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

Method and Experimental Results.

Background

We use 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).

Treatment heterogeneity implies that a simple average impact estimate is inadequate as a basis for generalization; it presents an opportunity for establishing generalizability by accounting for the heterogeneity through moderator effects.

Approaches to estimating impact at site *a

<table>
<thead>
<tr>
<th>Approaches to estimating impact at site</th>
<th>Estimates of the impact at site *a</th>
<th>Average mean squared error for estimates of impact at *a *b</th>
</tr>
</thead>
</table>

We compare performance at N-1 sites (b) (other than q) to that of the average single site of interest, q, where the intervention has not been used.

We assume that a randomized trial has not been carried out at q, but that within-site results have been obtained 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.*

* a 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, i.e., the posttest. We assume a balanced design.

** *b is the estimated between-site variance in site-average performance in the absence of treatment.

Method

The use of 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.)

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.)

Results

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 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 τq^2, τt^2, and τf^2 by modeling the interaction(s) of treatment with one or more school-level covariates. The corresponding variance components from models that include one or more moderators and their interaction(s) with treatment are denoted as: τq^2, τt^2, and τf^2. The potential moderators were convenience variables—simple demographic not theoretically tied to the intervention.

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 generalizability 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).

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: H1: τq^2 ≠ t, H2: τt^2 ≠ u, H3: τf^2 ≠ v)

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 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 τq^2, τt^2, and τf^2 by modeling the interaction(s) of treatment with one or more school-level covariates. The corresponding variance components from models that include one or more moderators and their interaction(s) with treatment are denoted as: τq^2, τt^2, and τf^2. The potential moderators were convenience variables—simple demographic not theoretically tied to the intervention.

As shown in the analyses, the mean of the average site impact, τq, is 11%, that is statistically significant (p-value). Black interaction(s) between covariates and treatment is not statistically significant (p-value). Empty model: neither of these conditions hold. The model with both main and interaction effect(s) reads as a better fit than the reference model (i.e., the model without any school-level main or interaction effects)."