


Generalizability of Randomized Trial Results to Target Populations: Design and Analysis Possibilities

Research on Social Work Practice
2018, Vol. 28(5) 532-537
© The Author(s) 2017
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1049731517720730
journals.sagepub.com/home/rsw


Elizabeth A. Stuart¹, Benjamin Ackerman¹, and Daniel Westreich²

Abstract

Randomized trials play an important role in estimating the effect of a policy or social work program in a given population. While most trial designs benefit from strong internal validity, they often lack external validity, or generalizability, to the target population of interest. In other words, one can obtain an unbiased estimate of the study sample average treatment effect from a randomized trial; however, this estimate may not equal the target population average treatment effect if the study sample is not fully representative of the target population. This article provides an overview of existing strategies to assess and improve upon the generalizability of randomized trials, both through statistical methods and study design, as well as recommendations on how to implement these ideas in social work research.

Keywords

literature review, evidence-based practice

Many questions of policy or practice interest involve estimates of the effect of some policy or program in a target population of interest. For example, a social work agency may be interested in predicting the average effects if all of their clients receive a new model of program delivery or a state may be deciding whether to invest in a new training program for social workers across the state. A challenge in estimating these effects, however, is that common existing study designs are often not well targeted for these target population effects. In particular, randomized trials are often conducted in study samples that are explicitly not representative of the target populations in which the policies or programs may eventually be implemented.

Randomized trials have played a critical role in informing evidence-based social work practice, used alongside physicians' expertise and patients' preferences to make the best practical decisions on an individual level (Soydan, 2008). Of interest in this article, however, is not how randomized trials can inform personalized decision-making but rather how average effects of interventions can impact policy and community-level outcomes in well-defined populations. Consider, for example, a randomized trial in which high school students at a public school are provided with training on how to prevent intimate partner violence and are then followed for 4 years. If the trial results suggest that, on average, students who received the training reported less violent behavior with their partners than the control group, then it may be of a state's interest to implement such programming on a larger scale. The methods discussed here can help a state assess how relevant the findings in the trial are to the state as a whole and what the average effects might be if the program were implemented statewide.

This article provides an overview of design and analysis methods for how we can assess and enhance our ability to estimate the effects of interventions in well-defined target populations. Because other work has primarily focused on analysis methods, we put somewhat more emphasis on study design options for estimating causal effects in well-defined target populations. Recently researchers have distinguished "generalizability," which involves generalizing results from a study sample to the population from which that sample was selected (potentially randomly but more commonly nonrandomly; Cole & Stuart, 2010), from "transportability," which involves estimating effects in a completely external population or one that the study sample was not drawn from (Bareinboim & Pearl, 2013; Hernan & VanderWeele, 2011). In general, the methods described in this article will be relevant for both scenarios—in part because it is sometimes difficult to draw a bright line between the two—but distinctions for the two scenarios are described when appropriate.

This article proceeds as follows. We first present background on the problem including some notation and a clear description of the goal of analysis and the setting. We then briefly describe analysis strategies for estimating target

¹ Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

² University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Corresponding Author:

Elizabeth A. Stuart, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA.

Email: estuart@jhu.edu

population treatment effects before turning to study design strategies to enhance the generalizability of trial results to well-defined target populations. We end with a broader discussion including relevance of the ideas for social work research.

Background on the Problem

The first step in examining generalizability or transportability is to identify the target population of interest. Discussing generalizability or transportability without that is in fact meaningless, and a particular study may be generalizable to one population but not to another (in fact that is essentially always the case). We find that all too often this initial step is not taken, however; researchers jump to discussing “the generalizability” of a study without clarifying to what population one is interested in generalizing. For example, two states (with very different populations) may both be interested in determining whether the Nurse Family Partnership (Olds et al., 1998) might be beneficial for the new parents in their state; the residents of these two states might be two different but both well-defined, target populations. Throughout the rest of this article, we will assume that “the population” has been well-specified and defined.

Clarification of Estimands

We assume that a randomized trial has been conducted in a sample of size n , and there is a well-defined target population of size N to which researchers would like to generalize the results from the randomized trial (e.g., a randomized trial of the Nurse Family Partnership; Olds et al., 1998).

The randomized trial can provide an unbiased effect estimate for the study sample: $\text{SATE} = \frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0) | S_i = 1)$, where n denotes the sample size of the trial, $Y_i(1)$ denotes the outcome for subject i if they receive treatment, $Y_i(0)$ denotes the outcome for subject i if they receive the control condition, and $S_i = 1$ if subject i is in the trial sample and 0 otherwise. However, ultimate interest is in a target (population) average treatment effect: $\text{TATE} = \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0))$. While the effect estimate in the trial is unbiased for the sample in the trial, it is not necessarily unbiased for the target population average treatment effect (TATE).

When Will The Sample and Target Population Effects Differ?

Intuitively and formally the sample and target effects will differ if there are factors that moderate (modify) treatment effects *and* if the distribution of those factors differ between the sample and the target population. For example, an intervention may be more effective among young adults, and different locations may have different age distributions. That

combination can lead to bias when trying to generalize the results of a trial from one location to another. Cole and Stuart (2010) present a formalization of this. Let α denote an estimate of the TATE and β an estimate of the SATE, such that the difference, $\beta - \alpha$, represents the bias of the sample average treatment effect (SATE) as a measure of the TATE. Consider the simple setting where there is only one pretreatment covariate, Z , which is binary. Cole and Stuart (2010) derive the formula for the bias of the SATE as a measure of the TATE:

$$\beta - \alpha = b_{xz} \times \left\{ \frac{P(Z=1)}{P(S=1)} \times [P(S=1|Z=1) - P(S=1)] \right\}$$

Here, b_{xz} denotes the coefficient for treatment effect heterogeneity due to Z obtained from the outcome model $E(Y_i) = b_0 + b_x X_i + b_z Z_i + b_{xz} X_i Z_i$, where X is a binary variable indicating treatment. Therefore, the bias depends on the magnitude of treatment effect heterogeneity (b_{xz}), the proportion of the target population sampled for the trial ($P(S=1)$), the overall prevalence of the pretreatment covariate Z ($P(Z=1)$), and the difference in the probability of participating in the trial across levels of Z , denoted as $(P(S=1|Z=1) - P(S=1))$. Note there will be no bias if the probability of being selected for the trial does not depend on Z ($P(S=1) = P(S=1|Z=1)$), if the sample consists of the entire target population ($P(S=1) = 1$), or if there is no treatment effect heterogeneity across levels of Z ($b_{xz} = 0$).

The equation above focused on a continuous outcome and an effect estimate parameterized as a difference in outcome means. One key point worth noting is that when the outcome is binary, sample and target effects can be expected to differ on at least one scale (e.g., risk difference or risk ratio), whenever the baseline risks differ between the two populations (a difference in baseline risks is a sufficient condition for moderation of treatment effects on at least one scale). Thus, any trial that overenrolls high-risk individuals from the target population (as is frequently done to enhance study power) will produce effect estimates that cannot be expected to generalize unconditionally on all scales.

There is growing evidence in practice that randomized trial samples are often not representative of target populations of interest (see, e.g., Rothwell, 2005; Stirman, DeRubeis, Crits-Christoph, & Rothman, 2005). Braslow et al. (2005) documented that randomized trials of psychiatric treatment often underenrolled minorities (relative to a target population of individuals with psychiatric disorders across the United States). Wisniewski et al. (2009) compared individuals in a large-scale pragmatic effectiveness trial of depression treatment to the subset of patients who would likely have been included in a more typical efficacy trial (with standard inclusion and exclusion criteria) and found large differences in both characteristics and effects. More recent work in studies

of drug abuse treatment documented that individuals in randomized trials of those treatments differ substantially from individuals seeking treatment for drug abuse in the United States in general, especially in terms of employment status and education levels (Susukida, Crum, Ebnesajjad, Stuart, & Mojtabai, 2017).

In education research, Stuart, Bell, Ebnesajjad, Olsen, and Orr (2017) detailed large differences between the types of school districts that participate in large-scale “national” evaluations of educational interventions and three plausible target populations: districts nationwide, disadvantaged districts nationwide, and, for federally funded programs, the districts nationwide implementing those programs. Stuart et al. found large differences between the districts participating in evaluations and all of these populations; for example, large-, low- or mid-performing urban districts represent approximately 48% of the study samples but only 4% of districts nationwide and 7.5% of disadvantaged districts nationwide. Bell et al. (2016) then showed that these differences can result in bias when trying to naively estimate the TATE using data from these trial samples, estimating that the external validity bias due to trial samples not representing the target population is on the order of .1 standard deviations.

In social work practice, interventions play a central role in improving conditions for clients, and the optimal method of evaluating the effectiveness of social work interventions is through randomized trials. While there has been limited quantification of the differences between trial samples and target populations in social work research, several studies have discussed the limitation of not having representative samples. Zhai et al. (2010) concluded that in order to better generalize the results from their trial examining dosage effect on school readiness of preschool-aged children, future studies should recruit samples more demographically similar to the national population of interest. In a review of Randomized Controlled Trials (RCTs) for parents of children with autism spectrum disorder (ASD), Dababnah and Parish (2016) observed that across studies, generalizability of trial results was weakened by the lack of racial, ethnic, and socioeconomic diversity that existed among the target population of parents of children with ASD. Bronstein, Gould, Berkowitz, James, and Marks (2015) also call for replication studies in more diverse communities in order to better address the generalizability of their results, indicating the importance of having representative study samples.

Analysis Methods for Estimating the Target Average Treatment Effect

Recent work has developed statistical approaches for estimating the target average treatment effect using data from a randomized trial and covariate information on the target population. Broadly, these primarily involve either (1) weighting methods that weight the study sample to resemble the target population on baseline characteristics that may moderate treatment effects, (2) flexible models of the outcome fit in the study sample and then used to predict impacts in the target

population, or (3) both methods combined. Kern, Stuart, Hill, and Green (2016) provide an overview of these approaches and simulation studies comparing their performance. Note that all of these approaches assume that there is a set of covariates that are observed consistently across trial sample and target population data sets.

The weighting approach to generalization involves stacking trial and population data on top of each other and fitting a model of participation in the trial as a function of observed characteristics; essentially, adjusting for sample and population differences by modeling the probability of participating in the trial. Individuals in the trial are then weighted by one over their probability of participating in the trial (similar to nonresponse weights in survey samples or propensity score weights in non-experimental studies) in outcome analyses; these weighted outcome models provide an estimate of the TATE, adjusting for the sample and target population differences in observed covariates. Cole and Stuart (2010) present an example of this approach, generalizing the results of a randomized trial of treatment for HIV to the population of individuals newly infected with HIV in the United States in 2006. Similar approaches are described in Hartman, Grieve, Ramsahai, and Sekhon (2015), O’Muircheartaigh and Hedges (2014), and Tipton (2013). This approach can be thought of as a smoothed version of poststratification, whereby effects might be estimated for specific subgroups in the trial (e.g., males and females) and then the subgroup effects weighted using the population distribution of that variable (male/female) to obtain a population effect estimate; the weighting version of this approach allows researchers to adjust for a larger set of factors than would be possible using direct poststratification (also known as standardization).

A second class of methods instead focus on using data in the trial to model the outcome as a flexible function of treatment status and the covariates (including potential interactions) and then using that model to predict outcomes (and thus effects) in the target population, based on the covariate distribution observed in the population. This approach was examined in Kern et al. (2016), using a specific modeling approach called Bayesian additive regression trees, which fits a very flexible outcome model using a nonparametric approach similar to random forests. Kern et al. (2016) found that this approach worked quite well even for somewhat complex outcome models.

A third broad class of methods combines these two approaches, similar in spirit to “doubly robust” approaches in nonexperimental studies (Kern, Stuart, Hill, & Green, 2016). In particular, with these methods both selection (trial participation) and outcome models are used, with the outcome models fit using weights generated as in the first approach.

The primary assumption underlying all of these approaches is that of conditionally unconfounded sample selection that we have observed the factors that moderate treatment effects and differ between sample and population. In other words, we have to be willing to assume that, once we adjust for the set of observed covariates, treatment effects are the same in the trial

sample and the population. This assumption, sometimes called “ignorability of sample selection,” is formalized in Hartman et al. (2015) and Kern et al. (2016) and differs depending on whether outcomes under the control condition are available in the population of interest. Huitfeldt et al. (2016) discuss variations on this assumption and implications for variable selection for modeling or outcome model-based approaches.

The assumption of conditionally unconfounded sample selection can be a heroic assumption in practice, especially given sometimes limited data on the population of interest (e.g., see Stuart & Rhodes, 2016). So what can we do instead? One key aspect is careful and thoughtful selection of covariates and attention to the comparability of measures across data sources. This selection can be greatly informed by theoretical models of participation in the randomized trials of interest and the interventions themselves, and, in particular, the factors that may relate to effects and participation. However, in practice, we often do not observe all of the factors that we would like to adjust for. For these scenarios, sensitivity analyses have been developed to assess how much the TATE estimates would change if there were an unobserved effect moderator (Nguyen, Ebnesajjad, Cole, & Stuart, 2017). However, another, perhaps better option, is to use smart design choices to make these assumptions less heroic. We turn to these designs now.

Design Options for Enhancing Generalizability to a Target Population

When the target population of interest is known in advance of a randomized trial being conducted, there are a number of design possibilities to better ensure that the results from the trial can be used to estimate effects in that target population. We note that these design options are not sufficient and the analysis strategies introduced above are often needed in addition, given that (1) there may be multiple target populations of interest from a given study (e.g., two U.S. states may both be interested in estimating effects in their own state population) and (2) the target population of interest may change after the trial is conducted, including due to general temporal changes and time trends.

Perhaps the “gold standard” for estimating the TATE is randomized trials conducted in formally representative samples (Imai, King, & Stuart, 2008). We are aware of a handful of studies that randomly sampled sites to participate from a well-defined target population (see Olsen et al., 2013). All evaluations in this category were of U.S. federal government programs, where program implementers (sites) could be mandated to participate in the evaluation: Upward Bound (Sefor, Mamun, & Schirm, 2009), Job Corps (Burghardt et al., 1999; in fact this study included *all* Job Corps sites across the United States), and Head Start (Puma et al., 2010). The possibilities for such designs may increase in the future, however, with more and more large-scale population administrative data sets. For example, a health system interested in studying a new warning system for potential drug interactions could be evaluated using a random sample of providers or patients in their population, through an electronic health record system. Olsen

and Orr (2016) present some of the considerations when setting up a study that aims for random selection from the target population. When there are concerns that some individuals may not agree to participate in a randomized trial, some studies conduct parallel randomized and nonrandomized arms, whereby the individuals who do not consent to randomization are allowed to choose their treatment condition but with their outcomes still tracked over time.

Another design approach that has been proposed does not use random sampling from a population but rather picks sites systematically in order to cover the target population (Shadish, Cook, & Campbell, 2002). One particular approach, formalized by Tipton et al. (2014), involves stratifying the population on factors strongly related to outcome. It requires a sample frame of potential study subjects, covariate information on them, and knowledge of the prognostic factors likely related to outcomes. Subjects are then selected for the study based on strata defined by those prognostic factors, with the goal of a final study sample that has representation from all strata. Tipton et al. (2014) illustrate the approach using the design of a scale-up study of mathematics and reading interventions.

There may also be a place for nonexperimental studies when primary interest is in a target population effect estimate. As formalized by Imai, King, and Stuart (2008), a well-done nonexperimental study in a data set representing the target population of interest may actually lead to less bias in the TATE than would a small-scale randomized trial in a very nonrepresentative study sample due to trade-offs between internal and external validity. Thus, a well-done nonexperimental study (such as described by Rosenbaum, 1999, or Rubin, 2001) that can be conducted in a sample representative of the target population of interest may be worth considering when interest is in informing decisions in that population.

Some of these design options may seem daunting, and in some contexts, it may not be feasible to consider random selection of subjects for a randomized trial. However, even in those cases, there are still important design lessons that can be taken from this literature. In particular, all randomized trials should collect data on variables that are likely to moderate effects and may relate to study participation. Studies should also consider their target population and show a table in the paper documenting the characteristics of study participants and the target population. One prerequisite for doing so will be the collection of variables in a consistent way between trial sample and population data sets; for example, with trials making an effort to use the measures that are available in common population data sets (e.g., large-scale national surveys). Najafzadeh and Schneeweiss (2017) discuss the importance of measure comparability in the context of medical trials and electronic health records to reflect target populations.

Conclusions and Recommendations for Future Work in Social Work

In summary, no trial is necessarily generalizable or even generalizable in expectation unless (i) sample = target or

(ii) sample = simple random sample of target. Otherwise, the assumption of generalizability is effectively an observational data analysis assumption. Until recently, this point has been underappreciated by nearly all fields, but it has important implications for the broader policy and practice relevance of research.

Thus, although generalization of results to target populations is often heroic, there are design and analysis choices to make it more plausible and believable. This includes careful choice of measures and efforts to provide measures comparable across studies. Stuart and Rhodes (2016) found it very difficult to find data on a trial and target population in the field of early childhood education with any comparable measures and, in fact, even the best example found had only seven measures in common between the trial and population. This makes the assumption of unconfounded sample selection particularly problematic and heroic. One way to think about this is that the analysis approaches above, which adjust for observed effect moderators, can help move from an assumption of missing completely at random to an assumption more like missing at random, but we can never eliminate the possibility of missing not at random, just as in nonexperimental studies, we cannot guarantee that there is no unobserved confounding. But careful selection and use of observed covariates can at least move us a step in that direction.

Researchers should also consider whether the design approaches described above are feasible for their work. And as noted above, even when, for example, random sampling from the target population is not feasible, efforts toward measure comparability with large-scale target population data sets will at least facilitate the use of analysis strategies to assess and enhance generalizability after the fact.

In this article, we have focused on situations with one randomized trial and one well-defined target population. In some contexts, there might be multiple trials available (e.g., Petrosino, Turpin-Petrosino, Hollis-Peel, & Lavenberg, 2013), or a combination of experimental and nonexperimental evidence, in which case other approaches may be more appropriate. Possibilities in that case include cross-design synthesis approaches also known as research synthesis (Pressler & Kaizar, 2013; Prevost, Abrams, & Jones, 2000). Broadly, this class of methods might model effects as a function of study characteristics and explicitly model the internal and external validity bias, for example, with prior distributions on the non-identified bias parameters (e.g., Turner, Spiegelhalter, Smith, & Thompson, 2009).

A number of fields are just beginning to understand the implications of these ideas in their fields and, for example, how representative (or nonrepresentative) their trials tend to be. Social work should begin to develop such an understanding, through documentation of the characteristics of individuals and sites that participate in rigorous evaluations and how they compare to potential target populations. Data needs are paramount, however, in particular: (1) population data to provide background information on target populations, (2) potentially population data to provide a sampling frame for selection of study

subjects, and (3) comparability of measures between those population data sets and randomized trials. The analysis approaches described in this article can only go so far if the data are not available or appropriate.

In conclusion, this article has provided a review of methods for enhancing the ability to draw target population inferences from randomized trials, attempting to bridge both internal validity and external validity and ensure that our research studies are as useful as possible for policy and practice.

Authors' Note

The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of the Institute of Education Sciences, the National Institutes of Health, or the Patient-Centered Outcomes Research Institute, its board of governors or methodology committee.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supposed in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305D150003 (Principal Investigators (PIs): Stuart and Olsen), by a Patient-Centered Outcomes Research Institute Award (ME-1502-27794; PI: Dahabreh), and by the National Institutes of Health, through grant DP2-HD-08-4070 (PI: Westreich).

References

- Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1, 107–134.
- Bell, S.H., Olsen, R.B., & Orr, L.L. and Stuart, E.A. (2016). Estimates of external validity bias when impact evaluations select sites non-randomly. *Educational Evaluation and Policy Analysis*, 38, 318–335.
- Braslow, J. T., Duan, N., Starks, S. L., Polo, A., Bromley, E., & Wells, K. B. (2005). Generalizability of studies on mental health treatment and outcomes, 1981 to 1996. *Psychiatric Services*, 56, 1261–1268.
- Bronstein, L. R., Gould, P., Berkowitz, S. A., James, G. D., & Marks, K. (2015). Impact of a social work care coordination intervention on hospital readmission: A randomized controlled trial. *Social Work*, 60, 248–255.
- Burghardt, J., McConnell, S., Meckstroth, A., Schochet, P., Johnson, T., & Homrighausen, J. (1999). *National job corps study: Report on study implementation*. Princeton, NJ: Mathematica Policy Research, Inc.
- Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations the ACTG 320 trial. *American Journal of Epidemiology*, 172, 107–115.
- Dababnah, S., & Parish, S. L. (2016). A comprehensive literature review of randomized controlled trials for parents of young

- children with autism spectrum disorder. *Journal of Evidence-informed Social Work*, 13, 277–292.
- Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178, 757–778.
- Hernan, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology (Cambridge, Mass.)*, 22, 368–377.
- Huitfeldt, A., Swanson, S. A., Stensrud, M. J., & Suzuki, E. (2016). Effect heterogeneity and variable selection for standardizing experimental findings. *arXiv preprint arXiv:1610.00068*.
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 481–502.
- Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9, 103–127.
- Najafzadeh, M., & Schneeweiss, S. (2017). From trial to target populations—Calibrating real-world data. *New England Journal of Medicine*, 376, 1203–1205.
- Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., & Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, 11, 225–247.
- Olds, D., Henderson, C. R., Jr., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., . . . Powers, J. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *Journal of the American Medical Association*, 280, 1238–1244.
- Olsen, R., Bell, S., & Orr, L. and Stuart, E.A. (2013). External Validity in Policy Evaluations that Choose Sites Purposively. *Journal of Policy Analysis and Management* 32(1): 107–121.
- Olsen, R. B., & Orr, L. L. (2016). On the “where” of social experiments: Selecting more representative samples to inform policy. *New Directions for Evaluation*, 2016, 61–71.
- O’Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 195–210.
- Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013). ‘Scared Straight’ and other juvenile awareness programs for preventing juvenile delinquency: A systematic review Cochrane Systematic Reviews 2013:5 DOI: 10.4073/csr.2013.5.
- Pressler, T. R., & Kaizar, E. E. (2013). The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Statistics in Medicine*, 32, 3552–3568.
- Prevost, T. C., Abrams, K. R., & Jones, D. R. (2000). Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. *Statistics in Medicine*, 19, 3359–3376.
- Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., . . . Ciarico, J. (2010). *Head start impact study* (Final Report). Washington, DC: Administration for Children & Families.
- Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science*, 14, 259–278.
- Rothwell, P. M. (2005). External validity of randomised controlled trials: “To whom do the results of this trial apply?” *The Lancet*, 365, 82–93.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2, 169–188.
- Seftor, N. S., Mamun, A., & Schirm, A. (2009). *The impacts of regular upward bound on postsecondary outcomes seven to nine years after scheduled high school graduation* (Final Report). Washington, DC: U.S. Department of Education.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage learning.
- Soydan, H. (2008). Applying randomized controlled trials and systematic reviews in social work research. *Research on Social Work Practice*, 18, 311–318.
- Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Rothman, A. (2005). Can the randomized controlled trial literature generalize to nonrandomized patients? *Journal of Consulting and Clinical Psychology*, 73, 127.
- Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness*, 10, 168–206.
- Stuart, E. A., & Rhodes, A. (2016). Generalizing treatment effect estimates from sample to population a case study in the difficulties of finding sufficient data. *Evaluation Review*. doi:10.1177/0193841X16660663.
- Susukida, R., Crum, R. M., Ebnesajjad, C., Stuart, E. A., & Mojtabei, R. (2017). Generalizability of findings from randomized controlled trials: Application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction*, 112, 1210–1219.
- Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38, 239–266.
- Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., 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, 114–135.
- Turner, R. M., Spiegelhalter, D. J., Smith, G., & Thompson, S. G. (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 21–47.
- Wisniewski, S. R., Rush, A. J., Nierenberg, A. A., Gaynes, B. N., Warden, D., Luther, J. F., . . . Trivedi, M. H. (2009). Can phase III trial results of antidepressant medications be generalized to clinical practice? A STAR* D report. *American Journal of Psychiatry*, 166, 599–607.
- Zhai, F., Raver, C. C., Jones, S. M., Li-Grining, C. P., Pressler, E., & Gao, Q. (2010). Dosage effects on school readiness: Evidence from a randomized classroom-based intervention. *Social Service Review*, 84, 615–655.