Article

Evaluation Review
2017, Vol. 41(4) 326-356
© The Author(s) 2016
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0193841X16655656
journals.sagepub.com/home/erx
SAGE

A Design-Based Approach to Improve External Validity in Welfare Policy Evaluations

Elizabeth Tipton[1] and Laura R. Peck[2]

Abstract

Background: Large-scale randomized experiments are important for determining how policy interventions change average outcomes. Researchers have begun developing methods to improve the external validity of these experiments. One new approach is a balanced sampling method for site selection, which does not require random sampling and takes into account the practicalities of site recruitment including high nonresponse. **Method:** The goal of balanced sampling is to develop a strategic sample selection plan that results in a sample that is compositionally similar to a well-defined inference population. To do so, a population frame is created and then divided into strata, which "focuses" recruiters on specific subpopulations. Units within these strata are then ranked, thus identifying "replacements" similar to sites that can be recruited when the ideal site refuses to participate in the experiment. **Result:** In this article, we consider how a

[1] Department of Human Development, Teachers College, Columbia University, New York, NY, USA
[2] Social and Economic Policy, Abt Associates Inc., Bethesda, MD, USA

Corresponding Author:
Elizabeth Tipton, Department of Human Development, Teachers College, Columbia University, 525 W. 120th St., Box 118, New York, NY 10027, USA.
Email: tipton@tc.columbia.edu

balanced sample strategic site selection method might be implemented in a welfare policy evaluation. **Conclusion:** We find that simply developing a population frame can be challenging, with three possible and reasonable options arising in the welfare policy arena. Using relevant study-specific contextual variables, we craft a recruitment plan that considers nonresponse.

Keywords
job training, content area, methodological development

Randomized experiments are important for determining the extent to which an intervention changes outcomes. Random assignment to treatment conditions ensures that such an evaluation can estimate unbiased *causal* impacts, resulting in high internal validity. Indeed, Orr (2015) asserts that the field of evaluation has been highly focused on internal validity to the detriment of external validity. In most large-scale experiments, sites are not selected randomly, resulting in low external validity. In fact, very few randomized experiments have successfully taken place within a randomly sampled set of sites, with the *Digest of Social Experiments* identifying only 7 of its 273 as meeting this criterion (Greenberg & Shroder, 2004; Olsen, Bell, Stuart, & Orr, 2013). This makes it difficult to generalize beyond the composition of sites in the experiment, leaving policy makers interested in evidence-based practice with few tools to determine whether the causal effect found in the experiment applies to a specific population.

As evaluation results begin to play a more prominent role in shaping policy, this question of generalizability becomes increasingly important (e.g., Cook, 2014). As one feature of external validity, generalizability has to do with the ability to relate the sample of units and settings found in the experiment to the set of units and settings found in the population. More specifically, the results of an evaluation are considered generalizable to a particular inference population only when the sample of sites or units that took part in the experiment is *compositionally similar to* or *representative of* the sites or units in the inference population on covariates that explain variation in treatment impacts (e.g., Orr, 2015; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013). When the sample is compositionally similar to the population on *all* the covariates that moderate treatment impacts, the average treatment effect (ATE) estimated in the evaluation is *unbiased* for the population ATE. Furthermore, this means that when the composition of sites or units in the evaluation *differs* from that in the population, the estimated treatment impact is biased. As Bell, Olsen, Orr, and Stuart (2016) indicate, the degree of bias that results from nonrepresentative sample selection can be as large as that which results from nonrandom assignment of units in observational studies seeking to identify causal effects of interventions. For policy makers, this means that the results from a randomized experiment may not indicate well the performance of a particular intervention for its particular inference population.

As attention in the evaluation community has shifted toward thinking about generalization, statisticians have begun developing alternative methods for achieving representativeness (i.e., covariate balance) without random sampling. This literature began with the development of retrospective methods based on propensity score matching methods (O'Muircheartaigh & Hedges, 2014; Stuart et al., 2011; Tipton, 2013). These methods are generalizations to the post hoc nonresponse adjustments found in survey sampling, including poststratification and inverse probability weighting. The goal is to reweight the evaluation's achieved sample to be compositionally similar to a well-defined inference population on a set of specified covariates. These covariates must be measured on all units in the sample and population and when exhaustive—in the sense that all covariates that explain variability in treatment impacts across units are included—allow for an unbiased estimate of the treatment impact in the population to be calculated. Importantly, this *population* average treatment impact estimate can differ from the usual *sample* ATE estimate, and, as Tipton (2013) illustrates, typically has a larger standard error. Additionally, in some cases, this reweighting approach fails, for example, when there are *coverage errors* (Tipton, 2013)—parts of the population without similar sample units that represent them.

In response to the limitations of post hoc adjustments, Tipton (2014a) and Tipton et al. (2014) focus on the development of methods for selecting samples that are compositionally similar to a population *without requiring* random sampling. These methods are stratification based and have two goals: first, to ensure that coverage errors are minimized (thus allowing post hoc reweighting if needed), and second, ideally to enable the selection of a sample that is compositionally similar to the population. Like the retrospective adjustment methods, these methods require evaluators to identify a set of covariates that are likely to explain variation in treatment impacts and for these covariates to be measured in the population (thus enabling the creation of a sampling frame). These methods build on the ideas found in model-based sampling, specifically *balanced-sampling*, and are often used when there is survey nonresponse (e.g., Valliant, Dorfman, & Royall, 2000).

Importantly, these methods work within the constraints found in real-world site selection, including eligibility concerns, high refusal rates, and the use of recruitment teams (see Tipton, 2014a, for a full discussion). The approach takes into account the very real resource constraints that recruiters face. For example, the decision regarding the number of strata used in the study balances what is statistically ideal with what is practically feasible. When combined with post hoc adjustments, these methods result in an ATE estimate that, like the post hoc approach, has less bias than the standard nonrandom sample, but—when used with proportional allocation—has little cost in terms of variance inflation (Tipton, 2013). This article contributes to this line of research.

Much of this recent discussion about external validity has focused on the evaluation of education programs. The U.S. Department of Education's Institute for Education Sciences has shifted toward an emphasis on randomized experimental evaluations over the past 10 years. The related focus on generalizability has been a natural outgrowth of the improvements in internal validity as well as the availability of national and state population frame data on districts, schools, and teachers. More than 600 randomized experiments have been conducted on health and social welfare: Greenberg and Shroder's (2004) count of 118 plus the *Randomized Social Experiments eJournal*'s early 2015 count of at least 500 experiments since 2004. Given the large number of health and social welfare randomized experiments, in this article, we argue that generalization not only *should* be, but *can* be, addressed in randomized social welfare evaluations, thereby increasing the relevance of their results.

In social welfare program evaluations, generalization is often of interest, though no formal method for making generalizations is used. As an example, "national" is included in the title of the National Evaluation of Welfare to Work Strategies (NEWWS), highlighting the desire for generalizable results. Yet, the evaluation sites (selected counties each with multiple local offices within selected states) were recruited based on a proven history of Job Opportunities and Basic Skills program administration, substantial size in number of welfare recipients served, and sufficient funding to run the program (Hamilton, Brock, & Farkas, 1994). Further, the sites decided to participate based on their willingness to be part of a rigorous impact evaluation. While the sites were chosen to maximize a diversity of welfare reform approaches, locations, and labor market conditions, they were never intended to be nationally representative (Hamilton, 2002). That is, national in this example (and a number of other examples) does not mean that the results would apply to the target population nationally were it subject to the same treatment. Despite the NEWWS sample's limitations, the results from the NEWWS's study findings had a role in subsequent welfare policy changes nationally (Weaver, 2000). The labor force attachment model gained traction and became the preferred model for states to implement, whereas the human capital development model lost traction, being viewed as both less effective and more costly (Weaver, 2000). The results from this large, multisite evaluation were interpreted to predict similar impacts in many more places among many more populations than those from which the original findings were derived (Peck & Mayo, 2011).

Generally in social welfare policy, multisite experimental evaluations are used to estimate causal treatment impacts, while qualitative, narrative interpretations are used to discuss the generalizability of the results. For example, researchers routinely informally compare the characteristics of a study sample to that of a population of interest to conclude that the study sample "seems" similar to populations that live in urban areas or to a diverse set of places. It is our position—and that of other analysts struggling with increasing the external validity of experiments—that we must be able to do better than this sort of narrative interpretation.

In response, it is the intent of this article to show how new ideas and strategies for establishing external validity in experimental evaluations can be brought into the realm of social welfare policy; particularly, we emphasize new ideas and strategies that involve the use of stratified selection and the development of a targeted recruitment plan (Tipton, 2014a; Tipton et al., 2014). Evaluations in social welfare, as compared to education research, provide three unique challenges for this methodology which we address here. First, in education experiments, the inference population of interest in many studies is the population of *schools*. In social welfare studies, however, we show that there are often multiple possible inference populations, requiring the development of a strategy either to accommodate each or to focus on one of interest. Second, while national and state censuses of schools and districts are readily available for the creation of population frames in education, no such preexisting frames exist in social welfare. In this article, we develop such a frame for a social welfare study, illustrating the method for this policy arena. Finally, while several papers in education now exist and provide insight into the selection of covariates and development of strata, no such work exists in the welfare policy community. In response, we develop an example, discussing issues that may be encountered and implications for future welfare research and policy.

Sampling Sites for Generalizability

In order to address the increasing concerns regarding the generalizability of impact study results in education research, Tipton (2014a) and Tipton et al. (2014) recently proposed a new framework for sample selection aimed at generalization as well as strategies for achieving a balanced sample. In this section, we provide a review of this method, highlighting places where the application in social welfare may differ in marked ways from those in education.

The framework begins by requiring researchers to specify an inference population. This involves a discussion regarding which population is relevant or important as well as the creation of a population frame and the determination of inclusion and exclusion criteria. In some evaluations, it may make sense to define the population broadly, while in others it may be more sensible to define the population more narrowly. The basic principle here is that, by definition, the set of sites in a study is a sample from *some* population, and the goal is to determine in advance what the population should be. The inference population might be defined to include specific individuals—such as students, welfare recipients, and single mothers—or it might be defined in some aggregate-level unit—such as schools, districts, welfare offices, counties, or states, for instance. Importantly, this inference population may differ from the population frame used for recruitment. For example, schools may be recruited for the study, while the results consider the ATEs for a population of students.

Importantly, the determination of the relevant population and of inclusion–exclusion criteria will often have to do with features of the intervention under study. For example, the population may focus only on charter schools or on schools similar to those currently using the program under study (e.g., Tipton et al., 2014). In education, for example, existing data frames—state longitudinal data systems or the Common Core of Data (a national census of schools available through the National Center for Education Statistics)—are commonly used for this purpose. In social welfare studies, however, no such national or state frames exist. Instead, in this article, we develop such a frame for use in social welfare studies; to do so, we combine county-level aggregates of census data (including from the American Community Survey) with state-level data from agencies (such as the U.S. Department of Health and Human Services [DHHS] or the U.S. Department of Labor's Bureau of Labor Statistics [BLS]). Combining information across these sources is required here both because of the unique nature of administrative units (some states, some counties) and because the variables of interest are not available in a single source.

In defining the inference population, the researchers also highlight the ATE of interest for the study. For example, the goal may be to estimate the average effect of a welfare intervention across all counties in the United States. An alternative effect of interest might be that on welfare recipients. Next, researchers must select a set of covariates that is likely to explain variation in site-specific treatment effects: As previous work shows, in education research, this is typically school-ATEs or district-ATEs. In social welfare research, in contrast, this may be welfare-office-, county-, or state-level ATEs. This covariate list may be based on previous research on similar interventions (for example, that treatment effects are larger or smaller for minorities or sites in large urban areas) or on knowledge of the causal mechanism (for example, a program developed for unemployed men may not perform as well when administered to unemployed women). Ideally, this list should be based on empirical information on treatment effect heterogeneity, though often information of this type is not available. Of note, it is impossible to test in advance which covariates matter using this approach prospectively when a study has not yet been conducted. Tipton (2014a) argues that researchers should aim to be *bias-robust*: they should err on the side of supposing an impact estimation model that includes too many covariates rather than too few. Doing so helps guard against making the wrong assumptions.

Once a population and a set of covariates have been defined, the goal is to select a *balanced sample* (which differs from a *random* sample, as we will describe; Valliant et al., 2000), where by "balanced," we mean one that is like a "miniature" of the population (on this particular set of covariates). For example, if the population includes 20% urban sites, the goal would be for the experiment to have 20% urban sites. Should the average site in the population serve 9,000 people each year, then the goal would be for the average site in the experiment to serve 9,000 people. Likewise, if the population does not include small programs, then the goal would be for the sample to exclude small programs. One way to think of this is that the goal is to *approximate* in distributional terms a large random sample, with the caveat that a truly random sample would provide balance in expectation on unmeasured covariates as well. Because random sampling is generally infeasible (as evidenced by the dearth of random-selection/random-assignment studies), we assert that a sample balanced on a large set of covariates is a next best option, reasonable to implement in practice.

After establishing a framework and goals for sample selection—population coverage and compositional similarity—Tipton and colleagues recommend developing sample selection *strategies* that meet the balanced sampling goals. These methods have been implemented in the selection of school districts in evaluations of Open Court Reading and Everyday Math by the Southwest Educational Development Laboratory (Tipton et al., 2014) and are being used currently for the recruitment of schools in West Virginia for an evaluation of a mathematics program. Also, these methods are in the planning stages for use in the selection of schools in three studies evaluating science programs in New Jersey, Washington, and Virginia. In this approach, balance is achieved through the following steps: creating strata, allocating and ranking within strata, and making post hoc adjustments.

Creating Strata

The first step is to divide the inference population into relatively homogenous strata (in terms of the selected covariates). Strata play an important role in the strategic sample selection methods introduced by Tipton and colleagues for three reasons. First, researchers are familiar with stratification. In current practice and desire for diversity, researchers often select sites with varying levels of urbanicity (e.g., urban and rural), for example, or from various regions of the country (e.g., southeast, northeast, and west), as was the case in the NEWWS, discussed earlier. The goal here is to push researchers to stratify on a larger set of covariates and to do so with the clear goal of achieving a bias-robust and balanced sample. Second, stratification is simple to explain and understand, which is important for policy makers and general consumers of evaluation research. Experiments are often prized for their transparency and simplicity, which result from the use of random assignment, not complex statistical adjustments. The idea of stratifying a population is equally simple to explain and commonly used in survey sampling; this makes it appealing when sharing the results with a broader audience.

Third, stratification is a statistical tool that, under certain conditions, leads to covariate balance between groups. The most important of these conditions is that each individual stratum is relatively homogenous on the set of covariates that are hypothesized or empirically shown to explain treatment impact variation. For example, if we knew that site-ATEs varied only in relation to a single covariate, urbanicity, we might create two homogenous strata (urban and rural). Because the goal of the method is to be bias-robust, we need a method that works with a large number of both categorical and continuous variables. One method—and the method we will use throughout this article—as proposed by Tipton (2014a) is to use *cluster analysis* methods to create these strata (for more information on cluster analysis, see, e.g., Aldenderfer & Blashfield, 1984). With this cluster analytic method, as our illustration will show, a "distance" measure is specified and various numbers of strata are investigated with the goal of dividing the population into nearly homogenous groups. In practice, the number of strata chosen is based on a combination of factors, including the degree of homogeneity within the clusters, the total sample size (i.e., sample size $\geq$ number of strata), and the feasibility for recruiters.

Finally, it is important to highlight that strata are used here merely to increase the overall similarity between the resulting sample and population. As we will elaborate in the Conclusion, researchers may also be interested in estimating stratum-average treatment impacts and exploring variability across the strata. However, typically the overall sample size for a study is powered for estimation of the average treatment impact in which case stratum-specific estimates may be underpowered. This is not a shortcoming unique to the approach given here but is a limitation of most small and moderately sized evaluations.

Allocating and Ranking Within Strata

After the strata are defined, the second step is to allocate the number of sites needed in the experiment (n) to the strata. The total number of sites (n) typically is determined from a power analysis based on the population ATE estimator that will be used in the analysis, often based on a multilevel model (Raudenbush & Bryk, 2002).[1] Importantly, the approach to sample selection we define here impacts the power analysis through the selection of the intraclass correlation, which, in social welfare policy evaluations, tends to be relatively small (on the order of 0.03; Nisar, Klerman, & Juras, 2012).

After the power analysis and strata creation, the total sample size (n) is allocated to the strata. The minimal goal is that coverage is achieved—that at least one site is selected from each stratum. Tipton (2014a) shows that the ideal goal is to allocate proportionally the sample based on the proportions of the population in each stratum; doing so results in a self-weighting sample and the smallest standard error. For example, if Stratum 1 contains 15% of the population, then $n_1 = 0.15\ n$ sites will need to be recruited from Stratum 1 for the experiment. This is ideal since it means that analyzing results post hoc to represent some desired population is unnecessary. If, however, proportional allocation is not possible (which often occurs, given the real constraints and difficulties experienced in field recruitment) by ensuring that each stratum recruits at least one site, poststratification adjustments can be implemented.

Finally, within each stratum, units (e.g., schools, districts, states, counties, and offices) are ranked in terms of similarity to the average unit in the population in that stratum. This means that the site ranked 1 is the most "average" unit in the stratum while the site ranked 2 is the second most similar unit (perhaps deviating on one or two covariates). This ranking is based on a measure of statistical distance (e.g., standardized variables with Euclidian distance; see Tipton, 2014a, for a discussion) and includes only those sites that are eligible to be in the study. For example, the population might include sites that are currently using the program under study (Tipton et al., 2014), yet these sites are clearly ineligible to take part in an evaluation (because they already use the program). The recruitment team then is given a number of ordered lists (one for each stratum) and an associated sample size goal. For example, Stratum 1 may contain 47 units from within the population, and of these, 43 are eligible to be in the study. Recruiters then would be given a ranked list of the 43 eligible units (also including contact information) and directions to find four units (e.g., if that is the number stipulated in the cross-stratum allocation of n) that will agree to participate in the evaluation. This process is then repeated for the remaining strata.

Making Post Hoc Adjustments

Finally, at the end of recruitment, the resulting sample of sites taking part in the experiment can be compared to the inference population posited at the beginning of the study. The degree of balance between the experimental sample and population can be assessed for each of the covariates using the methods provided by Stuart, Cole, Bradshaw, and Leaf (2011) or Tipton (2014b). When imbalances remain, they can be adjusted for using the post-stratification methods provided by Tipton (2013) and O'Muircheartaigh and Hedges (2014). Importantly, because the sample was selected with these covariates and balance in mind, the resulting sample typically has less undercoverage than what would be expected when no planning has taken place. Additionally, the imbalances that remain are likely to be small, which is helpful since reweighting performs best when there is full coverage and the sample is already somewhat similar to the population (see Tipton, 2013).

Welfare Policy Illustration

To illustrate how this sampling approach would work in the welfare policy evaluation arena, we develop an example that shares features with many social welfare studies. These studies typically involve a large number of individuals across many sites. In the welfare arena, these sites might be offices, counties, regions, or states, and the individuals might be those receiving some form of income support, such as Temporary Assistance for Needy Family (TANF), or at risk for receiving support. For example, this article was stimulated by work on the DHHS/Administration for Children and Families–funded evaluation of job search assistance (JSA) strategies (JSA Strategies Evaluation, 2015), though in practice that evaluation went in a different direction than the design proposed here.

While the JSA evaluation ultimately used a different site selection strategy, we use it throughout as an example because it situates the discussion of inference populations, covariate selection, and strata development. The characteristics of the study provide useful context for considering external validity in the social welfare policy arena where historically evaluations have not drawn random samples or been deliberate about site selection in a way that supports generalization, either broadly or specifically to a designated inference population.

Throughout the illustration, we keep in mind that any evaluation of JSA must take into account that the scope and nature of services varies widely across states and localities. Variation in resource availability, caseload size, and service philosophy can result in quite divergent approaches to delivering JSA. For example, some TANF programs offer JSA in a group setting for a limited period of time, while other states or counties may engage in more individualized, longer term assistance. Some provide job search prior to TANF application, while others provide assistance only to those who are receiving TANF. The scope and nature of these activities varies further depending upon the "job readiness" of the TANF recipients.

The remainder of this section applies the steps described above for selecting sites for generalizability to this example. As noted, throughout we use the example of a JSA study to illustrate the implementation of this approach more broadly, including issues unique to the social welfare context (as compared to education research). First, we define an inference population that may be appropriate for an evaluation such as this one; we discuss the fact that there may be *multiple* inference populations of interest in social welfare and provide an approach that may be applicable more generally. Second, we define variables potentially associated with cross-site variation in the magnitude of impacts that a program related to welfare JSA might cause. Third, we create strata that are relatively homogenous on these variables within each stratum and relatively heterogeneous between strata, discussing the implications of the specific strata for site recruitment to the study.

Inference Population

Use of the balanced sampling methodology requires that we start by defining carefully the population of interest. This includes considering *to whom and/or where* study findings might usefully be generalized. Recalling that large-scale evaluations are focused on the estimation of an ATE, this requires us to ask: *for whom* is the ATE of interest? Answering this question in any particular study largely has to do with federal policy priorities (particularly when they are funding the research) and the administrative structure of the program. We now introduce three possible inference populations, focusing on the case of a welfare program's JSA. While these populations are similar, they differ in how administrative units are weighted in the final analysis and these weights have implications for stratum allocations (see Online Supplementary Materials for greater detail).

With this in mind, the first possible inference population is that of the United States. Focusing on this as the inference population would allow an evaluation to estimate the ATE of a program across the United States,[2] which in welfare policy tend to be the drivers of policy decisions. This framing would be preferred if the goal was to inform *state* policy decisions regarding the program. However, as noted, states are not the only decision-makers, leading to the second possible population, that of TANF administration units (because these units typically administer welfare policies and programs). The TANF program is governed at the state, county, and local level depending upon the particular state or region. Defining the inference population as administrative units would allow the estimation of the ATE across TANF administrative units. This would be of policy interest if the audience for evaluation findings was specifically TANF administrators. While the first (state) inference population is relevant to state-level decision-making, this second inference population would be relevant to TANF administration-level decision-making. For example, policy makers at the administration level might be interested in knowing if they should adopt a particular welfare program variant. The third possible frame is that of the individual TANF recipients (because they are the "consumers" of welfare JSA programs). From the federal perspective, it seems warranted to want to generalize to all individuals who receive TANF nationwide, regardless of the jurisdiction where they reside. If individuals form the inference population, then the specific state or administrative units where they access welfare are incidental to their effects and moreover become a nuisance in the sampling process.

The choice of inference population is also driven by the ability to build an appropriate population frame using existing data. In education research, the most comprehensive frame is the Common Core of Data, which includes demographic information on the population of schools, districts, and local education agencies throughout the United States. Alternative frames include state-specific data from state longitudinal data systems (e.g., Stuart et al., 2011; Tipton, 2013) or information on a program's current user base often available from a program's publisher (e.g., Tipton et al., 2014). In social welfare policy, abundant state-level data are available,[3] but administrative unit frames do not exist in the same form. A focus of this article, therefore, is the development of one such frame for the evaluation of a welfare JSA program; we speculate that this frame will be useful for other social welfare studies as well and suggest that a contribution of this article is to provide that population frame for others' application (see Online Supplementary Materials). To do so, we combine state- and county-level data from various government agencies—for example, the Census and the BLS. Importantly, as in education, these frames focus on the aggregate level since no publicly available frames that list individual clients exist.

Given that the populations under consideration are nested (i.e., individuals within administrative units within states), the decision to make one population primary over the others largely has implications for the degree of stratification and for the optimal allocation of sites to strata (i.e., since each of the 50 states do not have equal sized TANF populations). In this illustration, we focus on the administrative units as the *primary* inference population, while leaving open the possibility of state-level generalization (a *second* inference population) using post hoc reweighting adjustments.[4] The population of individual TANF recipients seems inappropriate, given that most welfare programs occur at the administrative unit level (not the individual level)—that is, TANF offices run the programs. In order to establish this sample with primary and secondary inference populations, we employ a two-stage sample selection method: first, stratifying states, and second, administrative units. This two-stage approach allows for administrative units within the same state to remain in the same stratum, which is important for practical and cost reasons in an evaluation's recruitment work (e.g., travel logistics and costs for evaluators). Future users of the sampling frame that we have created may decide to prioritize states over administrative units or choose to use only some of the counties in some states or others depending on the specific application.

Considering administrative units as the sampling frame could inform future social policy evaluations because it explicitly recognizes the administrative unit as key to policy decisions and that these units exist within a state policy structure. The geography of TANF administrative units, including state- and county-administered regions, is shown on the map in Figure 1. The map also identifies the stratum to which each administrative unit (state or county) belongs; the development of these strata and further explanation follows. In total, 40 administrative units are states and 762 are counties within states. Some of these counties are proxies for a locality (e.g., "New York County" for New York City in NY). For simplicity, we include each county within a state as an administrative unit if that state is designated as one where welfare policy has devolved administration. We recognize that in practice, the balance of power may differ and that using our frame might overstate counties' weight within the population.

Site Characteristics for Stratification

After selecting the population frame of TANF administrative units, we turn to the selection of the covariates needed for stratification. Recall that the goal is to include all factors that explain variation in the differential impacts of the JSA programs across administrative units. Ideally, one would define strata based on evidence of impact heterogeneity from studies of similar or related programs. In practice, however, little is known about treatment effect heterogeneity, leaving researchers to select these features based upon theory. When this condition has been met and sites are picked to ensure compositional similarity with the population, the ATE estimated in the experiment is unbiased for the population. In situations when this condition has not been met fully—when there is one or more omitted dimensions of impact variation—this bias is typically reduced relative to other nonrandom site selection approaches but not eliminated. With this in mind, we select variables in the following categories: labor market conditions (e.g., availability of jobs for low-skilled workers, unemployment rates), geography (e.g., measures of urbanicity and population density), and the state's policy regime. Certainly, empirical evidence exists to justify the selection of these variables for framing the sample (e.g., Card, Kluve, & Weber, 2010). Conceptually, the following justify why these are the suitable variables upon which to group states for the purpose of a specific program evaluation:

- *Labor market (unemployment; conditions for low-skilled workers)*: We speculate that in places where there are relatively more opportunities for low-skilled workers and unemployment is low, TANF recipients served through JSA programs will be more likely to find

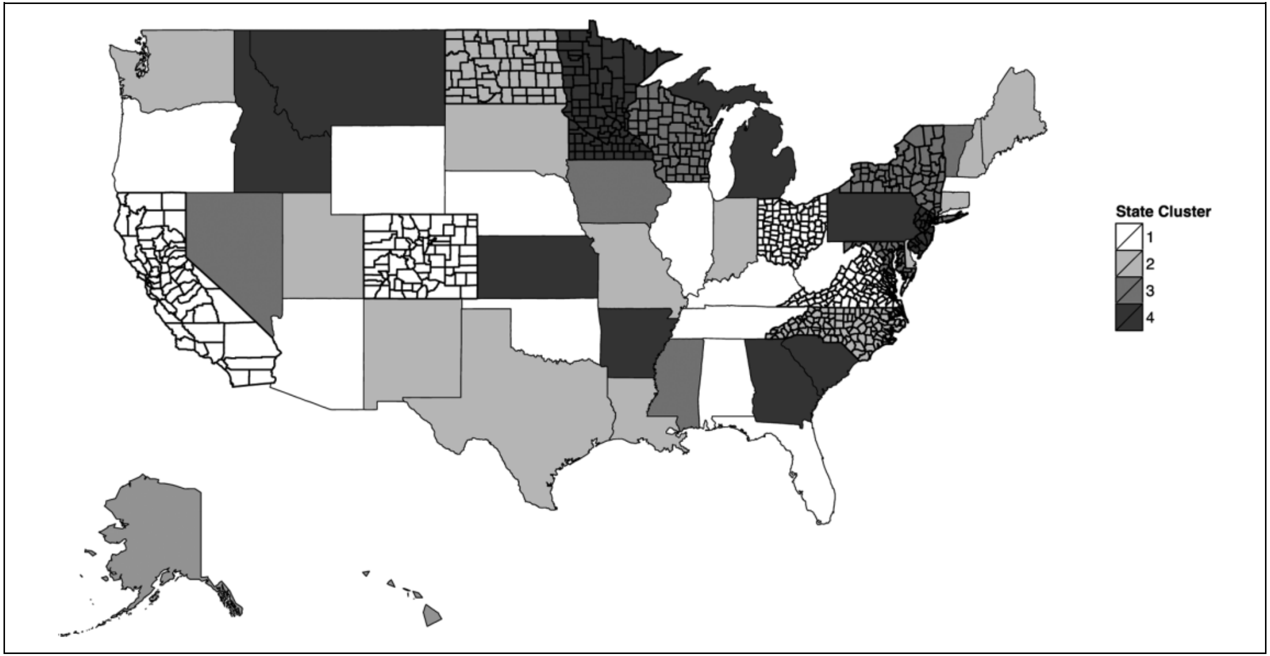

Figure 1. Map of Temporary Assistance for Needy Family administrative units by sampling stratum.

work sooner and possibly also at better wages than those who do not have access to such services.

- *Geography*: It is possible that people who live in denser places or urban areas will have greater access to job opportunities, potentially increasing the relative effectiveness of a TANF program. Conversely, it may be that greater density instead could decrease the effectiveness of a program.
- *Policy regime*: We speculate that individuals in places that offer more supportive job search programs, in the context of more supportive (less punitive) TANF programs, should show greater success in finding work.

While these are the categories we believe are important, one of the largest difficulties in sample selection is determining how to operationalize it using existing data. In building the population frame, we pool data from several sources for the units available, be they state or county (as TANF administrative units). For the labor market measures, for example, the BLS provides current employment numbers by job category and state (i.e., Occupational Employment Statistics Survey) as well as national projections for job growth (through 2022) by job category and years of education (BLS, 2014). We matched these two data sources to calculate the number of jobs projected to be available for low-skilled workers (those with a high school [HS] degree or less) and for skilled workers (those with greater than a HS degree) by 2022 within each state. Based on these numbers, we created two measures: (1) the projected proportion of jobs in a state in 2022 to be held by low-skilled workers and (2) the percentage increase in jobs for low-skilled workers (from 2012 to 2022). As relative numbers, these measures do not depend on the size of the state. While these figures are state-level measures, we apply them to every administrative unit in a state (the entire state or all counties). We also include the county-level or state-level unemployment rate for 2012 as a variable for stratification.

Next, we posit that geography matters. The ability to find and secure a job is often a function of the density of a state in terms of both population and degree of urbanicity. We used two state-level measures to capture this: (1) the population density (i.e., total population divided by area) and (2) the degree of urbanicity (i.e., total urban area divided by total land area of state). The difference between these measures is highlighted by the state of New York versus the District of Columbia: While both have high population densities, only a small portion of New York state is actually urban (23%) compared to DC (100%).

The third set of measures is about the state policy regime. State TANF-sanctioned and job search–specific policies shape each program's features and therefore its potential effectiveness. We use the DHHSs' data on two specific policy measures for stratification: (1) whether or not a job search program is mandatory for TANF recipients designated as "work eligible" and (2) the type, severity, and duration of TANF sanctions for lack of adherence to program rules. We classify the TANF sanction policies into three groups: partial, full, and closed. States that issue partial sanctions are the least strict in that recipients who do not comply with program rules forfeit only a portion of their TANF benefits. Those that impose full family sanctions result in complete loss of benefits until the recipient comes into compliance and some states close the welfare case for some period until compliance and reapplication occur. We anticipate that the effectiveness of programs might vary depending on their location within a state that uses each of these three approaches to sanction.

We believe that our audience of potential users of the data set—will be interested in the general demographic traits of the strata created for sampling purposes. In response, we compile data on selected measures—in addition to the above-discussed labor market, geographic, and policy measures—within our sampling frame, including the 2012 population size, number of families projected to enter TANF in a 12-month period, and total number of TANF recipients as of 2012. Also, included are proportion of the population that was non-White, ratio of projected job growth for low-relative to high-skilled workers, proportion of the low-skilled population aged 25 years or older in poverty, and the proportion of low-skilled population by age. We do not include these variables in the stratification design because we had limited reason to believe that they would moderate the *effectiveness* of a JSA program (though they may be related to *outcomes*). Descriptive information on these variables and their data sources appears in Table 1, and the full data set is available as an Online Supplementary Materials.[5]

Stratum Creation: Cluster Analysis

While we selected the population of administrative units as our primary inference population, we also wanted to potentially generalize to the population of states and therefore stratified the population in two stages. First, at the primary sampling unit stage, we divided the 51 states (including the District of Columbia) into four clusters. Second, within each of these four clusters, we then divided the administrative units (some states, some counties) into two strata each. This strategy guaranteed that administrative units

Table 1. Characteristics of Units in the Sampling Frame.

Variables	Level	Source	Mean	Minimum	Maximum	*SD*
Labor market						
• Available jobs	State	BLS (2012a)-Employment	39.4	7.8	67.6	15.2
	State	BLS (2012a)-Employment	7.7	5.8	9.7	1.1
• Job growth	State	BLS (2012a)	6.7	2.9	9.8	1.6
• Unemployment rate	County	BLS (2012b)-Occupational, using	7.2	0.9	24.5	2.4
• Relative low-skill job growth[a]	State	NCAIS codes	81.6	55.7	153.8	23.8
Geography						
• Density	State	U.S. CB-Geography (2010a)	74.1	38.7	100.0	14.9
• Urbanicity	State	U.S. CB-Geography (2010a)	9.2	0.05	100.0	16.6
Policy conditions						
• JSA mandatory policy (% yes)	State	Kassabian, Huber, Cohen, and Giannarelli (2013, Table L7)	39.2	0.0	1.0	0.5
• TANF sanction policy	State	Kassabian et al. (2013, Table I.A.2)				
– % Partial			9.8	0.0	1.0	0.3
– % Full			39.2	0.0	1.0	0.5
– % Closed			51.0	0.0	1.0	0.5
• TANF caseload size (000s)[a]	State	HHS-OFA	78.6	1.1	1,367	195
• 12-month TANF JSA intake (000s)[a]	State	Authors' estimates from HHS-OFA	5.7	0.7	154.2	21.5

(continued)

Table 1. (continued)

Variables	Level	Source	Mean	Minimum	Maximum	*SD*
Demographics						
• Population (000s)[a]	State	AFF-People	6.2	0.5	38.3	7.0
	County	AFF-People	0.5	0.0	38.3	2.3
• Percentage of White population[a]	State	AFF-People/race	79.5	26.1	95.4	13.3
	County	AFF-People/race	14.6	0.2	95.4	30.4
• Percentage of poor population[a]	State	AFF-People/poverty	14.7	9.7	23.3	3.2
	County	AFF-People/poverty	14.3	3.0	34.1	5.2
• Age distribution among those w/HS or less[a]	State	AFF-People/education				
– 18–24 years, %			3.8	2.1	5.6	0.8
– 25–44 years, %			45.5	42.7	48.4	1.0
– 45+ years, %			43.1	40.5	45.0	1.0
Education (% < HS)[a,b]	State	AFF-People/education	13.0	2.3	47.0	5.8

Note. Number of observations for states is 51. Descriptive statistics for counties aggregate over administrative units which includes 40 states and 762 counties. *SD* = standard deviation; BLS-Employment = Bureau of Labor Statistics, Employment Projections; BLS-Occupational = Bureau of Labor Statistics, Occupational Employment Statistics; CB = Census Bureau; HHS-OFA = U.S. Department of Health and Human Services, Administration for Children and Families, Office of Family Assistance; AFF = American Fact Finder Survey; JSA = job search assistance; TANF = Temporary Assistance for Needy Family; HS = high school.

[a] Included for descriptive purposes and not as part of the cluster analysis to create strata. [b] Among those 25 years and older.

in the same state would be in the same cluster, which matters to the evaluation for practical and cost reasons. This resulted in a total of eight strata. In what follows, we further explain how these strata were created and their properties.

In both stages, following Tipton (2014a), we create nearly homogenous strata using *k*-means cluster analysis methods implemented using the statistical program R (R Development Core Team, 2014). The *k*-means method requires that a distance metric be specified, and since our covariate list includes both continuous and categorical measures (i.e., sanction policies), we use Gower's distance (Gower, 1971) implemented using the *daisy* function in the *cluster* package in R Version 2.0.1 (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2015). Gower's distance is preferable when there are many categorical variables (as are found in this example) because it prioritizes these variables in the creation of strata (we refer the reader to Tipton, 2014a, for further discussion). This approach ensures that states and administrative units are essentially "replacements" on key categorical variables, differing only on continuous covariates. This is important since it is easier to adjust for differences in continuous covariates than on categorical variables.

Further, although multiple cluster "solutions" are feasible, our goal for the number of state clusters is three to seven. While more clusters typically leads to greater homogeneity, the range we focus on here is based on recruitment feasibility. For recruiters, adding strata increases their constraints, requiring them to focus more effort on getting particular sites, not just achieving the required sample size. In response, we focus on the range of three to seven clusters because we perceive this range to be accommodating to recruiting demands and consider that there will be an additional level of stratification into administrative units. At first, we examined the three-, four-, and five-cluster solutions and observed that these clusters explained 68.8%, 75.9%, and 82.7%, respectively, of the variation in the state-level covariates. Because the four- and five-cluster solutions differed only minimally, we decided to focus on the four-cluster solution; these clusters contained 9, 16, 15, and 11 of the states, respectively.

Within these four state-based clusters, we further divided the administrative units into two strata each. This means that within clusters, states, and counties were treated equally. For example, in Stratum 3 of Table 2, while Alabama is state-administered, Colorado is administered at the county level. In this data set, most of the covariates were constant at the state level, with only unemployment varying at the county level. Using two within-cluster strata further explained approximately 63%, 46%, 84%,

Table 2. Characteristics of Eight Administrative Unit Strata.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	Low Pressure		Low–Moderate Pressure		Moderate Pressure		High Pressure	
Variables	1	2	3	4	5	6	7	8
Labor market								
• Available jobs	50.4	55.3	45.3	36.0	59.9	10.1	54.0	51.6
• Job growth	7.6	6.6	7.0	8.2	6.5	8.7	7.8	6.5
• Unemployment rate	7.6	7.2	7.0	10.2	8.6	3.8	8.0	5.6
Geography								
• Density	57.0	46.9	48.2	72.3	40.9	23.4	85.7	38.3
• Urbanicity	20.6	8.6	16.7	11.9	9.0	2.0	43.6	2.7
Policy conditions								
• JSA mandatory policy (% yes)	100.0	100.0	0.0	0.0	0.0	0.0	100.0	100.0
• TANF sanction policy[a]								
– Partial, %	98.0	0.0	0.0	97.0	0.0	0.0	0.00	0.00
– Full, %	1.5	100.0	100.0	3.3	0.0	0.0	0.00	0.00
– Closed, %	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0
State administered	DC, MD, MO, and VT	HI, IA, and NV	AL, AZ, FL, IL, KY, MA, NE, OK, OR, TN, WV, and WY	n.a.	IN, LA, ME, MT, TX, UT, and WA	CT, DE, NH, NM, RI, and SD	AK, AR, GA, ID, KS, and PA	MI, MS, and SC
County-level administered (number of counties)	NY (62)	WI (72) and MD (23)	CO (62), OH (88), and VA (133)	CA (58) and CO (2)	NC (100)	ND (53)	NJ (20) and MN (2)	NJ (1) and MN (85)

Note. Variables reported in earlier tables for descriptive purposes are excluded here for parsimony. JSA = job search assistance; TANF = Temporary Assistance for Needy Family.

[a] Values may not be equal exactly 100% due to rounding.

and 70% of the administrative unit–level variation within Clusters 1–4, respectively. Altogether this resulted in eight strata, which are defined in Table 2. Interestingly, with the exception of two counties in Colorado (i.e., Broomfield and Denver), one county in New Jersey (i.e., Hunterdon), and two counties in Minnesota (i.e., Hennepin and Ramsey), all of the counties in county-level administered states were stratified together. This implies a recruitment approach that can be practical in the field.

At the administrative unit level, these eight strata vary in several regards. First, note that the four clusters can be characterized in relation to the policy regime variables (low pressure, low–moderate pressure, moderate pressure, and high pressure). Within the "low–moderate pressure" state cluster (Cluster 2), the administrative Stratum 4 contains all of the 58 counties in California and 2 of the counties in Colorado, with the remainder of the states classified into Stratum 3. Within the "moderate pressure" state cluster (Cluster 3), the 15 states were roughly split in half between Strata 5 and 6. States in this cluster typically have the smallest populations (not shown), and of these, those with the smallest administrative units and the least population density are found in Stratum 6. Of note, these administrative units have the lowest projected proportion of low-skilled jobs but the highest projected job growth (higher even than that projected for high-skilled workers). In the "low pressure" state cluster (Cluster 1), the nine states were split roughly in half across the two strata. Stratum 1 includes states with larger urban areas (e.g., NY and DC) and higher expected job growth relative to Stratum 2, whose units have lower degrees of urbanicity and more moderate job growth. Finally, in the "high pressure" state cluster (Cluster 4), Stratum 7 includes states and administrative units that are the most population dense and urban, while Stratum 8 includes those that are the least. Both have low-projected 12-month TANF enrollments, however (not shown). In Table 3, we further characterize these strata by their most salient features, which, we think, combine with the low-to-high pressure designation of the state-formed clusters to add nuance to the characterization of the strata of administrative units.

Stratum Details: Sample Allocation, Site Selection, and Post Hoc Adjustment Costs

Sample allocation. In order to determine how many units should be included in the experiment from each stratum, the first step is to determine the population proportions in each stratum. Table 4 lists these proportions for the population of administrative units (our primary inference population). In

Table 3. Characterization of Eight Administrative Unit Strata.

Variables	1	2	3	4	5	6	7	8
Sanctions	Partial or full				Closed			
JSA	Mandatory		Not mandatory		Mandatory		Not mandatory	
Population	M		Largest		Smallest		M	
Density	M	L	L	H	L	Very L	Very H	Very L
Urbanicity	H	L	H	H	L	Very L	Very H	Very L
Unemployment	M	M	M	H	H	Very L	M	L
Job growth	H	M	M	H	L	H	H	M
JSA 12-month enrollment	M	M	M	Very H	M	L	L	L

Note. L = low (light shading); M = moderate (no shading); H = high (dark shading); JSA = job search assistance.

Table 4. Population Proportions and Sample Allocation.

Variables	1	2	3	4	5	6	7	8
Total administrative units/states	66/5	97/4	295/15	60/1	107/8	59/7	28/7	89/4
Administrative unit population	0.082	0.121	0.368	0.075	0.134	0.074	0.035	0.111
Allocation	6	5	18	1	9	8	8	5

Note: This table assumes total of $n = 60$ administrative units in the experiment.

the Online Supplementary Materials, we also include the proportions for the other two possible inference populations as well as a discussion of alternative allocation schemes. Importantly, based on the eight strata, the ideal in our hypothetical exercise would be to select sites for inclusion into the evaluation, so that the same proportions in the population are found in the sample. Table 4 proposes how those units might be selected under each allocation for a study with 60 sites. For example, in Stratum 3, 18 administrative units (including counties or states) would need to be selected out of a possible 295 for the administrative unit population.

Finally, note that if proportional allocation is not possible, then post-stratification reweighting could be used instead to adjust for compositional differences. To do so, at least one site would need to be selected in each of the eight strata. In field application, this minimal requirement would be more achievable than proportional allocation for any given resource investment. For example, these strata could then be used at the tail end of a study's recruitment process to make a final recruitment push in order to ensure complete coverage. If this were not possible—that is, if one or more strata were not represented in the final sample—it would be possible to redefine the inference population to a still-meaningful concept, a point we return to later in the Conclusion section.

Site selection. While the final eight strata deduced here are relatively homogenous in terms of their policy conditions, there remains some within-stratum variability in the labor market and geography-related covariates. Tipton (2014a) shows that in order for the resulting sample and population to be balanced, the soundest strategy is to give priority to stratum "average" units when selecting the sample. Within each stratum, units are ranked from most to least similar to the stratum average unit. In the ideal selection process, recruiters would first approach top ranked units, and if they decline participation, move on to lower ranked units. These rankings can be used to provide us with a measure of how "representative" the units actually selected in each stratum are of the stratum-specific population of administrative units.[6]

Post hoc adjustments. In an example, Tipton (2014a) illustrates a case with a high refusal rate in which the resulting sample remains compositionally similar to the population. It is possible, however, that in practice very high refusal rates can result in imbalances; this is particularly true if strata are very heterogeneous. Providing that there is at least one site in the experiment in each of the eight strata, post hoc adjustments can be used to reweight the final sample to be compositionally similar to the inference population(s). One simple tool for this purpose is that of poststratification (Tipton, 2013), which we mention in our introductory section and is further detailed in the Online Supplementary Materials.

Discussion of a Middle Ground Approach

The stratified recruitment plan we developed here as a welfare evaluation example is distinctive in that it considers both generalization and evaluation practicalities. As such, it might be helpful to understand how the strategy compares to others in the field. We see this approach as somewhat of a compromise strategy, allowing improved generalizability but with fewer barriers or costs to implementation. This is in contrast to random sampling, which has the best statistical properties but is very rarely used, especially in social welfare program evaluations, given the high refusal rate commonly found in practice. It also is in contrast to another extreme, convenience sampling, which is much more commonly implemented. In convenience sampling, units are often selected based entirely on cost. For example, one method to achieve a large sample of TANF recipients is to recruit a small number of sites with very large 12-month JSA intake. As Table 2 shows, however, this strategy would likely result in very few units being selected in Cluster 3 (moderate pressure), where the intakes are typically low. Another variant of convenience sampling also might involve selecting only those places that "want" to participate in the evaluation. The size criterion makes the resulting sample unrepresentative as does the "wants to participate" criterion. It should be obvious that the unmeasured characteristics of the sites that want to participate are likely highly correlated with the program's likely success in those places, implying that no amount of post hoc adjustment could overcome the bias in generalizing from those sites to some larger population.

In between the random sampling and convenience sampling approaches is a third commonly used approach, which is to stratify based upon a single covariate, typically geographic region or urbanicity. Unlike convenience sampling, this approach is used explicitly to improve generalizations. However, while this approach increases the "face validity" of the study, in the framework provided here, it only reduces bias if treatment effects are thought to vary purely in relation to region or urbanicity. As Figure 1 illustrates, however, when creating strata in relation to a theory of treatment effect variability, these strata do not align well with region at all.

As a compromise strategy, the success of this method hinges on how well it is implemented. First, in order for the treatment effect estimated to be unbiased for the population, the set of covariates used for stratification must be complete—it must contain all those that explain treatment effect variability. Clearly random sampling does not require this, though as noted, is rarely used. Convenience sampling and geographic sampling also do not require the specification of stratifier variables, but only because they lead to ill-formulated generalizations based on analogies and description instead of statistics. For this reason, it is important to qualify all claims of generalization with this method by stating the stratifying variables used and potentially omitted variables.

Finally, the purpose of the strata is to generate a sample that is compositionally similar to the population on these covariates. When this similarity is sufficient, the population average treatment impact estimated in the study is unbiased (assuming all variables explaining treatment effect heterogeneity have been included). The success of the method thus depends on both how homogenous the strata are and, when there is within-stratum variability, on how similar those units selected within strata are to the stratum mean.

Conclusion

This article has described the application of a new, design-based approach for crafting a sample in experimental evaluations of social welfare-related programs, so that generalizing beyond the study sample to a larger population of interest is possible. In this article, we focus on the implementation of these methods in welfare policy. Here, we briefly review two unique issues that arose as well as future directions for research in this area.

Population Definition

When we set out to implement this sample selection plan in a social welfare example, we were surprised that defining the inference population was not straightforward. This first step, in other contexts, is generally straightforward; in this case, we easily identified three possible inference populations (states, administrative units, and welfare recipients), which are associated with three different ATEs of potential interest to distinct audiences. The decision about inference population is an important one, and it requires considering the potential implications of the research findings for different policy audiences during a study's design phase, not post hoc. Although we create an administrative unit sampling frame, we also recognize that the frame could be modified for use as a state-level frame as well.

Additional Uses for Strata

Strata are used primarily as a tool for achieving a balanced sample in the method detailed here, with the goal of reducing bias in the estimate of the population average treatment impact. It is easy to see, however, that two additional uses for the strata may be as important.

First, if full coverage is not possible—as may be the case in welfare population evaluations—then the clear definitions of the strata allow for a new, restricted inference population to be clearly defined. For example, if no units were available in Stratum 8 in our example, then we could simply define the ATE estimated as one that could generalize to the population of administrative units, excluding units in nonmandatory closed-sanction states with moderate nonurban, low-density populations.

Second, in addition to a single ATE estimate for the population, stratum-specific treatment effect estimates can be provided. If the units selected within each stratum are representative of the population units in the stratum (as occurs if they are highly ranked), and if these estimates differ, then it may be prudent for administrative units to turn to their stratum-specific ATE to make decisions regarding the policy implementation and implications. While in most studies these stratum-specific effects will be underpowered greatly, they come closer to answering questions about whether a program will work in the particular conditions encountered in a particular administrative unit. This approach could be helpful in situations in which the treatment is highly effective in one stratum but not effective in another (thus leading to a small population ATE).

Future Directions

We believe that this method has the potential to improve welfare policy research and evaluation. One of the largest impediments to implementation in this field is the lack of a clear population frame; this article, we hope, provides researchers with the tools to build an appropriate frame in social welfare studies and a model of how one might go about doing so in other domains (beyond education and social welfare studies). Another concluding lesson is that timing is important for the implementation of any sampling method. It is most successful when developed early in the study design process. This is important whether the strategy would be carried out in advance or not; indeed, engaging in this work in advance may reveal how best to consider even post hoc adjustments to enhance a study's external validity. Even if not implemented as a sampling strategy, the elements of the framework can help facilitate clarity in the discussion of generalizing study results to various potential populations, interested stakeholders, and possible shortcomings.

It is our hope that the framework and analysis presented here will provide future evaluations of social welfare policies and programs with a way to think about sampling that also considers generalization. Even without a purely random sample, opportunities for cluster-stratified sampling exist, where—with or without full coverage in practice—study findings have the potential to have greater external validity than is currently and commonly the case.

Acknowledgments

We recognize input from attendees at the 2014 Welfare Research and Evaluation Conference the Abt Associates Journal Author Support Group, including Jacob Klerman and Rob Olsen; and three anonymous reviewers. We thank Lauren Fellers for assistance in creating the population data frame found in the Online Supplementary Materials. We are grateful to Stephen Bell for substantive and editorial input, especially with regard to the Job Search Assistance Strategies Evaluation, for which he serves as coprincipal investigator. All errors or omissions remain the authors' own.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online data supplements are available at http://journals.sagepub.com/doi/suppl/10.1177/0193841X16655656

Notes

1. In cluster randomized designs, an additional benefit of this approach is that if researchers base their power analysis on a design with no strata but then later actually randomly assign units within the strata, the resulting design has even greater power.
2. We use the term "state" here, but should the population frame include other U.S. territories as units (e.g., the District of Columbia, Puerto Rico, Guam, and the Northern Mariana Islands), then those would be included here too as units that are part of the frame.
3. These include Urban Institute's Welfare Rules Database (http://anfdata.urban.org/wrd/WRDWelcome.cfm), University of Kentucky's Poverty Center's state data (http://www.ukcpr.org/AvailableData.aspx), Annie E. Casey Foundation's Kids Count data (http://datacenter.kidscount.org/), and National Center for Children in Poverty's 50-State Policy Tracker (http://www.nccp.org/tools/policy/).
4. In the Online Supplementary Materials, we provide a discussion of optimal allocation methods when a study needs to be designed to estimate average treatment effects for two or more populations.
5. A recent meta-analysis of job search assistance (JSA) studies shows that age of participants typically matters to the impact of JSA (Liu, Huang, & Wang, 2014).

However, because age distributions of Temporary Assistance for Needy Family participants vary minimally across states and counties, we found that it was not helpful for stratification purposes.

6. In the Online Supplementary Materials, we include a table listing summary measures for the first and last 10 units within each stratum. To see how this would influence site recruitment in a hypothetical welfare policy evaluation, in Online Table S2, we include summary measures for the first and last 10 units within each stratum.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage.

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*, 318–335.

Bureau of Labor Statistics. (2012a). *Employment Projections* [data file and code book]. Retrieved from http://data.bls.gov/projections/occupationProj

Bureau of Labor Statistics. (2012b). *Unemployment Estimate* [data file and code book]. Retrieved from http://www.bls.gov/lau/lastrk13.htm

Bureau of Labor Statistics. (2014). *May 2013 State Occupational Employment and Wage Estimates* [data file]. Retrieved from http://www.bls.gov/oes/current/oessrcst.htm

Card, D., Kluve, J., & Weber, A. (2010). Active labour market policy evaluations: A meta-analysis. *The Economic Journal*, *120*, F452–F477. doi:10.1111/j.1468-0297.2010.02387.x

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*, 527–536.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857–872.

Greenberg, D., & Shroder, M. (2004). *The digest of social experiments* (3rd ed.). Washington, DC: The Urban Institute Press.

Hamilton, G. (2002). *Moving people from welfare to work: Lessons from the National Evaluation of Welfare-to-Work Strategies*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation and U. S. Department of Education Office of the Under Secretary, Office of Vocational and Adult Education.

Hamilton, G., Brock, T., & Farkas, J. (1994). *The JOBS evaluation: Early lessons from seven sites*. Washington, DC: U.S. Department of Health and Human

Services, Administration for Children and Families, Office of the Assistant Secretary for Planning Evaluation and U.S. Department of Education, Office of the Under Secretary, Office of Vocational and Adult. .

Job search assistance (JSA) strategies evaluation; Notice of proposed information collection activity; Comment request. (2015, July 15). *Federal Register*, *80*, 41505–41506.

Kassabian, D., Huber, E., Cohen, E., & Giannarelli, L. (2013). *Welfare rules databook: State TANF policies as of July 2012: Final Report, Table I.A.2 (pp.42-44) & Table L7* (pp. 214–220). Washington, DC: The Urban Institute.

Liu, S., Huang, J. L., & Wang, M. (2014). Effectiveness of job search interventions: A meta-analytic review. *Psychological Bulletin*, *140*, 1009–1041.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2015). *Cluster analysis basics and extensions*. R package version 2.0.1.

Nisar, H., Klerman, J. A., & Juras, R. (2012). *Estimation of intra class correlation in job training programs* (working paper). Bethesda, MD: Abt Associates.

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*, 107–121.

O'Muircheartaigh, C. A., & 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.

Orr, L. L. (2015). 2014 Rossi Award lecture: Beyond internal validity. *Evaluation Review*, *39*, 167–178. doi:10.1177/0193841X15573659

Peck, L. R., & Mayo, A. (2011, November 3). *An empirical assessment of external validity*. Presented at the Association for Public Policy Analysis and Management (APPAM) Conference (Panel: External Validity in Randomized Experiments), Washington, DC.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Volume 1 of Advanced Quantitative Techniques in the Social Sciences. Thousand Oaks, CA: Sage.

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, Part 2*, *764*, 369–386.

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. (2014a). Stratified sampling using cluster analysis: A balanced-sampling strategy for improved generalizations from experiments. *Evaluation Review*, *37*, 109–139.

Tipton, E. (2014b). How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics*, *39*, 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*, 114–135.

U.S. Census Bureau. (2010a). *2010 Urban and rural classification and urban area criteria* [data file and code book]. Retrieved from https://www.census.gov/geo/reference/ua/urban-rural-2010.html

U.S. Census Bureau. (2010b). *American Fact Finder Survey: Community facts—Education* [data file]. Retrieved from http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: A prediction approach* (Wiley Series in Probability and Statistics). New York, NY: John Wiley & Sons.

Weaver, R. K. (2000). *Ending welfare as we know it*. Washington, DC: Brookings Institution Press.

Author Biographies

Elizabeth Tipton is an assistant professor of applied statistics at Teachers College, Columbia University. Her research focuses on issues of generalizability in the design and analysis of large-scale experiments and meta-analysis.

Laura R. Peck is a principal scientist at Abt Associates Inc. in the social and economic policy division. Dr. Peck specializes in innovative ways to estimate program impacts in experimental and quasi-experimental evaluations and applies this to many social safety net programs.