Improved generalizability through improved recruitment: Lessons learned from a large-scale randomized trial

Matlen, B.J. & Tipton, E.

Introduction

Over the past 10 years, the Institute of Education Sciences has funded over 100 randomized control trials (RCTs) in education. When implemented well, the results of these RCTs are included in the What Works Clearinghouse (WWC), ideally providing policymakers with the information to make evidence-based decisions in practice. However, while the WWC provides the results of RCTs, these studies may have taken place in contexts quite different from sites in which policymakers wish to make decisions. This *generalizability* problem is made even more difficult by the fact that the majority of RCTs conducted in education take place in convenience samples, with very little information on these samples provided in reports (Fellers, 2016; Olsen, Orr, Bell, & Stuart, 2013).

This concern has led to a new wave of methodological developments focused on improving generalizations. One method is a stratified recruitment plan, developed by Tipton (2014a, Tipton et al, 2014), which helps to ensure that the research sample is similar to the broader population from which the sample is drawn with respect to covariates that could potentially moderate the impact of the treatment.

While details of this plan has been described in other work (e.g., Tipton & Peck, 2017), the goal of this poster is to provide some on-the-ground “lessons learned” from a concrete illustration of the method in a large scale RCT.

The Evaluation

This IES funded RCT was designed to determine the effect of a Web-based Activity and Testing System (WATS) in the population of community colleges in California. The goal of the study was to estimate the average impact of this WATS platform in supporting community college students’ Algebra I knowledge. In this study, instructors were either randomly assigned use the WATS platform as a part of their instruction (treatment condition) or to teach using their normal practices (control condition).

Stratified Recruitment

We implemented a stratified recruitment plan in this study which involved: (1) Defining an inference population; (2) Selecting possible treatment effect moderator variables; (3) Creating and describing the strata; and (4) Developing a within-stratum recruitment strategy.

The inference population in this study was determined to be the population of community colleges that offered semester-long developmental algebra courses in California. The state of California was selected in part because the state is large and diverse, and in part to decrease variability from differing mathematics standards and graduation requirements across states. Using publicly available information about California Community Colleges, we created a database that consisted of relevant covariates (i.e., student demographics, faculty characteristics, etc.) of on all eligible colleges for the study. We then conducted a cluster analysis on this data which resulted in five distinct strata of colleges. Recruitment of sites into the study was then made to be similar to the proportion of each strata in the interference population.

Overall, the study was successful in recruiting a proportionate ratio of sites within each stratum (see Figure 1). The overall degree of similarity of the sample to the inference population was estimated to have a generalizability index of .93 (Tipton, 2014b), which is considered "very high", similar to what might be expected in a random sample of the same size. Furthermore, the average absolute standardized mean difference (ASMD) across these covariates was 0.09 (see Table 1), which is small, indicating a high degree of similarity between the resulting sample and population (see Stuart et al, 2011).

Lessons Learned

We identified five lessons from implementing the stratified recruitment method which may improve the likelihood that others can implement approaches aimed at improving the generalizability of the results of their experiments.

Lesson One: Learning about the population can take time. Building the population frame involved making conceptual decisions within the research team, collecting data, ensuring its accuracy, conducting statistical analyses, and finally communicating the method to recruitment staff. The entire process took place over the course of approximately 2 months. Future studies and evaluations should build in adequate time to complete this process, ideally before recruiting begins.

Lesson Two: Need buy-in from recruiters. Communicating the generalizability goals effectively to the recruiters is central to its effective implementation. Towards this goal, it is useful to have meetings with the recruiters and the methodologists, with the goal of making the generalizability efforts clear. Recruiters should be allowed to ask questions about the value of the recruitment method, as well potential roadblocks and strategies to incentivize participation. This dialogue can help to create shared understanding of the importance of the recruitment method in attaining the research goals.

Lesson Three: Strategizing is a dynamic process. Providing recruiters with targets for each stratum was not effective, as the proportion of sites recruited in one stratum depends on all other stratum. Thus, we learned that the recruiting process is more effective when it is communicated as a *dynamic process*, involving frequent checking of the distribution of recruited sites across the strata against the targets. In order to facilitate this process, the research team created a shared document that visually displayed for the recruiters the stratum proportions in the population (the recruitment targets) and their up-to-the-minute success in matching these targets. This strategy helped to ensure that the sample proportions did not end up wildly inconsistent with the target proportions.

Lesson Four: Stratum info can help guide recruitment. Tipton (2014a) suggests that an additional benefit of the stratification approach is that it can be used to *describe* the population as well, potentially helping with recruitment efforts. In the study, we followed this approach, providing descriptions of the strata and discussing with the recruitment team how these could be leveraged during recruitment (e.g., "we are really hoping that schools like yours are represented in our study so that we can appropriately generalize our findings to schools that offer lots of remedial courses").

Lesson Five: Stratification didn't make recruitment harder. Recruiters interviewed were unanimous that the method did not create additional work or make recruiting harder. Recruiters indicated that the method was helpful as it provided clear priorities on where to invest their recruitment efforts.

References

Fellers, L. (2016) Does IES funded research represent US schools well? An evaluation of issues of generalizability in grant funded research between 2005- 2014. *Doctoral dissertation: Columbia University.*

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*(1), 107-121.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174(2)*, 369-386.

Tipton, E. (2014a). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation review*, *37*(2), 109-139.

Tipton, E. (2014b). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478-501.

Tipton, E., Hedges, L.V., Vaden-Kiernan, M., Borman, G.D., Sullivan, K. & Caverly, S. (2014) Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness, 7*(1): 114-135.

Tipton, E., & Peck, L. R. (2017). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation review*, 41(4): 326 – 356.

Table 1. Means and standard deviations for the population and recruited sample, as well as the absolute standardized mean difference (ASMD) for each variable.

	Population		Sample		ASMD
	N = 113		*N* = 34		
	M	*SD*	*M*	*SD*	
TotalEnrollment	20168	13149	21146	13527	0.07
MathBasicEnrollment	1018	769	974	703	0.06
MathBasicSections	35	30	38	33	0.09
TotalAcadEmp	519	304	568	322	0.16
MathBasicFTES	150	116	143	92	0.06
Female	0.53	0.06	0.54	0.04	0.09
AfricanAmerican	0.07	0.08	0.05	0.04	0.30
Asian	0.13	0.10	0.14	0.13	0.13
Hispanic	0.42	0.17	0.4	0.17	0.10
WhiteNonHispanic	0.30	0.16	0.32	0.16	0.10
USCitizen	0.86	0.09	0.86	0.1	0.00
Age19orLess	0.26	0.06	0.26	0.06	0.06
Age20to24	0.34	0.06	0.34	0.05	0.01
Age25to39	0.26	0.05	0.25	0.04	0.17
Age40to49	0.07	0.03	0.07	0.02	0.03
Age50above	0.08	0.05	0.08	0.05	0.07
FirstTime	0.17	0.06	0.16	0.05	0.16
FirstTransfer	0.08	0.05	0.08	0.05	0.01
Returning	0.11	0.05	0.1	0.05	0.29
Day	0.74	0.08	0.75	0.07	0.11
Evening	0.18	0.05	0.18	0.05	0.09
Unit.1to9	0.47	0.10	0.47	0.08	0.05
Unit9to14.9	0.38	0.08	0.38	0.06	0.02
Unit_15plus	0.09	0.04	0.09	0.04	0.08
TenureTrackFac	0.20	0.05	0.21	0.05	0.15
TemporaryFac	0.50	0.08	0.5	0.08	0.01
MathBasicRetention	0.83	0.06	0.84	0.05	0.21
MathBasicSuccess	0.54	0.08	0.56	0.08	0.26
MedHouseIncome	61102	14291	60029	14457	0.08
CountyPov	0.17	0.05	0.17	0.05	0.06

Figure 1. Proportion of recruited colleges relative to the target proportion at the beginning of the study.