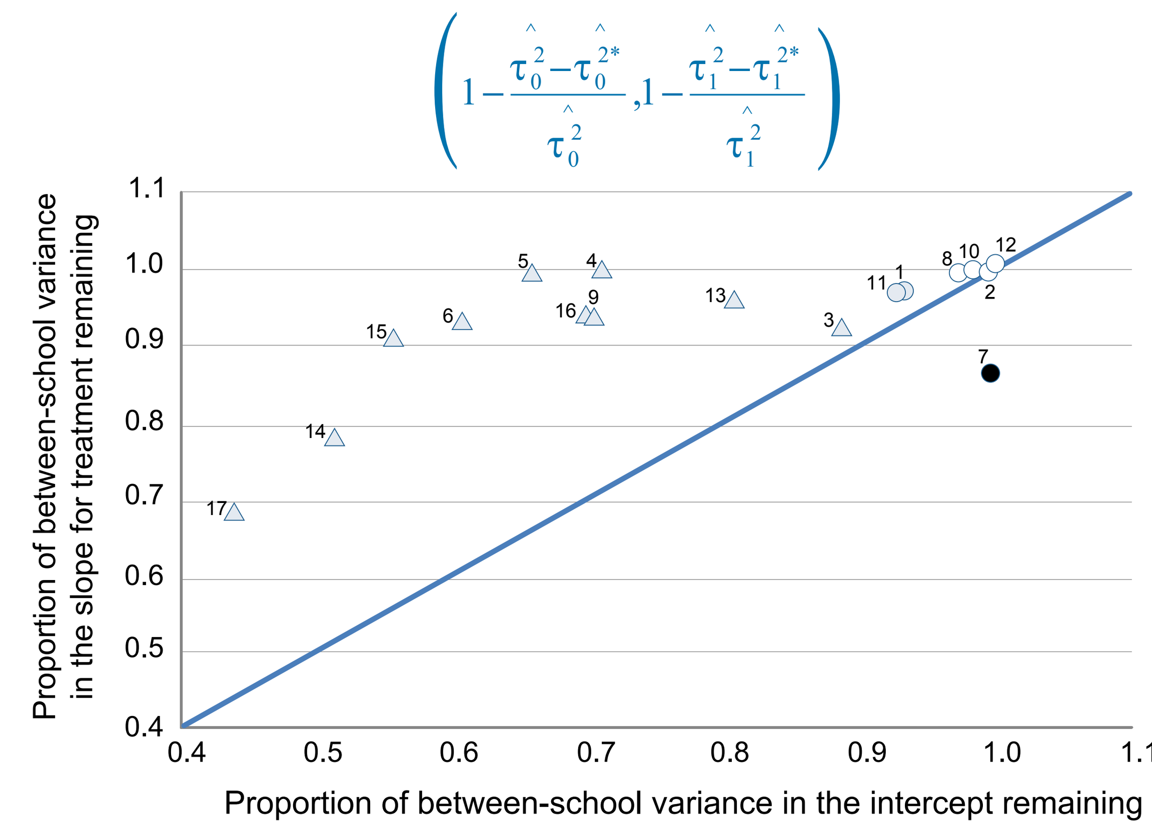
Findings

The experiment demonstrated variation in impact across sites. The current work asserts that due to this variation, the impact at one subset of sites in the sample does not necessarily generalize to a different subset. We started by modeling the impact and average performance as randomly varying across schools. Our goal was to reduce the quantities τ_0^2, τ_1^2, and τ_{10}^2 by modeling the interaction(s) of treatment with one or more school-level covariates (potential moderators of the impact.) The corresponding variance components from models that include one or more moderators and their interaction(s) with treatment are denoted as: τ_0^{*2}, τ_1^{*2}, and τ_{10}^{*2}. The potential moderators were 'convenience variables'–simple demographics not theoretically tied to the intervention.

Proportions of Variance Remaining for the Math Outcome



Proportions of Variance Remaining for the Reading Outcome



1) Mobility
2) Proportion of teachers with a high degree
3) Years of teaching experience
4) Proportion of students whose race matches teacher's
5) Proportion Black students (ST)
6) Proportion in free lunch program (ST)
7) Proportion male (ST)
8) Parent advocacy
9) Inner city school
10) Suburban school
11) Rural school
12) Urban school
13) Average residualized pretest (ST)
14) Student-based variables (all ST)
15) School-based variables (all non ST)
16) Urbanicity combined
17) All covariates combined

The analyses are numbered in the figures. 'The covariate(s) that is/are modeled is/are indicated for each analysis. The covariates are modeled one at a time with the exception of Analyses 14-17 where certain combinations of covariates (and terms for their interactions with the treatment indicator variable) are modeled; Analysis 17 uses the 'fully conditional model' which includes all school-level covariates and their interactions with treatment. (ST='student based covariate'; these are variables that are based on attributes of students or their designations)

Note: Gray: main effect(s) of the covariate(s) is/are statistically significant (p<.05); Black: interaction(s) between covariate(s) and treatment is/are statistically significant (p<.05); Empty marker: neither of these conditions holds; Triangle: model with both main and interaction effect(s) *results in a better fit* than the reference model (i.e., the model without any school-level main or interaction effects); Circle: model with both main and interaction effect(s) *does not result in a better fit* than the reference model.

Conclusion 1: For 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. **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.)**

Summary of Indicators of Accuracy: Math

	Square root of variation in average performance (expressed in standard devation units of the posttest)	Square root of variation in the treatment-control difference in performance (expressed in standard deviation units of the posttest)	The square root of the estimated average MSE of $\hat{\theta}_1$ (expressed in standard deviation units of the posttest)	The square root of the estimated average MSE of $\hat{\theta}_2$ (expressed in standard deviation units of the posttest)
Without adjustment (i.e., using the results from the model with no school-level covariates)	$\frac{\hat{\tau}_0}{\hat{SD}} = .43$	$\frac{\hat{\tau}_1}{\hat{SD}} = .31$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_1)}}}}{\hat{SD}} = .51$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_2)}}}}{\hat{SD}} = .31$
With adjustment (i.e., using the results of the model that includes all school-level covariates and their interactions with the indicator of treatment status.)	$\frac{\hat{\tau}_0^*}{\hat{SD}} = .34$	$\frac{\hat{\tau}_1^*}{\hat{SD}} = .26$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_1^*)}}}}{\hat{SD}} = .39$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_2^*)}}}}{\hat{SD}} = .26$

Summary of Indicators of Accuracy: Reading

	Square root of variation in average performance (expressed in standard devation units of the posttest)	Square root of variation in the treatment-control difference in performance (expressed in standard deviation units of the posttest)	The square root of the estimated average MSE of $\hat{\theta}_1$ (expressed in standard deviation units of the posttest)	The square root of the estimated average MSE of $\hat{\theta}_2$ (expressed in standard deviation units of the posttest)
Without adjustment (i.e., using the results from the model with no school-level covariates)	$\frac{\hat{\tau}_0}{\hat{SD}} = .38$	$\frac{\hat{\tau}_1}{\hat{SD}} = .23$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_1)}}}}{\hat{SD}} = .56$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_2)}}}}{\hat{SD}} = .23$
With adjustment (i.e., using the results of the model that includes all school-level covariates and their interactions with the indicator of treatment status.)	$\frac{\hat{\tau}_0^*}{\hat{SD}} = .25$	$\frac{\hat{\tau}_1^*}{\hat{SD}} = .19$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_1^*)}}}}{\hat{SD}} = .41$	$\frac{\sqrt{\widehat{\overline{MSE(\hat{\theta}_2^*)}}}}{\hat{SD}} = .19$

Conclusion 2: In this multi-site trial, using results of experiments done elsewhere, on average, does not allow us to make accurate predictions about the impact at a given site. Accounting for cross-site differences based on available demographic variables, does not improve the accuracy of our estimate. The level of inaccuracy continues to be as large as effect sizes often deemed to be educationally important (>.20 sd units).
(We reject the following hypotheses: $H_0: \tau_1^{*2} < \tau_1^2$, $H_0: \tau_1^2 = 0$, $H_0: \tau_1^{*2} = 0$***)***