Title: The past, present, and future of recruitment and generalization in education

Organizer: Elizabeth Tipton (tipton@northwestern.edu)
Discussant: Vivian Wong (vcw2n@virginia.edu)

Paper presentations: (* denotes presenting author)

1. Elizabeth Tipton*, Qian Wang, & Jessaca Spybrook. “Assessing the Relevance of IES Funded Goal 3 and 4 Studies to Important Policy Populations.”


Overview:
The ability to recruit schools and school districts and to secure their commitment to an evaluation is fundamental to the ability to conduct an RCT in education. For example, IES Goal 3 and 4 grant proposals require that researchers provide evidence (e.g. letters of commitment) that at least some schools have already agreed to take part in a study. Despite the importance of this recruitment, very little information on recruitment is provided in written reports and publications regarding evaluations (Fellers, 2017). Information on the types of schools that took part in a study and the target population to which results can (cannot) generalize are equally under-reported (Fellers, 2017). Disciplinary journals rarely require information on recruitment, sample characteristics, or target populations, and even the WWC provides little detail on these features.

At the same time, information on the sample – on the types and characteristics of schools that were actually involved in the evaluation – is essential to policymakers and practitioners seeking to interpret findings and make decisions about the potential use of an intervention in their own school or school districts. This is particularly true given recent evidence that schools and school districts included in some RCTs may not well represent important target populations (e.g., Fellers, 2017; Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017; Tipton et al., 2016).

In this session, we take a ‘longitudinal’ approach to the role of recruitment and reporting in RCTs in education. The first paper (‘past’) provides a review of the types of schools that took part in Goal 3 and 4 evaluations funded by IES between 2011 – 2015. This review includes information on 36 RCTs and the over 1,000 schools included in these studies, with a focus on revealing types of schools that are over- or under-represented in funded research. The second paper (‘present’) provides an evaluation of possible recruitment (sample-selection) methods to better understand how different approaches
might affect the ability to generalize. These include standard practices (convenience samples), recently proposed selection methods (stratified-balanced-sampling), and simple random sampling. The third paper (‘future’) provides suggestions for how IES might integrate sample and population data into the WWC, providing researchers with the tools they need to understand where results might apply best. The authors note that a by-product of such reporting may be that it encourages researchers to take these guidelines into account when planning their own studies, thus improving recruitment practices.

**References**


Paper #1
Assessing the Relevance of IES Funded Goal 3 and 4 Studies to Important Policy Populations

Background
There is increasing evidence that the impact of interventions depend on context and student characteristics (e.g., Weiss et al., 2017). At the same time, the vast majority of Goal 3 and Goal 4 randomized control trials are conducted in convenience samples, with samples based on proximity to researcher locations. This disconnect is important for policy, particularly given that there is evidence that the ATEs estimated in these convenience samples differ from population ATEs by as much as 0.10 standard deviations (Stuart, Olsen, Bell, & Orr, 2012). Thus better understanding the characteristics of schools that take part in IES funded RCTs – or in education RCTs more generally – is increasingly important (Fellers, 2017; Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017; Tipton et al., 2016).

Purpose
The goal of this study is to better understand the degree to which schools taking part in funded randomized controlled trials (RCTs) represent well various populations of schools nationwide. Our goal is to understand both the extent to which individual studies represent these populations well and the extent to which the overall field of research collectively represents these populations. We seek to both better understand current practice and to uncover populations that are not well-studied, upon which future research can build.

Methods
Populations. Educators and policy-makers often wish to understand the extent to which research findings apply to their schools and contexts. As a result, there are many possible populations of interest. In this paper, we focus on a handful of these based upon the goals of IES and the education research community. These include populations defined based upon:

- Urbancity: Rural, Urban, and Suburban
- School district size: Large, Medium, and Small
- Title 1 status: Schoolwide Title 1, Not

For each of these, we separate out each by elementary, middle, and /high school and intervention type. Population frames for each population were created with data from the Common Core of Data (NCES).

Sample. We focus on a sample of IES funded Goal 3 and 4 studies conducted between 2011 – 2015. Studies were selected if they recruited K – 12 schools or districts (and excluded if they were partnership studies) and randomized schools, teachers, or classrooms to an intervention. These inclusion criteria resulted in 41 intervention studies distributed across 33 unique Principal Investigators (PIs). We contacted each PI and asked to interview them regarding recruitment in their studies. Overall, we conducted interviews regarding recruitment in 38 of the studies (3 study PIs did not respond to requests).

Additionally, we requested that PIs share the school NCES IDs for each of the schools in their sample. Out of these 41 studies, we were able to obtain data from 31, with data from 5 more
studies promised to us (a total of 88% of studies). While we are still processing data, our current database of clean data includes 23 studies with 1,004 schools from 418 school districts that took part in one or more RCTs. Each of these schools were located within the Common Core of Data in the year prior to recruitment; in some cases, data from the year of recruitment was necessary (if the school was new).

While we collected both qualitative data (on recruitment) and quantitative data (schools), this presentation will only focus on the latter. Future presentations will focus on the former.

Assessment of similarity/ generalizability. In order to assess the degree to which the overall sample of schools represent well each population, we conducted several comparisons based on a set of 11 covariates (see Table 1). We calculated: 1) Covariate level measures of similarity using the absolute standardized mean difference (ASMD); and 2) Overall assessments of similarity via the generalizability index (Tipton, 2014) and the logit standardized mean difference (Stuart et. al, 2011). These latter indices are measures of distributional similarity between the distribution of sampling propensity scores (based on these 11 covariates) in the total sample of schools and in the population. The generalizability index takes values between 0 and 1 and, when multiplied by 100%, can be interpreted as the % similarity between the sample and population. Samples with similarity greater than about 90 – 95% are considered as similar to the population as that of a random sample, while those less than 50% similar are considered so different that any generalizations would require extrapolations (see Tipton, 2014). In addition, we will make comparisons between each study sample and population, summarizing their degree of similarity via the generalizability index.

Results
For the purposes of this proposal, we focus on a single population and on covariate level comparisons: Title 1 schools. In Table 1, we provide comparisons between the sample and population of elementary schools and the population of middle and high schools. More study schools are in urban areas (48%) and less of them are in town/rural areas (19%) than the population of Title 1 elementary schools nationwide (35% and 36% respectively). Study schools on average included 32% white students, with the population containing 45% in Title 1 elementary schools. For middle/high schools, those sample schools were smaller than those in the population, with a larger percentage of white students (49%) and a smaller percentage of black students (19%) than the Title 1 population. In addition, 30% of those middle/high schools were located in urban areas, while there are 47% of urban schools in population.

Conclusions
In the analyses reported in this proposal, we found that schools that have participated in IES funded trials differ from the population of Title 1 schools in terms of urbanicity, school size, and racial composition. In our paper and presentation, we will report similar analyses for other important populations. We will then use these analyses as a springboard for discussions of future Goal 3 and 4 studies, and how to use prior evidence of this type to inform recruitment to fill gaps.

1 In 22% of studies, the PIs were not able to directly provide us with a list of schools. Instead, they pulled the data from the CCD for us and provided us a list of demographics which we then used in our analyses.
References


Appendix

Table 1. Preliminary comparisons between overall sample of schools and Title 1 schools nationwide.

<table>
<thead>
<tr>
<th></th>
<th>Elementary</th>
<th>Middle/High</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Study School n = 467</td>
<td>Title I Schools N = 34,570</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Total Students</td>
<td>556.89</td>
<td>225.57</td>
</tr>
<tr>
<td>Pr. Urban</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td>Pr. Suburban</td>
<td>0.33</td>
<td>0.47</td>
</tr>
<tr>
<td>Pr. Town/Rural</td>
<td>0.19</td>
<td>0.39</td>
</tr>
<tr>
<td>Pr. White</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>Pr. Black</td>
<td>0.23</td>
<td>0.34</td>
</tr>
<tr>
<td>Pr. Hispanic</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>Pr. Other</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Pr. Female</td>
<td>0.49</td>
<td>0.02</td>
</tr>
<tr>
<td>Pr. FRL</td>
<td>0.63</td>
<td>0.28</td>
</tr>
<tr>
<td>Stud/Tch Ratio</td>
<td>18.31</td>
<td>4.64</td>
</tr>
</tbody>
</table>

Note: Pr. = proportion; FRL = Free and reduced-price lunch; ELL = English language learners; ASMD = Absolute standardized mean different (values larger than 0.25 are bolded).
Assessing sampling methods for generalization from RCTs: Modeling recruitment and participation

INTRODUCTION

Generalizability of randomized control trials (RCTs) has recently come under scrutiny. RCT external validity is limited if (1) study units don’t represent the population and (2) treatment effects vary across population units. Both conditions are plausible in large-scale evaluations of educational programs, where recruitment is often driven by convenience and cost effectiveness concerns rather than population representation. The potential result is biased estimates of population average treatment effects from unrepresentative samples. Post hoc statistical methods have been proposed to improve study generalizability, but their effectiveness is limited if segments of the population are unrepresented in the sample (i.e., coverage errors).

A series of recent papers advocate planning for generalizability at the recruitment stage (Tipton, 2013a; Tipton, 2013b). One method for doing so, Stratified Balance Sampling (SBS), has attracted attention from researchers. First, extant data is used to define a target population and create a sampling frame. Using cluster analysis, population units are then stratified on a set of observed covariates related to treatment heterogeneity. Units are ranked within each strata in order of similarity to the average unit in that strata. These rankings are then used to guide recruitment.

Potential advantages of SBS include reducing coverage errors and greater recruitment transparency. However, little methodological work has examined this method’s effectiveness. Furthermore, the additional resources required to recruit from all strata create concerns regarding practicality. Schools with certain characteristics are unlikely to participate in large-scale RCTs (Stuart et al., 2017; Tipton et al., 2016). If one or more strata are comprised of difficult schools, researchers may resort to convenience sampling within those strata.

In this study, we aim to lay a groundwork for studying sampling methods in the context of educational RCTs by developing a model for the processes of sampling and recruitment. We use the model in a Monte Carlo simulation to investigate the effectiveness and feasibility of implementing Stratified Balanced Sampling (SBS) relative to Simple Random Sampling (SRS), Convenience Sampling (CS), and stratified versions of these models (SSRS, SCS). In both cases, SRS represents a theoretically ideal sampling technique while CS represents the baseline of current practice. Our study is guided by two primary questions:

**RQ1**: How do samples recruited using various methods compare in terms of generalizability and feasibility?

**RQ2**: How are generalizability and feasibility affected by the population’s overall willingness to participate?

METHOD

We designed a Monte Carlo simulation that aims to emulate recruiting a representative sample of 60 schools for a large-scale efficacy trial. Elementary schools from 5 states served as the population of interest. We created a sampling frame of schools using information from three sources: the 2014-2015 Common Core of Data (CCD) for school-level demographics; (2) accountability data from State DOE websites; and (3) the 2010 U.S. Census for local median income.

We modeled each school’s propensity to participate, or the probability of participating if approached, as a linear combination of the covariates on the logit scale. We manipulated the
overall participation rate from 10% to 90%. Coefficients in the participation propensity model were tuned to generate a specified overall participation rate while maintaining characteristics of participating samples as described in Tipton et al. (2016). School participation for all population units was generated using the propensity to participate. We then simulated recruitment under 5 different sampling schemes. In SRS and SSRS, schools were sampled with equal probability. In CS and SCS, schools that were more likely to participate were more likely to be sampled. In the stratified methods, schools were grouped into strata and sampled until a proportionally representative number of schools was selected from each strata. In SBS, schools were grouped into strata and sampled in ascending order by distance from the stratum centroid.

RESULTS

Figure 1 displays a summary of the achieved generalizability of each sampling method, as measured using the average B-index (Tipton, 2014), a measure of overlap between the sample and the population that ranges from 0 (no overlap) to 1 (complete overlap). SBS and SSRS resulted in highly generalizable samples overall. SRS and SCS resulted in less generalizable samples, especially when the overall population participation rate was below 50%. CS resulted in the least generalizable samples.

Figure 2 displays the average standardized mean differences between the sample and population for each covariate. Again, SBS generally outperformed other models, followed by SSRS, SCS, SRS, and CS. More representative samples were easier to achieve at higher population participation rates.

Feasibility was assessed by comparing the total number of schools contacted in order to achieve a sample of N = 60 participating schools. Figure 3 displays the average number of schools contacted across all iterations, for each sampling method and population response rate. CS required the least effort, followed by SCS and SRS, SSRS and SBS.

CONCLUSION

Our findings illustrate the benefits of actively pursuing generalizability in the recruitment stage. Even if balanced sampling isn’t implemented fully, simply stratifying and selecting a convenience sample within each strata enhances representativeness on observed covariates. However, under our model, SCS is both as effective—and as difficult—as SRS, a method held up as a rarely achievable gold standard. Nevertheless, the benefits of implementing a sampling method designed for generalizability are clear and we would encourage practitioners to allocate additional resources towards recruitment.

Findings from this study are limited due to the speculative nature of the sampling and recruitment models. In practice, convenience sampling might involve factors that we have not considered, such as proximity to research institutions or existing relationships between researchers and educational agencies. Furthermore, our participation model is based on research that examines district and school characteristics related to participation in large scale trials. Ironically, samples in this research are collected through convenience and are therefore an imperfect representation of population trends in participation propensity. Finally, our model has considered only direct sampling of schools, ignoring districts-level gatekeepers. These limitations underline the importance of further research into recruitment practices and school participation, in order to build a base of empirical evidence on how to recruit samples that support generalization from RCTs.
REFERENCES


Figure 1: Average B-index across iterations by sampling method
Figure 2: Average SMDs for each covariate by sampling method. Covariates include school size (n), % economically disadvantaged students (pED), % proficient in ELA (pELA), % english language learners (pELL), % proficient in math (pMath), % minority students (pMin), and indicators of whether schools in urban, suburban or town/rural (ToRu) geographic classifications.
Figure 3: Average number of schools contacted to achieve a sample of N = 60
Paper #3
Incorporating External Validity into the What Works Clearinghouse

**Background/Context.** The What Works Clearinghouse (WWC) is an online collection of rigorous evidence on educational interventions. Its goal is to inform decisions by educators and policymakers about whether to adopt or implement educational interventions. In reviewing papers and reports, the WWC assesses the internal validity of each study, aiming to identify interventions with strong evidence of their efficacy. However, the WWC does not assess studies’ external validity—and does relatively little to help educators assess the likelihood that the impact estimated for an intervention would apply to their school or district.

This focus on internal validity to the exclusion of external validity is not problematic if we assume that the efficacy of each intervention is constant. However, researchers have increasingly found evidence that treatment effects do in fact vary across sites and subgroups (e.g., Weiss et al., 2018). This indicates that the interpretation of the “effect” of an intervention in a study must be contextualized with reference to the types of schools, teachers, and students in the evaluation.

**Purpose.** In this talk we make recommendations for how to incorporate external validity into the WWC. This work is motivated by the needs of state and local policymakers/educators seeking evidence to inform decisions about which educational interventions to adopt.

**Sources.** In considering options for enhancing the WWC, we reviewed the following sources:

- The WWC Procedures Guide (Version 4.0);
- The WWC Standards (Version 4.0);
- The WWC’s Find What Works webpage;
- Selected WWC intervention reports and practice guides.

In addition, we considered recommendations from the burgeoning literature on external validity of rigorous impact studies in education (e.g., Olsen et al., 2013; Stuart et al., 2011; Tipton, 2013; Tipton and Olsen, 2018), frameworks for external validity (e.g., Cronbach and Shapiro, 1982; Shadish, Cook, and Campbell, 2002) and evidence on variation in treatment effects (e.g., Weiss et al., 2017).

**Options.** There are two key functions of the WWC that could potentially be enhanced by a consideration of external validity: (1) the rating of evaluation studies and (2) the reporting of information about these studies, including their findings. Revising the rating system to account for the external validity of studies and their findings would be possible, particularly with respect to the target population for the evaluation. However, the WWC serves a broad audience, and educators reviewing evidence from the WWC will have their own populations of interest. Alternatively, the WWC could make some enhancements designed to serve the broader audience of educators that use the WWC to inform state and local policy decisions.
**Recommendations.** We recommend that the WWC focus on enhancements designed to help educators identify evidence that is relevant to their context and predict the impacts of adopting interventions locally. In particular, we propose that the WWC:

- **Define a small number of populations on which to focus.** In particular, we recommend defining a single national population of all schools nationwide that serve the grade ranges served by the intervention. We also propose to define 4-8 subpopulations based on factors that are known to moderate the effect of interventions (broadly) and that are vary across the population. For example, sub-populations might be classified according to urbanicity, school and district size, and school achievement (but these are just examples). The population definitions would be created based on characteristics that are available in public use data (e.g., Common Core of Data, Stanford Education Data Archive).

- **Map WWC users to one or more subpopulations.** In *Find What Works!*, users would either be asked to provide some information about their district or school to match them to one or more of the subpopulations, or they would be asked to select one or more of the subpopulations that best characterizes the schools that they represent.

- **Ask study authors to identify (for the WWC) the schools that participated in their studies.** This will allow the WWC to describe the study sample using the public use data described earlier. Lists of participating schools would be obtained through “author queries,” which the WWC uses to obtain the information required to assign study ratings. Using this system the WWC could request the list of participating schools—either for studies that receive the highest rating or all studies that pass. While this may not be possible for all studies, recent research suggests this is possible in the vast majority of cases (e.g., Fellers, 2017; Tipton & Spybrook, working paper).

- **Estimate and describe the likely generalizability of study findings to the national population and each subpopulation.** We propose that the WWC calculate an index of similarity based on Tipton (2014). This index can be used to assess the likely generalizability of the study findings to each of the WWC’s focal populations. For each of the 4-8 populations, the likely generalizability could be color coded—for example, green if the population is similar to the study sample (index >= 0.9), yellow if the population is somewhat similar to the study sample (index >= 0.05 but < 0.9), and red if the population is quite different from the study sample (index < 0.5). These classifications will help users identify research that is relevant to their context.

**Conclusion.** To make fully informed decisions, local decision makers need to have knowledge both about the potential efficacy of an intervention and about the similarity of the schools or districts in the sample that established that evidence of efficacy to their own locale. In this talk we provide some concrete steps that the WWC could take to help ensure local decision making is fully “evidence based.”
References


