External Validity in U.S. Education Research

Sean Tanner

Learning Policy Institute

Abstract

As methods for internal validity improve, methodological concerns have shifted toward assessing how well the research community can extrapolate from individual studies. Under recent federal granting initiatives, over $1 billion has been awarded to education programs that have been validated by a single randomized or natural experiment. If these experiments have weak external validity, scientific advancement is delayed and federal education funding might be squandered. By analyzing trials clustered within interventions, this research describes how well a single study's results are predicted by additional studies of the same intervention in addition to analyzing how well study samples match the target populations of interventions. I find that U.S. education trials are conducted on samples of students who are systematically less white and more socioeconomically disadvantaged that the overall student population. Moreover, I find that effect sizes tend to decay in the second and third trials of interventions.

I. Introduction and Prior Literature

As randomized and natural experiments continue to gain acceptance as the preferred standards for causal inference, the locus of methodological concern in applied social science is shifting toward external validity (Cook, 2014; Orr, 2015). An abiding concern is that results from these methods are local to the setting and population in which they are conducted, with a consequent lack of generalizability masked by a veneer of rigor. Scholars from fields as disparate as economics (Wolpin, 2011) and philosophy (Cartwright, 2007, Chapters 15–16) have commented specifically on the insufficiency of randomized trials (RCTs) alone to inform critical policy decisions. Cook (cite) offers a conceptual model for generalizing causal knowledge,

noting challenges of representation and extrapolation at the levels of units, treatments, outcome measures, settings, and time- features of research that are largely orthogonal to the details of internal validity. Without careful attention to these challenges, even the strongest inferential strategies faces grave epistemological shortcomings.

In education research these concerns are not merely academic, as the policy consequences of randomized and natural experiments can hinge crucially on scale and context. The oft-cited Tennessee STAR experiment improved the educational achievement of low-performing students through reduced pupil-teacher ratios (Schanzenbach, 2006), yet such gains eluded California students when the state implemented similar changes at a larger scale (Stecher, Bohrnstedt, Kirst, McRobbie, & Williams, 2001). Conversely, five field experiments of performance pay for teachers fail to find any effect of linking salary bonuses to student test scores, yet a recent natural experiment has revealed that such coupling of teacher salary and student test scores can have marked effects on subsequent teacher and student performance (Dee & Wyckoff, 2015). More broadly, for three of the four educational interventions awarded an Investing in Innovation Initiative (i3) scale-up grant in 2010 (KIPP, Success for All, Teach for America, Reading Recovery[1]), impacts on students were substantially lower than impacts from the initial experimental trials of each intervention (Clark, Isenberg, Liu, Makowsky, & Zukiewicz, 2015; May, Sirinides, Gray, & Goldsworthy, 2016; Quint, Zhu, Balu, Rappaport, & DeLaurentis, 2015; Tuttle et al., 2015). As these examples illustrate, extrapolating from a single study can be a dubious enterprise without careful attention to external validity.

Unfortunately, little is currently known about the external validity of rigorous education research in the United States. Recent evidence from impact evaluations in the developing world raises concerns about lack of protocol fidelity as well as poor external validity (Vivalt, 2015), while a study of U.S.-based energy conservation experiments reveals that even multiple RCTs do not yield average treatment effects that correctly predict future experimental results (Allcott, 2015). The limited evidence from domestic education research is not reassuring. An analysis of U.S. school districts that have participated in eleven randomized control trials of

[1] See http://www2.ed.gov/programs/innovation/awards.html

educational interventions found that those districts are poorer and more ethnically diverse, have a greater number of English language learners, and perform worse on standardized tests of literacy and numeracy than the national average of districts that could potentially implement any of the eleven interventions (Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2016). An additional analysis of those participating districts finds that they responded more weakly to a federal policy intervention than the typical district, suggesting a correlation between study participation and potential outcomes (Bell, Olsen, Orr, & Stuart, 2016). This is troubling in a research environment where a single trial can trigger hefty public investment in an intervention and shape legislation (Bogenschneider & Corbett, 2010; Haskins & Margolis, 2015).

I contribute to this literature by providing the first extensive assessment of external validity in U.S. education research by meta-analyzing over 300 education-related interventions conducted in the United States in the past two decades. Using the newly available data on interventions and trials from the Institute for Education Sciences' What Works Clearinghouse – a repository of rigorous random and natural experiments maintained by the U.S. Department of Education – I address the following two primary research questions: (1) how representative of U.S. students are the samples selected for participation in U.S. trials of education interventions and (2) how much can be extrapolated from these trials. Within the endeavor to generalize causal knowledge, the first question addresses the *representation function* (Cook, 2014, p. 527) by investigating the degree to which trial samples match the target populations of the intervention being studied along numerous dimensions such as geographic location, student demographic profiles, school financial resources, racial and socioeconomic segregation, and student achievement. The second question addresses the *extrapolation function* (Cook, 2014, p. 527) by analyzing how well a single trial's results are predicted by additional trials of the same intervention. Using trials clustered within interventions, a Bayesian prediction framework will provide summary measures of precision across trials. The analysis will also leverage the temporal ordering of the trials to determine whether or not effects decay in subsequent trials of an initially promising intervention, as anecdotal evidence suggests is common.

The results of this study will provide a comprehensive description of external validity in recent U.S. education research while also suggesting features of research on which the scientific community and funding agencies should focus.

To preview the results, I find that study samples are systematically unrepresentative of students in the national as a whole. Education trials are conducted on student samples that are less white, more socioeconomically disadvantaged, and more likely to have limited English proficiency than the national average of districts. Moreover, effect sizes exhibit decay over subsequent trials for interventions subject to two or three trials, though the pattern is less clear for interventions subject to between four and eight trials.

In the remainder of the paper, I present the analytic framework and describe the newly available What Works Clearinghouse data before moving on the results and a discussion of their implications.

II. Analysis

The analysis is structured around two primary questions: (1) how representative are studies of education interventions; and (2) how much can one extrapolate from a single study. Two complementary sets of analyses will be undertaken to evaluate the representativeness of the WWC trial samples. The first set assesses the degree of demographic similarity between the WWC trial samples and the population of students targeted by the interventions therein. The second set inverts the first by analyzing the relative achievement and socioeconomic conditions of the students in school districts that are demographically similar to the WWC trial samples. Operationally, similarity is measured as the standardized difference between each trial sample's characteristics and the mean of the population of schools and districts for each demographic variable (race, socioeconomic status, English language ability, individualized education plan), as well as the multidimensional distance between each trial sample's demographic composition and the multivariate centroid of the population of schools and districts. As sample compositions differ across outcomes within a trial, the median of each demographic variable across outcomes for that trial is used. Equation one presents the standardized, univariate distance measures calculated for each demographic variable in each

trial, where D is the k^{th} demographic variable in the i^{th} study, μ is the enrollment-weighted population mean of the k^{th} demographic variable, and σ is the population standard deviation of the k^{th} demographic variable.

$$s_{ik} = {(D_{ik} - \mu_k)}/{\sigma_k} \quad (1)$$

Equation two presents the Mahalanobis distance of a single trial's demographic composition from the centroid of the target population's demographic variables. After standardizing the demographic variables with equation one and setting the population means equal to zero, the Mahalanobis distance reduces to the following:

$$M_i = \sqrt{\sum_{i=1}^{n} s_{ik}^2} \quad (2)$$

S_{ik} and M_i are also calculated for each school and district in the SEDA and CCD. Each WWC trial receives corresponding scores π_{ik} and λ_i equal to the proportion of schools and districts in the population with absolute values of s_{ik} and M_i larger than their own. The primary summary statistics are $\bar{\pi}_k$ and $\bar{\lambda}$, the means of π_{ik} for each of k demographic variables and the mean of λ_i respectively. These summary statistics provide scale-invariant measures of how aberrant the trial samples are from the population of schools and districts.

Classical tests for the statistical significance of $\bar{\pi}_j$ and $\bar{\lambda}$ are inappropriate, as these trial samples are not likely to have been sampled randomly from the population of schools (Stuart et al., 2016, pp. 3–5). Instead, to provide a measure of how likely these values are to have arrived by a random sample of schools and districts, a bootstrap procedure is used by taking 5000 random samples (with replacement) of n schools and districts and computing $\bar{\pi}_j$ and $\bar{\lambda}$ for each sample. The number n of schools and districts selected are chosen to match the number of trials in the WWC appropriate for each demographic test. The resulting p-values are the proportion of bootstrapped summary statistics that are at least as large as the corresponding statistics from the WWC studies. In order to control the false discovery rate across multiple hypothesis tests, the entire set of p-values from this set of tests (group A in table 1) are transformed into sharpened q-values (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001).

In the second set of analyses, M_i values are used to identify districts with demographic characteristics similar to those of the WWC trial samples. Specifically, each trial is assigned a set of three districts from the SEDA whose M_i values are the closest to that trial's. The resulting set of matched districts serve as the represented population and are compared to the general target population of districts using the full SEDA data. Using the newly available SEDA rather than the CCD allows national, district-level comparisons to be made on the basis of standardized performance of literacy and numeracy, gaps in performance by ethnicity and socioeconomic status, the degree of segregation within the district, and socioeconomic conditions of the area in which the district resides. The key variables of interest from the SEDA are listed in table 3. The same procedures for calculating summary and test statistics from the first set of analyses are employed here as well, with subscripts s to denote use of the SEDA data. All p-values from this second set of tests (group B in table 1) are transformed into sharpened q-values.

The second primary phase of the analysis leverages the 91 multi-trial interventions from the WWC flat file to evaluate how well trials predict one another within interventions. The motivation for this section of the research is the extrapolation problem that frequently bedevils scholars, program funders, and policymakers alike: given all available information on prior trials of an intervention, what should one expect from an additional trial? The first primary question in this fundamentally Bayesian inquiry is whether effect sizes exhibit systematic decay over subsequent trials. Decay may result from difficulties scaling up interventions (Clark et al., 2015; Quint et al., 2015; Tuttle et al., 2015), expanding them to a group of students less amenable to the treatment,[2] or an increase in the performance of the counterfactual group (Lemons, Fuchs, Gilbert, & Fuchs, 2014). More cynically, decay may be the result of selection bias in the choice of which interventions receive subsequent trials. In this rendering, initially promising trials are merely the right tail of a sampling distribution centered closer to zero and thus any subsequent trial should be expected to revert to the mean (Ioannidis, 2005). The second primary question concerns the variance of trials within interventions. High variance may be the result of

[2] This is a perennial concern in early childhood education research. See Weiland and Yoshikawa (2013, pp. 2126–2127) for a brief discussion of impact estimates of pre-kindergarten across socioeconomic and ethnic groups.

heterogeneous treatment effects between groups or contexts that differ across trials (Feller, Grindal, Miratrix, & Page, 2016), or merely poor protocol fidelity (Ginsburg & Smith, 2016, pp. 8–11). Even in the absence of systematic decay, high variance across trials will limit the ability to draw strong inference around average treatment effects and their attendant social welfare and policy implications. Operationally, I analyze which study variables best predict effect size.

The estimation procedures begin, somewhat simply, by calculating, for each multi-trial intervention in the WWC file (n=85), the percent difference between the effect size of each trial and that of the prior trials of the same intervention on the same outcome domain. Most trials assess the impact of an intervention on more than one outcome domain, such as literacy and numeracy, and subsequent trials often use a slightly different measure (ex. SAT vs. ACT scores) within the same outcome domain. As one purpose of this analysis is to assess the stability of results across trials, the primary linkage of trials within interventions is outcome domains rather than the narrower outcome measures. The WWC assigns a domain to each outcome assessed in each trial, yet the full list includes domain labels that are broad enough to contain meaningfully different outcomes (ex. "labor market outcomes") as well as multiple domain labels for what appear to be similar outcomes (ex. "English language development" and "language development"). These "jingle" and "jangle" fallacies, respectively, are well known to scholars of child development and frustrate the task of causal generalization across trials. A trade-off must be made between overly specific domain labels that risk obscuring variance across similar outcomes and uselessly broad domain labels that are unable to differentiate among substantively distinct outcomes, thus inflating inter-trial variance. The WWC domain labels are used as a guide, but recoding does occur in some cases.

Equation 3 presents the formal calculation of the change, Δ_{ij}, where θ_{ij} is the effect size from the i^{th} trial of the j^{th} intervention, $\mu_{i-1,\,j}$ is the effect size from the prior trial of the j^{th} intervention.

$$\Delta_{ij} = \theta_i - \theta_{i-1,j} \qquad (3)$$

Interventions with at least two trials of the same outcome domain receive a value of Δ_{ij} for each trial, from the second to the n^{th}. The summary statistics of interest are the means and medians,

$\bar{\Delta}_i$ and $\tilde{\Delta}_i$ respectively, of Δ_{ij}, from i=2 to i=n. The statistical significance of $\bar{\Delta}_i$ and $\tilde{\Delta}_i$ will be assessed through permutation tests based on 5000 draws of the data wherein the order of trials within interventions is randomly assigned.

III. Sample Selection

The analysis sample is drawn from three data sets: the What Works Clearinghouse's (WWC) summary file of education interventions, the National Center for Education Statistics' Common Core of Data (CCD) files, and the Stanford Education Data Archive (SEDA) of district characteristics and student achievement. The WWC file contains data on study characteristics (year, intervention, outcome measure, domain, setting, geographic location), average student characteristics (grade-level, gender, race, socioeconomic status, English language ability), and statistical test information (sample sizes, post-treatment means and standard deviations by treatment status, effect size, statistical significance). The CCD files contain district- and school-level demographic and financial information for all public K-12 schools in the United States. The SEDA files contain district-level data on student achievement (means and gaps on a common scale by year-grade-demographic group), district characteristics (geographic location, segregation by race and socioeconomic status), and socioeconomic data for the region in which the school district is located (education levels, employment, SNAP participation, household income levels and inequality).

The WWC data is the focus of the analysis, whereas the SEDA and CCD data are used to compose the universe of schools and districts from which intervention trial samples are drawn. The units of analyses are individual studies of education interventions as well as the interventions themselves (many of which have been subject to multiple trials). The interventions and studies used are those contained in the WWC data which involved elementary and secondary students in the United States from 1998 to 2016. As can be seen in table 1, the WWC data contains tests ("findings") from 534 trials of 309 interventions. Most are randomized control trials, but a substantial portion come from natural experiments. Regardless of the inferential method, all trials analyzed here have been given one of the top two ratings in the WWC's evidence review system and thus meet the Institute of Educational Sciences standards for causal evidence. As table 2 reveals, 85 of the interventions have been subjected

to more than one trial, while 40 interventions have been subjected to at least three trials. The publication dates of the trials are contained in figure 2. Figure 1 contains the distribution of effect sizes for the WWC trials. Each results from the trials contained in the WWC is standardized into the effect of the treatment on the outcome in terms of the outcomes standard deviations, Cohen's D. The trial-weighted mean effect in this data is .24, with a standard deviation of .37.

V. Results

To begin with the representation function, I find that study samples are less white, more socioeconomically disadvantaged, and less proficient in English than districts as a whole. Table 3 lists simple comparisons of the proportion of each demographic group in the districts contained in the Common Core of Data (CCD) and the trials in the WWC. The main results are presented in table 4. The student composition in the WWC trials are in the upper tercile of proportion Hispanic and free-lunch eligibility (the common metric of socioeconomic disadvantage). For the rest of the demographic variables, the WWC trials are in the upper quintile district, meaning that over 80% of districts in the United States have demographic profiles closer to the national average than the WWC trials. This raises serious concerns about the generalizability of the rigorous causal information contained in the WWC.

Moving to the extrapolation function, I find that effect sizes exhibit systematic decay in the first two follow-up trials, but am unable to distinguish a pattern in further trials. Figure 3 displays the pattern of results for interventions subject to more than one trial. The mean and median effect in the second trial decrease by .097 and .042 standard deviations, respectively. The mean and median effect decrease even more by the third trial, to .121 and .057 standard deviations less than the original effects. All results from the second and third trial are distinguishable from random noise, however I am unable to do so for the rest of the follow-up trials. This is largely due to the diminishing number of interventions subject to further trials.

While these results point to serious challenges in generalizing causal knowledge from rigorous trials of education interventions in the U.S., several questions are yet to be answered. First, while demographic variables are likely correlates of important moderators and mediators for

education interventions, it is still unclear whether or not trial samples are aberrant on those elements themselves. Second, further work should be done to examine, to the extent possible, why effect sizes decay and whether this is attributable to a difference in outcomes, sample characteristics, settings, or the like.

Table 1: Trial Data in the What Works Clearinghouse

WWC Data	RCT	QED	Total
Interventions	246	101	309
Trials	393	141	534
Findings	1921	682	2603
Intervention-Outcome Matches	399	148	506

Table 2: Multi-Trial Interventions in the What Works Clearinghouse

Number of Trials	Number of Interventions	Intervention-Outcomes
1	268	409
2	45	60
3	18	21
4	6	7
5	10	4
6	2	2
7	3	2
8	1	1

Figure 1: All Effects from the WWC data

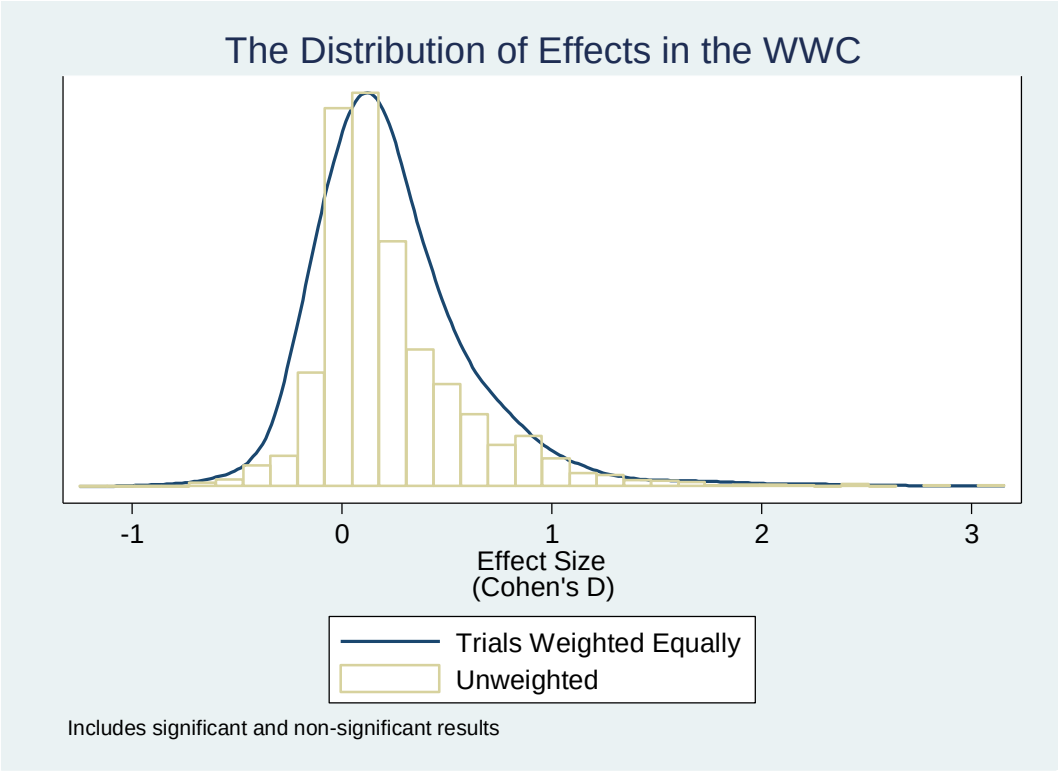


Table 3: WWC vs. CCD

Demographic Mean Proportion (SD)	Districts in U.S.	WWC Trials	p-value
African-American	.16822498 (.20928031)	.28700856 (.0723077)	<0.001
Hispanic	.20653194 (.24377424)	.27605177 (.09985937)	<0.001
Asian	.04707013 (.07947539)	.03306675 (.02224536)	<0.001
White	.55917652 (.31147693)	.39372461 (.10072037)	<0.001
Eligible for Free Lunch	.44058706 (.23065425)	.55955783 (.15323555)	<0.001
English Language Learner	.09537689 (.11591922)	.06224852 (.05549068)	<0.001

Table 4: Poor Representation in WWC Trials

Demographic Variable	Proportion of Districts more Aberrant than WWC Trials
African-American	.117
Hispanic	.332
Asian	.112
White	.172
Eligible for Free Lunch	.369
English Language Learner	.145

Figure 2: Tests in the WWC

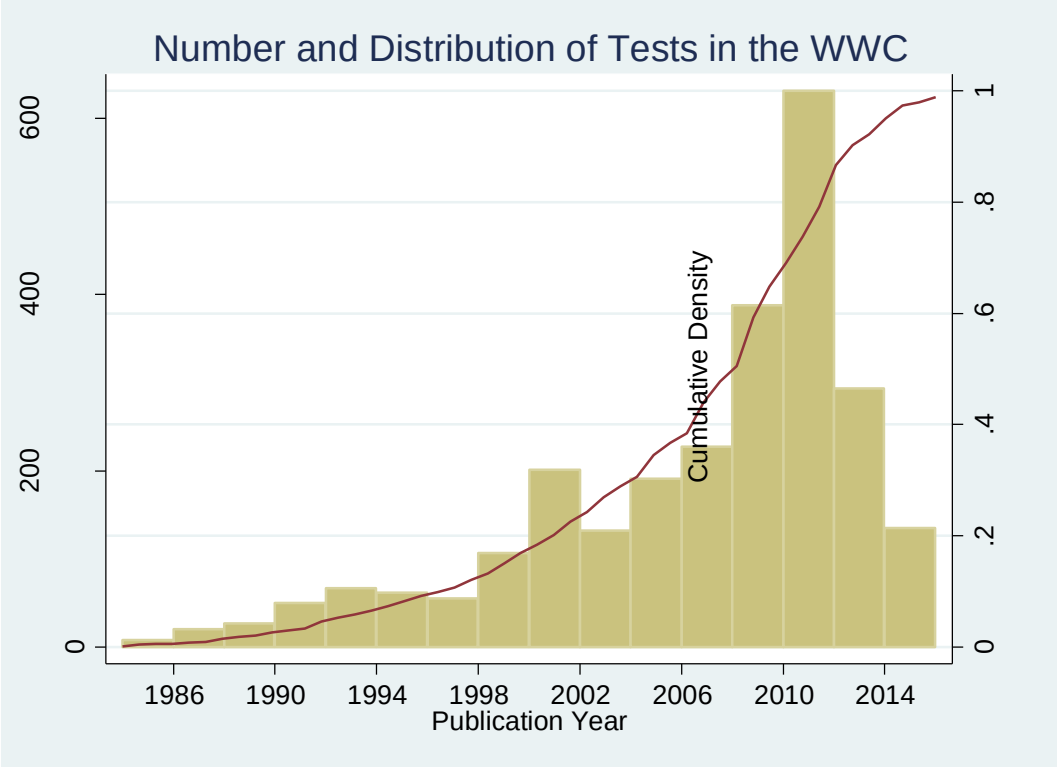
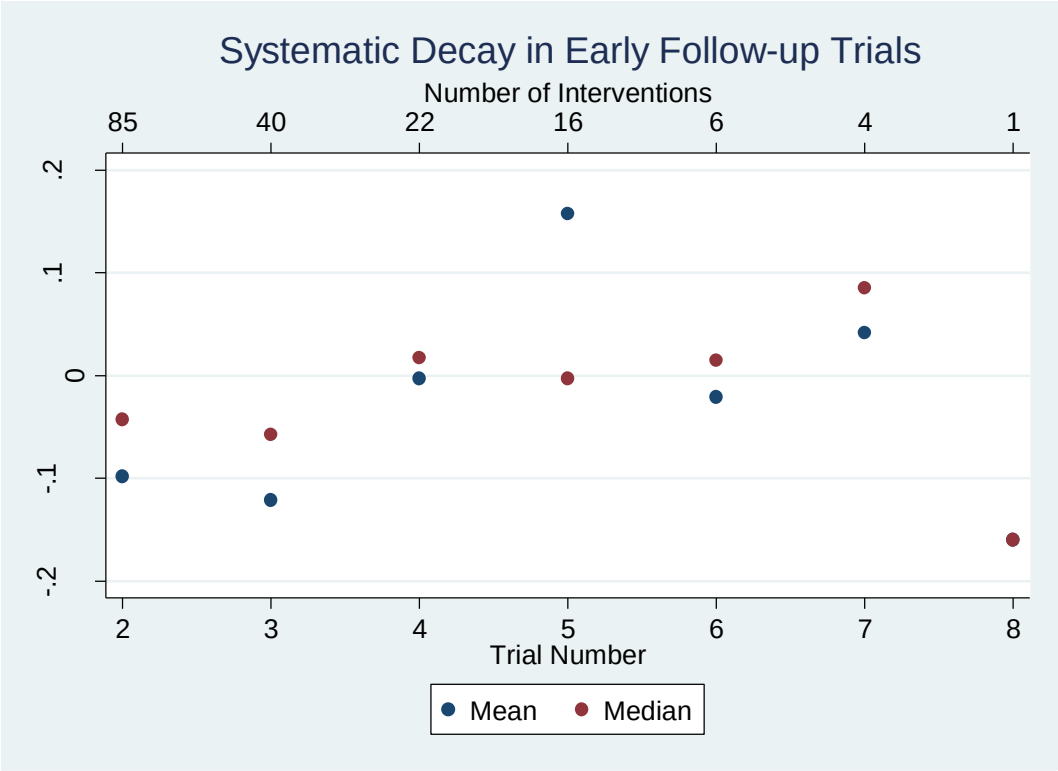


Figure 3: Systematic Decay in Early Follow-up Trials



Allcott, H. (2015). SITE SELECTION BIAS IN PROGRAM EVALUATION. *Quarterly Journal of Economics*, *130*(3), 1117–1165. http://doi.org/10.1093/qje/qjv015.Advance

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of External Validity Bias When Impact Evaluations Select Sites Purposively. *Educational Evaluation and Policy Analysis*, *38*(1), 1–18. http://doi.org/10.3102/0162373715617549

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Bogenschneider, K., & Corbett, T. (2010). *Evidence-Based Policymaking: Insights from Policy-Minded Researchers and Research-Minded Policymakers*. New York: Routledge.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge: Cambridge University Press.

Clark, M. A., Isenberg, E., Liu, A. Y., Makowsky, L., & Zukiewicz, M. (2015). *Implementation of the Teach for Amercia Investing in Innovation Scale-Up*. Washington, D.C.

Cook, T. D. (2014). GENERALIZING CAUSAL KNOWLEDGE IN THE POLICY SCIENCES: EXTERNAL VALIDITY AS A TASK OF BOTH MULTIATTRIBUTE REPRESENTATION AND MULTIATTRIBUTE EXTRAPOLATION. *Journal of Policy Analysis and Management*, *33*(2), 527–536.

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297. http://doi.org/10.1002/pam

Feller, A., Grindal, T., Miratrix, L., & Page, L. (2016). *Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Settings*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2534811

Ginsburg, A., & Smith, M. S. (2016). *Do Randomized Controlled Trials Meet the " Gold Standard "? A Study of the Usefulness of RCTs in the What Works Clearinghouse*.

Haskins, R., & Margolis, G. (2015). *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. Washington, D.C.: Brookings Institution Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8).

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-Based Practices in a Changing World: Reconsidering the Counterfactual in Education Research. *Educational Researcher*, *53*(5), 242–252. http://doi.org/10.3102/0013189X14539189

May, H., Sirinides, P., Gray, A., & Goldsworthy, H. (2016). *Reading Recovery : An Evaluation of the Four-Year i3 Scale-Up Reading Recovery : An Evaluation of the Four-Year i3 Scale-Up*. Philadelphia, PA.

Orr, L. L. (2015). 2014 Rossi Award Lecture: Beyond Internal Validity. *Evaluation Review*, *39*(2), 167–178. http://doi.org/10.1177/0193841X15573659

Quint, J., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015). *SCALING UP THE SUCCESS FOR ALL MODEL OF SCHOOL REFORM: Final Report from the Investing in Innovation (i3) Evaluation*. New York.

Schanzenbach, D. W. (2006). *What Have Researchers Learned from Project STAR?* (2006/2007 No. 9). *Brookings Papers on Education Policy* (Vol. 2006). Was.

Stecher, B., Bohrnstedt, G., Kirst, M., McRobbie, J., & Williams, T. (2001). Class Size Reduction in California: A Story of Hope, Promise, and Unitended Consequences. *Phi Delta Kappan*, *89*(2), 670–674.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2016). Characteristics of School Districts That Participate in Rigorous National Educational Evaluations. *Journal of Research on Educational Effectiveness*, *5747*(August). http://doi.org/10.1080/19345747.2016.1205160

Tuttle, C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., … Goble, L. (2015). *Understanding the Effect of KIPP as it Scales: Volume 1, Impacts on Achievement and Other Outcomes*. Washington, D.C.

Vivalt, E. (2015). *How Much Can We Generalize from Impact Evaluations?* (Unpublished Manuscript). Retrieved from http://evavivalt.com/wp-content/uploads/2014/12/Vivalt-JMP-latest.pdf

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112–2130. http://doi.org/10.1111/cdev.12099

Wolpin, K. I. (2011). *The Limits of Inference without Theory*. Cambridge, MA: MIT Press.