Symposium title: New Evidence on the Generalizability of Evidence from Impact Evaluations

Session organizer: Robert Olsen (research@robolsen.org)
Chair: Elizabeth Stuart (estuart@jhu.edu)
Discussant: Duncan Chaplin (DChaplin@mathematica-mpr.com)

First and second choice of conference section:
(1) Research methods
(2) Effects of educational policies

Symposium justification:

While impact evaluations provide evidence on causal effects for the study sample, the key question is whether the evidence allows us to accurately predict the consequences of potential policy decisions—such as whether to adopt or scale-up an intervention, or whether to cancel a program. To date, there has been some research on the theory of the problem (e.g., Olsen et al., 2013, Tipton et al., 2013) and improved designs and analysis methods to address the problem (e.g., Cole and Stuart, 2010; Kern et al., 2016, Olsen and Orr, in press; Tipton 2014). There has been little empirical evidence on whether impact evaluation evidence generalizes to broad populations—see Allcott (2015) and Bell et al. (2015)—and no evidence on whether impact estimates from evaluations generalize to local populations affected by local decision-making.

This session will offer additional evidence on the extent to which the findings from impact evaluations—in particular, impact evaluations of educational interventions—generalize to other populations. The first paper, by Fellers and Tipton, assess the generalizability of impact studies funded under grants from the Institute of Education Sciences over the last 10 years to test the effects of curricular interventions in elementary school. The second paper, by Orr, Olsen, Bell and Stuart, presents empirical evidence on the generalizability of results from several multi-site impact evaluations in education to individual sites. The third paper, by Jaime Thomas, Thomas Cook, Alice Klein, Prentice Starkey, and Lydia DeFlorio, provides evidences on whether an intervention that has shown to be effective when implemented on a small scale retains its effectiveness when “scaled up” statewide in California—and assesses the reasons for the success or failure to scale up.

References:


Paper #1

Title: Does IES funded research represent U.S. schools well? An evaluation of issues of generalizability in grant funded research between 2005-2014

Authors:

Lauren Fellers (Teachers College, Columbia University)* &

Elizabeth Tipton (Teachers College, Columbia University)

*presenting author, laf2156@tc.columbia.edu

Background:

Since its creation in 2002, the Institute of Education Sciences (IES) has aimed to increase the rigor of education research. To do so, IES has focused on increasing the number of high-quality experiments and quasi-experiments in education, with an emphasis on determining if programs and interventions cause changes in student learning. These efforts have resulted in over 100 efficacy (“Goal 3”) trials and 12 effectiveness/scale-up (“Goal 4”) studies. When implemented well results of these trials are included in the What Works Clearinghouse (WWC), allowing policy makers and practitioners to apply findings from trials to make curricular decisions in districts and schools.

Researchers and policy makers have become increasingly concerned with determining if results from educational experiments can be (or should be) generalized to different inference populations. If these generalizations are desired, how should policy makers and practitioners determine if their school is similar enough to study schools to warrant generalizations? Information on school context is rarely mentioned in WWC reports, making it difficult to determine if programs might work well in one context or another, impacting the ability of individuals to make decisions about how, where, and with which students an intervention may work.
Currently there are no studies that comprehensively review these issues of generalizability in IES grant-funded studies (however, see Stuart, et al. (2016) for a discussion of related concerns in contract research). An important question is how well current IES grant-funded studies represent the greater U.S. population of schools, as well as specific sub-populations of schools in need of high-quality evaluations (e.g. Title I schools). Furthermore, little is known collectively about recruitment practices in these studies or common barriers encountered by the field (for an exception, see Tipton et al, in press). Knowing the nature of schools who are likely, and conversely those not likely, to participate in studies could lead to better recruitment practices, better documentation of contexts, and ultimately better use of findings.

Purpose:
This work aims to address this concern with the generalizability of results from educational experiments by investigating the types of schools that participate in a subset of Goal 3 and Goal 4 trials funded by IES-NCER over a 10-year period. To do this we answer three questions:

1. How similar are participating schools (across all studies) to the population of public elementary and middle schools in the U.S., as well as to various sub-populations (e.g., Title I schools)?
2. How are issues of generalization reported and discussed in research articles and reports (thus available to policy makers and practitioners)?
3. What issues arose in recruitment, (e.g., constraints, resources, best practices) that impacted the generalizability of findings?

To answer Question 1, the sample of schools taking part in IES funded research was compared to various populations using statistical methods for generalization (e.g., Stuart et al, 2011; Tipton, 2014). To answer Question 2, a content-analysis of published articles and reports was conducted. To answer Question 3, a content-analysis of interviews with project PIs was conducted.

Population/Subjects/Participants:
Studies in this review include those funded under NCER between 2005 – 2014 (ensuring completed recruitment), focusing on interventions aimed at curricular changes for elementary school students. This leads to a population of 34 studies.

Research Design and Data Collection:
For each study, we gathered three sources of data:
- A 20-40 minute interview with project PIs regarding study recruitment processes;
- Any study related published articles (or unpublished, gathered from PIs);
- A list of participating study schools (reported by PIs).
Following IRB protocol, PIs were guaranteed that school names would not be reported in any published documents and that no analyses would be linked directly to any individual studies.
To answer Question 1 – regarding similarity between the collective study schools and various inference populations – a data frame for the population of elementary was created based upon the Common Core of Data (NCES 2015-16). Only non-charter, non-magnet, public, operational elementary schools (serving students in K-6 grades) are included. We gathered the same data on schools taking part in each study from the CCD one year prior to grant funding.
To assess similarity, we began by selecting a set of demographic covariates from the CCD. Then the combined sample of study schools were compared to the inference population using logistic regression to create a propensity score (see Stuart, et al., 2011). We calculated standardized mean differences for each covariate studied and calculated several combined metrics, including logit-standardized mean difference (Stuart et al, 2011) and the generalizability index (Tipton, 2014).

To answer Question 2 – regarding reporting standards – a content-analysis was conducted using Nvivo Pro software with 36 articles across studies included in the sample. The analysis focused on how the field reports and discusses samples and generalization. Finally, to answer Question 3, interview notes were coded and analyzed using the same content-analysis methods, focusing on the process of recruitment and sampling for studies.

Findings:

Preliminary findings (based on 19 of the 34 studies, with interviews scheduled for the remaining studies) indicate that on the 12 covariates studied, the collective sample of study schools was in fact representative of the population of elementary schools in the United States. State level analyses indicated that the schools collectively did not represent any state populations well (Question 1). Across these studies, participating schools were more often located in urban areas, had higher rates of FRL students, had more minority students enrolled, and had more total students, in both district and school, than those schools in the population of U.S. schools. Preliminary findings also suggest that reports and publications rarely provide much discussion of the sample or generalization (Question 2) and that very often recruitment does not go as planned in studies (Question 3).

Conclusions:

This study shows that some types of schools are not well represented in funded research, that PIs encounter many problems during recruitment, and that publication norms do not include the rich discussion necessary for practitioners and policy makers to make decisions regarding generalizability.

References:


Paper #2
Title: Using Rigorous Evaluation Results to Improve Local Education Policy Decisions
Authors:
Larry Orr, Bloomberg School of Public Health, The Johns Hopkins University, lorr5@jhu.edu *
Stephen Bell, Abt Associates, Stephen_Bell@abtassoc.com
Robert Olsen, Rob Olsen LLC, research@robolsen.org
Elizabeth Stuart, Bloomberg School of Public Health, The Johns Hopkins University, estuart@jhu.edu
* Presenting Author

Abstract:

Background

Many evaluations of education interventions are primarily intended to inform local education decision makers. Adoption of specific curricula or teaching techniques, provision of services to special needs students, mentoring programs, and teacher assignment are all the province of local school administrators or state officials. To help local decision makers with these choices, the Institute of Education Sciences has sponsored numerous randomized controlled trials (RCTs) to obtain rigorous evidence on the impact of educational interventions. The Investing in Innovation Grant Program has sponsored over 100 impact evaluations between 2010 and 2016 that, under their grant agreements, were expected to meet the standards of the What Works Clearinghouse (WWC). Furthermore, as of June 2016, the WWC has reported the evidence from studies “…to provide educators with the information they need to make evidence-based decisions” (front page of the WWC website).

However, when local decision makers wish to use the results of multi-site rigorous evaluations conducted elsewhere to inform policy decisions, they are faced with a question: If we were to implement locally the intervention that was tested in the evaluation, would it produce impacts similar to
those estimated in the evaluation? At this point it is unknown whether evidence from prior evaluations can be used to accurately predict what intervention impacts would be in local areas outside the study sample.

**Purpose**

This paper presents empirical evidence on how accurately we can predict the impacts of state or local adoption of educational interventions using the evidence from multi-site impact evaluations. In particular, it estimates the Mean Squared Prediction Error for individual sites and assesses whether the confidence intervals for those estimates adequately capture the uncertainty in those predictions. The goal is to assess whether evidence-based policy decisions by states and localities, using the type of impact evidence that is currently available, is likely to improve student outcomes.

The paper addresses two questions:

1. How well can schools and districts predict the impacts of adopting interventions locally on the basis of the published results from randomized trials conducted in other areas?
2. Can application of standard statistical adjustments improve the predictive power of evaluation results for this purpose?

**Data**

For the analysis, we use data from completed random assignment studies of the impacts of (1) charter schools, (2) education technology interventions, (3) Teach for America, and (4) Head Start. Each of these studies randomized either students or classrooms to receive the intervention or to a business as usual control group. As a result, each of these studies allow us to construct unbiased estimates of the impact of the intervention in each site.
Methods

To predict the impact of the intervention in individual sites, we use three approaches:

1. *Estimate the average, pooled impact for sites in the study sample.* This impact estimate is usually the main finding from an impact analysis.

2. *Estimate the impact for a subgroup in which the excluded site falls.* Some impact analyses estimate impacts for subgroups of sites. For example, if the excluded site is located in an urban area, the estimated impact for urban sites in the study sample could be used to predict the impact of the intervention in that site.

3. *Estimate an equation that models the variation in impacts across sites as a function of multiple site-level variables.* This equation is then used to predict the impact in the excluded site.

To simulate the process of making out-of-sample predictions for individual schools that did not participate in the study, we pretend that one of the sites that participated in the study did not participate (the “excluded” site), and we use the data from the remaining sites to predict the impacts of the intervention in that site. This process is repeated for each site. This approach allows us to predict the impact of adopting the intervention for each site, calculate its confidence interval, and estimate the prediction error. The prediction error is calculated by comparing the predicted impact for site j to the unbiased impact estimate for site j using only the data for that site. While this approach overestimates the prediction errors due to sampling error in the unbiased impact estimate for site j, we have developed a correction which addresses that concern.
Findings

Evidence from one of the four studies has found that the prediction errors from all of the methods are very large—too large to expect local policy decisions based on the evidence to improve student outcomes substantially. Analysis of the data from the other three studies will be used to determine how robust these conclusions are across different interventions and studies with different attributes (e.g., sample sizes, availability of rich data on potential moderators).

Conclusions

If the preliminary results are supported by the analyses of additional data sets, it would raise questions about whether multi-site impact studies in education can reliably inform local policymakers about whether to adopt an educational intervention. It would also create some urgency to considering how educational impact studies could be conducted differently to produce more accurate predictions of the impacts of adopting educational interventions locally.
Paper #3

Title: Scaling Up an Evidence-based Intervention to the State Level: A Case Study
(New Title: The Sequential Scale-up of an Evidence-based Intervention: A Case Study)

Authors and Affiliations: Jaime L. Thomas, Mathematica Policy Research; Thomas D. Cook, Northwestern University and Mathematica Policy Research; Alice Klein, WestEd; Prentice Starkey, WestEd; Lydia DeFlorio, University of Nevada, Reno

Background and Context

Policymakers face many dilemmas when choosing a program to implement at the national, state, or local level. Public health researchers have proposed a sequential approach to identifying interventions that are worthy of broader adoption, starting with small pilot studies, extending to larger efficacy and effectiveness studies, and culminating in scale-up research. Scale-up research examines the effectiveness of a previously studied intervention when it is broadly disseminated to the target population under conditions that seek to resemble what would pertain if the intervention were official policy. The present study takes an intervention, Pre-K Mathematics, that has been shown to be effective in several relatively small evaluations and asks whether it continues to be as effective when implemented at the state level. We propose a novel framework whereby five dimensions of the scale-up process might influence the effect sizes an intervention achieves: (1) settings and study participants, (2) program content and delivery, (3) timing and the counterfactual condition, (4) measurement quality, and (5) evaluation design. This framework is useful for external validity because it explicates some of the major conditions under which we might expect to replicate a study’s findings.

Objectives and Research Questions

This paper (1) provides theory to explain why effects might be smaller as scale increases, (2) tests whether the version of Pre-K Mathematics that is evaluated in this study raises average math achievement at the larger state scale, and (3) tests whether the program effects are robust across two cohorts and a wide range of student and site differences. We then (4) describe how effect sizes differ between the present study and earlier studies, and (5) identify which dimensions of scale-up are responsible for impact differences observed at different scales.

Setting and Study Sample

The baseline sample consisted of 1,373 children who were 4.5 years old on average. They came from 140 pre-kindergarten (pre-K) classrooms throughout Northern, Central, and Southern California. These classrooms represented a racially, ethnically, and linguistically diverse population of low-income families. Overall, nearly 20 percent of pre-K classrooms were in urban areas, about half were in suburban areas, and approximately one-third were in rural areas.
**Intervention**

*Pre-K Mathematics* targets the classroom and home learning environments of young children, especially those from families experiencing economic hardship. Designed to provide systematic support and foster the development of informal mathematical knowledge, the intervention consists of math activities with concrete manipulatives that teachers deliver in small groups in the pre-K classroom and that families can use one-on-one with the child at home. Teachers receive professional development to learn about early mathematical development and how to implement *Pre-K Mathematics* activities. The curriculum targets a broad range of early math concepts, including number, operations, space, geometry, and measurement. The math activities are designed to prepare children for the Common Core State Standards for Mathematics in kindergarten, and are explicitly linked to the Curriculum Focal Points proposed by the National Council of Teachers of Mathematics.

**Research Design**

This study used a cluster RCT design in which pre-K classrooms were randomly assigned to the treatment or control condition. *Pre-K Mathematics* was the intervention and the control group entailed business-as-usual instruction in pre-K. Eligible children were 4-year-olds who were age-eligible to attend kindergarten the following year, who spoke either English or Spanish, who did not have an identified developmental disability, and for whom we obtained consent to participate. Up to 12 children per pre-K classroom were selected. The intervention was implemented across two cohorts: the first began pre-K in 2013–2014 school year, and the second in the next year.

**Data Collection and Analysis**

The outcomes we examined were the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Mathematics Assessment and the Test of Early Mathematics Ability, Third Edition (TEMA-3), measured at the end of the pre-K year. Baseline measures, collected at the beginning of the pre-K year, included the ECLS-B Mathematics Assessment, the Test of Preschool Early Literacy (TOPEL), and demographic characteristics (age, gender, race/ethnicity, and language).

**Findings / Results**

At the end of the pre-K year, *Pre-K Mathematics* had a positive and significant effect on math achievement as measured by both the ECLS-B and TEMA-3. Overall effect sizes were 0.37 for ECLS-B and 0.28 for TEMA-3 ($p < 0.01$). These impacts were comparable for children in different racial/ethnic categories and across urban, suburban, and rural locations. Effect sizes differed by cohort, however: for Cohort 1 they were 0.33 for ECLS-B and 0.16 for TEMA-3, while for Cohort 2 they were 0.41 and 0.38 respectively. When we compared the results of this study and those of past efficacy and effectiveness studies of *Pre-K Mathematics* that used the TEMA-3 as an outcome measure, a downward trend in effect sizes was observed. Comparing the TEMA-3 results by
cohort, however, the Cohort 2 result was similar to the past results but the Cohort 1 impact was smaller.

**Conclusions**

The hypothesis this study set out to test is that scale-up lowers achievement gains but demonstrates these smaller gains for a larger and more heterogeneous population than in earlier research. Positive benefits of *Pre-K Mathematics* continued to be manifest even after scale-up to the state level, and the TEMA-3 results across the two cohorts were indeed smaller than in past studies. However, the smaller impact was due to only one of the two cohorts, and the cohort difference in impact cannot be readily ascribed to the organizational learning reasons postulated in the theory guiding this paper. The best, but only partially validated, explanation we can offer is that lower performing children may have learned more from *Pre-K Mathematics* and that, for no intended reason, Cohort 2 happened to have lower math performance even before the program began.