Abstract Title Page

Not included in page count.

Title: How generalizable is your experiment? An index for comparing samples and populations

Authors and Affiliations: Elizabeth Tipton, Teachers College – Columbia University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:
Recent research on the design of social experiments has highlighted the effects of different design choices on research findings. For example, the within-study comparison design literature has illustrated how and why the results from randomized experiments may differ from those found using the regression discontinuity design (Shadish, Galindo,Wong, Steiner, & Cook, 2011). Similarly, this literature suggests that another reason that research findings may vary has to do with differences in the target populations of different studies (Cook, Shadish, & Wong, 2008). By exploiting a case in which both an experiment and a wide scale policy change using the same curriculum were rolled out, Stuart, Olsen, Bell, and Orr (2012) were able to show that external validity bias can be as large as that due to the internal validity problems found in observational studies.

Since experiments rarely collect their samples using random selection, in order to address these external validity problems and design choices, recent research has focused on two areas. The first area is on methods for adjusting experimental treatment effect estimates for known population characteristics (Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, in press). A second research area has been on how to select a sample (non-randomly) to improve generalizations to a well-defined inference population (Tipton et al., 2012; Tipton, 2012).

Purpose / Objective / Research Question / Focus of Study:
Like the method for assessing generalizability developed by Stuart et al (2011) this paper proposes a method for evaluating differences between the experimental sample and population based upon a propensity score. Unlike the Stuart et al (2011) method, however, the index we propose does not require outcome information of any type (which is often unavailable in the population) and offers not just an assessment of baseline differences, but also the degree to which any propensity score based adjustment might improve the estimate of the population average treatment impact. We define this generalizability index so that it takes values in [0,1], with 1 indicating that the sample is "representative" of the population, and a 0 indicates that the sample and population are maximally different.

The index is useful both retrospectively – allowing samples from already completed experiments to be compared to many different potential inference populations – and prospectively— as a tool for evaluating a sample in relation to a specified inference population during recruitment. The fact that this index does not focus on a particular outcome measure or treatment effect estimator is useful, since many experiments collect multiple outcome measures and many treatment effect estimators are available. By focusing only on pre-treatment covariates, this index allows the role of differences between the sample and population (i.e. generalizability issues) to be isolated and separated from other design choices.

Significance / Novelty of study:
As the literature on experimental generalization develops, it is important for there to be statistical summary measures that allow for the quick comparison of the potential for generalization from a sample to a particular population. In order to be useful when an experiment is in the planning

stages or when multiple outcomes are collected in an experiment, it is ideal for such an index to be independent of the treatment impact measure. Additionally, it is ideal for this metric to be one that can be calculated easily and can be compared across many samples and populations. Ideally researchers could use this index when reporting the results of an experiment to give a sense to the limits and scope of generalization for the study.

Statistical, Measurement, or Econometric Model:

The goal of this paper is to develop a measure that summarizes differences in the propensity score distributions in the experimental sample and population. This means that in order for the index to be calculated, first a propensity score model must be specified and estimated. Based on the results of Stuart et al (2011) and Tipton (in press), the covariates used in this model should include those that are likely to explain variation in the effectiveness of a treatment or intervention across schools. After these variables are selected, a data frame is constructed which includes both units in the experimental sample and those in a particular inference population. This can be easily calculated using a logistic regression model. Based on this model, propensity scores (or their logits) can be estimated for each unit in the experimental sample and population. The generalization question that thus remains is to what degree these distributions are similar?

A natural first step is to do this visually by examining their histograms. For example, we might find two distributions similar to those in Figure 1. When creating a histogram, an important question is how the bins should be defined. The question of how to create the bins is very similar to the question of how to define the strata in a post-stratification or subclassification estimator. Applying results from Cochran (1968), we suggest that a good strategy is to create the bins in relation to the distribution of logits in the population, so that each of the *k* bins contains 1/*k*th of the population units.

More formally, our goal is to develop a measure that summarizes differences between these two histograms. We have defined the *k* histogram bins so that each bin *j* contains $N_j = N/k$ of the population units and n_j experimental sample units, where it is possible that some $n_j = 0$. Let $w_{pj} = 1/k$ and $w_{js} = n_j/n$ be the associated weights for each stratum j=1…k. The population distribution can therefore be summarized as the vector of weights $\mathbf{w}_p = (w_{p1}, w_{p2}, \ldots. w_{pk})$ and, similarly, the experimental sample distribution by the vector of weights $\mathbf{w}_s = (w_{s1}, w_{s2}, \ldots, w_{sk})$. Unless the two distributions are completely identical, unlike the w_{pj}, the values of the w_{sj} will vary from stratum to stratum. The statistical question at hand is how to compare the $\mathbf{w}_p$ and $\mathbf{w}_s$ vectors.

Definition: Generalizability index

Let the *generalizability index* B_k based on the covariates in **X** be defined as

$$B_k = \sum_{j=1}^{k} \sqrt{w_{sj} w_{pj}} == \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \sqrt{w_{sj}} \quad .$$

**

The generalizability index B_k is based upon the Bhattacharyya coefficient (1943, 1946), which is related to the Matusita measure of affinity (Matusita, 1967) and the Chernoff distance (Chernoff, 1952). Bhattacharyya first proposed this coefficient as a method for comparing two probability densities, and Rao (1949) noted that it could be used as an alternative to the Mahalanobis distance to compare two populations. It is widely used in pattern recognition for comparing or tracking images in video using histograms of color pixels (e.g. Comaniciu et al, 2000; Nummiaro, Koller-Meier, & Van Gool, 2003; Khalid, Ilyas Sarfaraz, & Ajaz, 2006), as well as in genomics for comparing the alleles in different populations (e.g. Chattopadhyay, Chattopadhyay & Rao, 2004; Shen et al, 2006; Zhang & Wang, 2009). There are many ideal properties of this index, which we will discuss in the next section.

This index has several important properties:
1) Takes values in [*0,1*]

2) Takes into account common support problems (*θ and τ*) *symmetrically*
Let θ is the proportion of the population with experimental sample units "like" them, while τ is the proportion of the experimental sample with population units "like" them. These two parameters highlight the two types of common support problems that can arise in generalization. Importantly, unlike other divergence measures (e.g. Pearson's chi-square test, Kullback-Liebler distance), this index can accommodate situations in which $\theta < 1$ and/ or $\tau < 1$ (Ullah, 1996). Additionally, it can be shown that when *k* is sufficiently large (i.e. when $\theta \geq 1/k$ and $\tau \geq 1/k$),

$$B_k = \theta\tau \sum_{i=1}^{k} \sqrt{w_{s_0 j} w_{p_0 j}} \leq \theta\tau$$

where $w_{s0j} = w_{sj}/\theta$ and $w_{p0j}=w_{pj}/\tau$ are the renormalized distributions on the common support region. This means that when either θ or τ are small – which are associated with large bias and increased sampling variance problems – the generalizability coefficient will be small too. The fact that θ and τ are weighted equally here is important, since an experimental sample is not *useful* for generalization if an estimator of the PATE based on it *either* exhibits large biases (when $\theta < 1$) or very large standard errors (when $\tau < 1$).

3) Does not require distributional assumptions
The fact that the B_k index can be calculated based on the empirical densities (i.e. histograms) of the logits is ideal, since it makes the calculations simple and the interpretations transparent.

4) Is invariant under monotonic functions
It is easy to show that the B_k index based upon the logits gives the same results as that based upon the propensity scores directly. While this is also true for the KS test, it is not true for the standardized mean difference balance test (which is proposed by Stuart et al (2011) as a method for assessing generalizability). This, too, makes the index more interpretable.

5) Special case: Normal distribution case
A central feature of the B_k index is that can be calculated based upon the empirical distributions and does not require any distributional assumptions to hold. Examining the special case in which the distributions of the propensity score logits are both normally distributed, however, offers insight into the relationship between this index and other measures of balance. If we assume that the distributions of the logits are $N(\mu_s,\sigma_s^2)$ in the experimental sample and $N(\mu_p,\sigma_p^2)$ in the population, it can be shown that

$$B = \exp\left(-\frac{1}{8}\left(\frac{\mu_s - \mu_p}{\sqrt{\left(\sigma_s^2 + \sigma_p^2\right)/2}}\right)^2\right)\sqrt{\frac{\sigma_s \sigma_p}{\left(\sigma_s^2 + \sigma_p^2\right)/2}}$$

where *B* does not have a subscript *k* since it is an exact, analytic result. This is a function of the standardized mean difference (*d*), commonly used as a measure of propensity score balance. However, unlike the propensity score balance measure, this function also takes into account differences in variances; when the variances of the propensity score logits differ, the effect is to reduce the size of *B*, even if the mean difference is small.

Usefulness / Applicability of Method:
In order to display the usefulness of the index, we include an example based upon a 2008-9 IES sponsored Goal III study of the middle school mathematics program SimCalc. SimCalc is a computer based program which teaches linearity, proportionality, and rates of change. The cluster randomized trial of this program was conducted on 73 schools in Texas. Previous work by Tipton (in press) evaluated how well the results of this study generalized to the state of Texas; this work gave an improved estimate of the population average treatment impact based upon propensity score subclassification.

In this paper, we return to the SimCalc experiment and look to see how well the sample of 73 schools used in the experiment represent the populations of non-charter non-magnet schools serving 7th graders in each of the 50 states. In order to do so, we use the publicly available Common Core of Data (CCD) from 2008-9. For this example we select the following covariates: total school size (MEMBER08); location of the school (urban, rural, suburban, town; ULOCAL08); Title I status of the school; the proportion of students that were white (WHITE08); on free lunch (FRELCH08); on reduced priced lunch (REDLCH08); were eligible for Title I services (STITLI08); or were male (MALE08). We chose these covariates since they included very little missing data and broadly included the types of covariates which we expect might impact the effectiveness of the SimCalc program on student achievement.

For each of the 50 states, we calculated a separate propensity score comparing the schools in the state to the schools in the SimCalc study. We then calculated the generalizability index for each of these states. The results can be found in Figure 2, which maps the degree to which generalizations are warranted in each state.

Conclusions:
The primary purpose of this paper is to develop an index that can be used to evaluate how generalizable an experimental sample is for a particular inference population. This index is based entirely on pre-treatment variables, enabling the index both to be calculated and recalculated throughout recruitment but before an experiment has commenced (as a sampling tool) and to be calculated after an experiment has been completed for many different inference populations and outcomes.

Appendix A. References

References are to be in APA version 6 format.

Bhattacharyya, A. (1943) On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society, 35*: 99 --109.

Bhattacharyya, A. (1946) On a measure of divergence between two multinomial populations. *Sankhya: The Indian Journal of Statistics*, *7*(4): 401-6.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4), 724-750.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., et al.. (2010). American Educational Research Journal. *American Educational Research Journal*. doi: 10.3102/0002831210367426.

Shadish, W.R., Galindo, R., Wong, V.C., Steiner, P.M., & Cook, T.D. (2011). A randomized experiment comparing random to cutoff-based assignment. *Psychological Methods*, 16(2), 179-191.

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, 369-386.

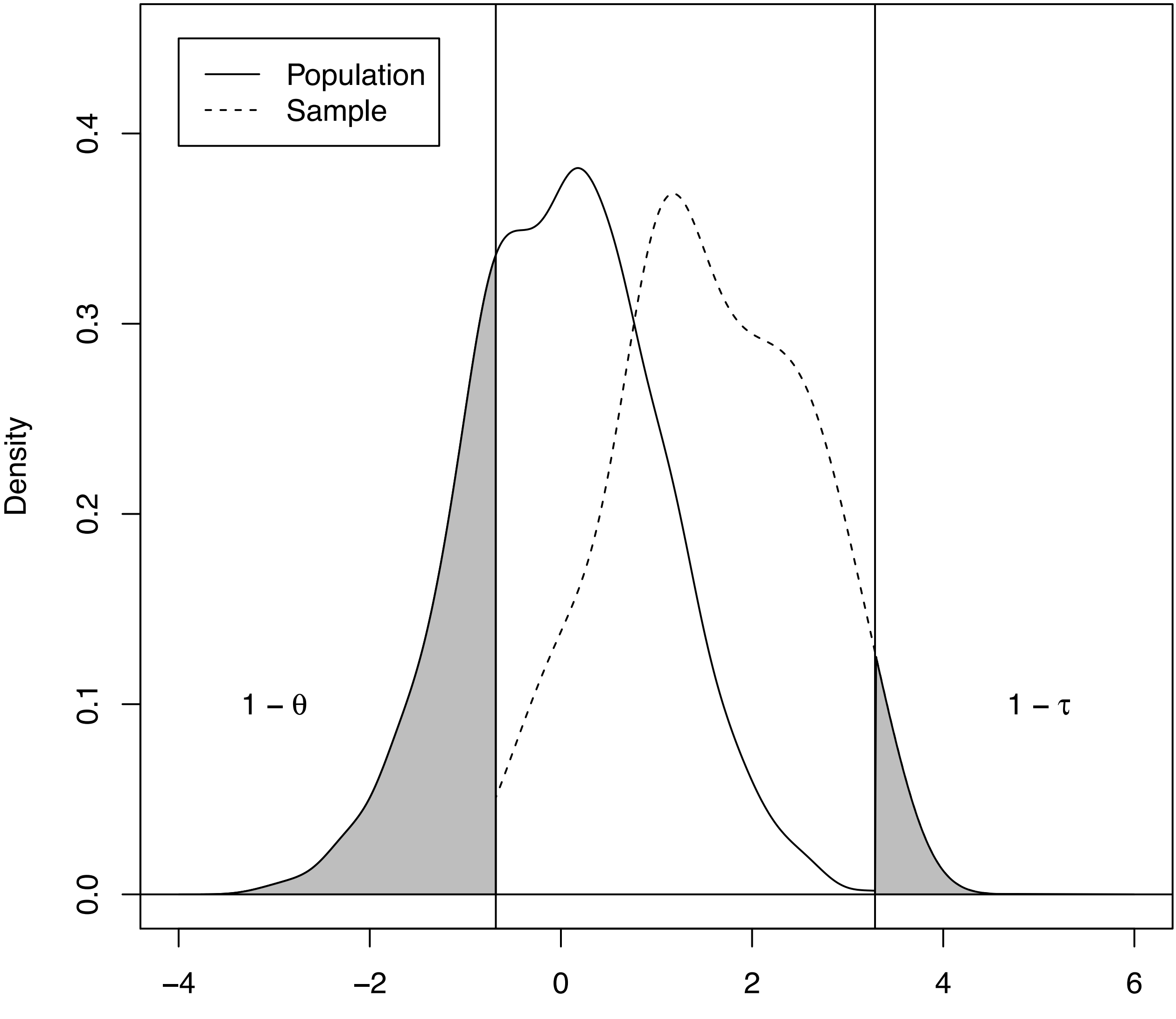
Stuart, E., Olsen, R.B., Bell,S.H., & Orr,LL. (2012) Estimates of External Validity Bias When Impact Evaluations Select Sites Purposively. Presented at the Spring 2012 Meeting of the Society for Research on Educational Effectiveness. Washington, D.C.

Tipton (in press) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*.

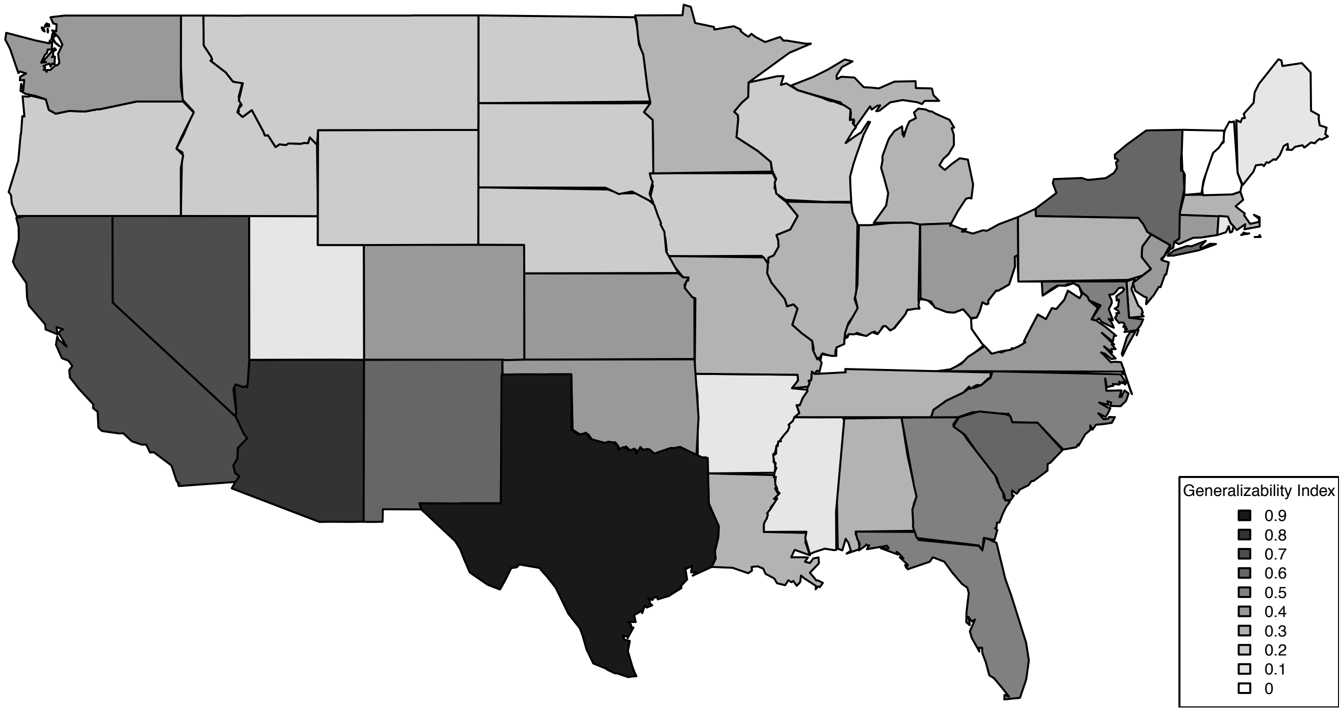
Tipton, E., Sullivan, K., Hedges, L.V., Vaden-Kiernan, M., Borman, G., & Caverly, S. (2011) Designing a sample selection plan to improve generalizations from two scale-up experiments. *Abstracts of papers, Fall Meeting of the Society for Research on Educational Effectiveness,* Washington, D.C.

Tipton, E. (2012) Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Working Paper. Teachers College, Columbia University*.

Figure 1: Three regions when comparing densities



Where does the Simcalc experiment generalize?



Note: Darker colors and larger indices indicate a higher degree of similarity between the experimental sample and population, where similarity is based on the following school level variables: # of students, proportion minority, proportion male, proportion free lunch, proportion reduced lunch, proportion Title I eligble, Title I status, location (4 categories).