Abstract Title Page

Title: Implications of small samples for generalization: Adjustments and rules of thumb

Authors and Affiliations:

Elizabeth Tipton*, Teachers College, Columbia University*
Kelly Hallberg*, American Institutes for Research*
Larry V. Hedges*, Northwestern University*
Wendy Chan*, Northwestern University*

Abstract Body

Background / Context:

Policy-makers are frequently interested in understanding how effective a particular intervention may be for a specific (and often broad) population. In many fields, particularly education and social welfare, the ideal form of these evaluations is a large-scale randomized experiment. The fact that sites or units within sites are randomly assigned to different interventions (or a control group) allows the causal impact of an intervention to be assessed without bias. However, recent research has highlighted that sites in these large-scale experiments are typically not randomly sampled from the population, making generalizations difficult (Olsen, Orr, Bell, & Stuart, 2013). For example, Stuart, Cole, Bradshaw and Leaf (2011) and Olsen et al. (2013) provide methods for *assessing* the similarity between samples and populations, while Hedges & O'Muircheartaigh (2011) and Tipton (2013) develop methods for *adjusting* for differences between the achieved sample and the population. All of these approaches extend the propensity score methodologies (Rosenbaum & Rubin, 1983; 1984) originally developed for observational studies (where causality is at issue) to the problem of generalization.

Purpose / Objective / Research Question / Focus of Study:

A problem not addressed by this literature is the effect of *small* sample sizes in generalization. For example, multi-site experiments can have as few as 10 – 15 sites[*], while "large-scale" cluster randomized experiments typically have fewer than 70 sites. In contrast, the inference population is typically much larger – often over 100 times larger. In this paper, we address three questions regarding the effect of these small sample sizes on: 1) assessments of generalizability; 2) rules of thumb for covariate balance; and 3) properties of estimators and estimation strategies. We investigate these issues in relation to sample sizes that vary from 30 to 70 clusters and on studies that are cluster-randomized or multi-site (random block) in design.

Significance / Novelty of study:

To date, the literature on generalization has not addressed the implications of small sample sizes on propensity score methods.

Statistical, Measurement, or Econometric Model:

Stuart et al (2011) and Tipton (2013) outline the assumptions necessary for generalization and situate these assumptions in relation to the propensity score literature. These methods require that a *population frame* can be developed that includes a list of all units (e.g. schools) in a well-defined population, as well as those units in the experiment. This frame also needs to include all covariates that explain variability in site-average treatment effects (the *sampling ignorability condition*). In order to make these comparisons, a *sampling propensity score* is estimated using a logistic regression model. If the sample selection is *strongly ignorable* (see Tipton, 2013) then an *unbiased* estimate of the population average treatment effect is possible. When it is not met, the goal is to provide an estimate with *less bias* than the sample average treatment effect typically calculated in experiments.

[*] In the final paper, we will also report results including studies using multi-site (randomized block) designs. In this abstract we describe the study designed to simulate cluster randomized trials randomizing schools to treatments.

Assessments of generalizability: RE-logistic

The sampling propensity score can be used to *assess* the similarity between a sample and population. Stuart et al (2011) argue that two statistics are often of interest, the absolute standardized mean difference (|SMD|) of the estimated propensity score logits and the difference in estimated propensity scores in the two groups. Importantly, these measures of assessment depend directly on estimates provided by a logistic regression model. In small enough samples, however, logistic regression is known to produce biased estimates of both the coefficients (β_i) and the associated probabilities. This *rare-events* problem arises when the number of 1's (here sample units) is small relative to the number of 0's (here population units). Various solutions have been proposed for reducing this bias, including the use of *rare-events* logistic regression (RE-), a profile-likelihood method proposed by King and Zeng (2001). In this paper, we compare results from RE- and standard- logistic regression to determine if and when these small sample corrections matter.

Assessments of generalizability: Balance

An additional concern, also of found in assessing generalizability is in the determination of rules of thumb. In the ideal, a sample would be a *miniature* of the population, though what counts as "miniature" is not clearly defined. In observational studies, rules of thumb for similarity (i.e., balance) have been proposed; the most common of these include either $|SMD| < 0.25$ or $|SMD| < 0.10$. Here, balance is typically assessed not only in terms of the |SMD| of the logits, but also in terms of the underlying covariates (i.e., $X_1, X_2, \ldots, X_p$), with the goal being to minimize the SMD for all of these. An important question, therefore, is if these rules of thumb are reasonable in generalization.

Post-hoc adjustments: Estimation methods

In many instances, researchers are not only interested in assessing generalizability but also in creating a better estimator of the average treatment effect. These methods use propensity score estimators for reweighting, including the post-stratification or subclassification estimator (Tipton, 2013; Hedges & O'Muircheartaigh, 2011),

$$T_{sub} = \Sigma w_{pj} T_j.$$

In this estimator, the distribution of estimated propensity scores in the population is divided into k equal sizes, each with $w_{pj} = 1/k^{th}$ of the population. Within each of these k strata, a separate treatment effect (T_j) is estimated, based on the n_j sample units that fall within the stratum. One question is to what degree post-stratification is useful in small samples and if better results could be gleaned through use of other methods, such as inverse-probability-weighting (IPW), defined as

$$T_{IPW} = \Sigma Y_{iT}/s(\mathbf{X}_i) - \Sigma Y_{iC}/s(\mathbf{X}_i).$$

This can be viewed as the limit of the post-stratification estimator, where each stratum contains only one school.

Research Design:

In this paper, we situate our investigation of small samples in generalization in relation to a particular example. The data we examine were drawn from a cluster randomized controlled trial (Konstantopoulos, Miller, & Van der Ploeg, 2013) that was designed to study the effect of Indiana's benchmark assessment system on student achievement in mathematics and English

Language Arts (ELA) base on annual Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) scores. Fifty-six K-8 schools volunteered to implement the system in the 2009-10 school year. Of these, 34 were randomly assigned to the state's benchmark assessment system while 22 served as controls.

Data from the experiment were supplemented by data on all of the other K-8 schools in the state of Indiana, which were used to define the inference population. The original Indiana dataset was truncated so that there were no charter schools nor any schools whose proportion of free reduced priced lunch, male, special education and limited English proficiency exceeded 95%. Schools whose enrollment were fewer than 100 students were also removed. This resulted in a population frame of 1514 schools. The locations of the 56 experimental and 1514 population schools are indicated in **Figure 1**. Initial analyses revealed several problems when using the subclassification or IPW approaches to generalize, including a limited number of strata, and SMDs outside the standard rules of thumb.

To examine these issues, using the population frame of 1514 schools, we conducted a simulation study to understand the relationship between sample size and propensity score estimation method (logistic vs. RE-logistic), degree of similarity/balance, and the effectiveness of various estimation strategies. To develop these adjustments and rules of thumb, we drew random samples of n (= 30, 50, 70) schools out of these 1514 schools. For each simulation, we included 1,000 iterations. We focused on random sampling since it is the ideal site selection method, in terms of both simplicity and bias.

In each repetition, after randomly selecting n schools, half of the schools were assigned to receive treatment and half were assigned to receive control. Next, a single propensity score model was estimated using both RE- and standard-logistic regression with fifteen covariates; these covariates were selected to achieve the ignorability condition and are listed in the first column of **Table 1**. For each iteration, we calculated the |SMD| between the sample and population for each of the 15 covariates, as well as the associated logits and RE-logits. For each of the sample sizes, across the 1,000 simulations, we calculated the value such that 95% of the |SMD|s were less than this value. This allowed us to answer our first two questions regarding assessment and rules of thumb.

Once these propensity scores were estimated, in each iteration, the logit and RE-logit of the propensity scores were used to stratify the population into three, four and five strata, and then post-stratification and IPW balance were assessed, using the T_{sub} estimator given above. For each covariate, the |SMD| was calculated.

Findings / Results:

Based on these simulation results (as well as analytic work included in the paper, but not here), we have three important findings with implications for practice.

Assessments of generalizability: RE-logistic

In all three sample sizes studied here, the |SMD| for the RE-logits were typically much smaller than those for the logits, and more importantly, were in line with the |SMD|s for the individual covariates. The differences were largest for n = 30, and decreased with sample size. This is an important finding: as a measure of assessment, in small samples the SMD of logits makes samples appear *less* similar to a population than they actually are. For example, with a sample of size n = 30, on average the |SMD| of the RE-logits is 0.14, versus 0.64 for the logits. These comparisons can be seen at the bottom of **Table 1**.

Assessments of generalizability: Balance

The second finding, illustrated in **Table 1**, is that the degree of imbalance between a sample and population is much larger under random-sampling than would be expected by the rules of thumb commonly in place in propensity score methods. For example, when $n = 30$, on average the |SMD| is roughly 0.15 and in 5% of samples, the |SMD| can be greater than 0.35. Even with samples as large as $n = 70$, on average the |SMD| is 0.10, with 5% of samples having values greater than 0.23. This means that the rule of thumbs for assessing similarity in generalization need to be *larger*, with |SMD| < 0.25 being the *most* stringent of these requirements, not the least.

Post-hoc adjustments: Estimation methods

In our previous experience, small sample sizes often limit the number of equal-populations strata possible in generalization. The results of this simulation study indicate that this problem is also likely to arise simply by change in random samples. In **Table 2**, for a sample size of $n = 30$ (the paper will include $n = 50, 70$) we indicate the proportions of samples (out of 1,000) in which there were enough schools in each of 3, 4, or 5 strata for estimation of a treatment impact in each stratum. We investigated this using RE-logit as well as logistic regression, and for both cluster-randomized designs (in which at least 1 treatment and 1 control site was required in each stratum) and multi-site designs (in which only 1 site is required per stratum).

Here there are two main results. First, using RE-logistic regression versus logistic regression greatly improves the ability to use post-stratification. For example, in a cluster-randomized design, under random sampling three strata would be possible for post-stratification in only 80.1% of samples using logistic regression, but 99.5% using RE-logit. Second, even using RE-logit, the use of five strata – the standard in the post-stratification literature – is not possible in over 30% of samples. This unfortunately suggests that in cases in which balance is not good, there are limitations to all post-hoc approaches.

In addition, we also investigate the performance of IPW relative to post-stratification, particularly in cases in which the maximum number of strata is small (e.g., 3, 4). (We include these results in the full paper, but do not discuss them here.)

Usefulness / Applicability of Method:

After the simulation study, we return to the Indiana example and apply our findings. In **Table 3** we first compare baseline differences between the sample and population, as well as remaining differences after post-hoc adjustments with different estimators. At baseline, if we use the standard |SMD| < 0.10 rule, only 5 out of the 15 covariates meet this assessment of similarity; however, if we use the |SMD| < 0.28 rule developed in Table 1, instead 10/15 of the covariates can be considered balanced. Using IPW increases this balance, as does sub-classification. In the paper, we further discuss similarities and differences between these results.

Conclusions:

Propensity score matching methods can be used to improve generalizability of findings from randomized experiments with non-probability samples, but adjustments and new rules of thumb are necessary in the application of these methods in this context.

Appendices

Appendix A. References

Hedges, L.V. & O'Muircheartaigh, C.A. (2011) Improving generalizations from designed experiments. Northwestern University. Manuscript submitted for publication.

King, G., & Zeng, L. (2001) Logistic regression in rare events data. *Political Analysis 9*: 137–163. Copy at http://j.mp/lBZoIi

Konstantopoulos, S., Miller, S., van der Ploeg, A. (2013). The Impact of Indiana's System of Interim Assessments on Mathematics and Reading Achievement. *Education, Evaluation and Policy Analysis.*

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

Rosenbaum, P.R. & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1): 41-55.

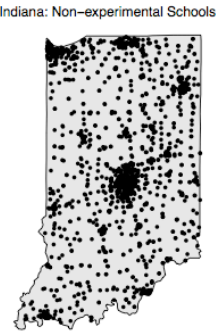
Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, Part 2, 369-386.

Tipton, E. (2013) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*: 239-266.

Appendix B. Tables and Figures

Figure 1: Map of Indiana experimental & population schools

(a) Experimental Schools



(b) Non-experimental Schools

Table 1: Empirical sampling distribution of |SMD| by sample size

	n=30		*n*=50		*n*=70	
Covariates	|SMD|	95% Rule	|SMD|	95% Rule	|SMD|	95% Rule
English Language Arts Test	0.15	0.36	0.11	0.27	0.10	0.24
Math Test	0.15	0.37	0.11	0.28	0.10	0.24
Attendance	0.14	0.35	0.11	0.28	0.09	0.23
FTE	0.15	0.37	0.11	0.28	0.10	0.24
Enrollment	0.14	0.36	0.11	0.29	0.10	0.24
Pupil Teacher Ratio	0.15	0.34	0.11	0.28	0.09	0.23
Couny Population	0.14	0.34	0.12	0.27	0.09	0.22
Title I (proportion)	0.15	0.38	0.11	0.28	0.09	0.24
Student Title I (proportion)	0.15	0.33	0.11	0.28	0.09	0.23
Male	0.14	0.36	0.11	0.27	0.09	0.22
White	0.15	0.36	0.12	0.26	0.09	0.22
Special Education	0.14	0.36	0.11	0.27	0.09	0.23
Free/Reduced Lunch	0.15	0.35	0.12	0.27	0.09	0.22
ELL/LEP	0.15	0.37	0.12	0.28	0.09	0.23
logits	0.64	0.86	0.5	0.66	0.43	0.57
RE-logits	0.14	0.34	**	**	**	**

*Note: All covariate values come from 2008, while the intervention occurred the following year. Importantly, the outcome in the experiment was the ELA test (first item in the list) in the following year. RE-logits result from a rare-events logistic regression model. Those items marked ** will be included in the final paper but are not yet available.*

Table 2: Number of equal-population strata possible in random samples (n=30)

	CRT (#T>0, #C>0)		RBD (# > 0)	
Number of Strata	Logits	RE-logits	Logits	RE-logits
3	0.801	0.995	0.997	1.000
4	0.498	0.896	0.958	1.000
5	0.214	0.697	0.874	0.997

Note: In the Cluster Randomized Trial (CRT) design, at least one treatment and one control school had to be available in each stratum, while in the Random Block Design (RBD) or multi-site trial, only one school needed to be in each stratum. RE-logits result from a rare-events logistic regression model

Table 3: Indiana example comparison of |SMD| for different estimators

Covariates	Baseline	IPW	3 Strata Sub Logits	3 Strata Sub RE-Logits
English Language Arts Test	**0.438**	0.249	0.174	0.155
Math Test	0.032	0.123	0.092	0.098
Attendance	0.068	0.279	0.249	0.253
FTE	**0.348**	0.23	0.16	0.167
Enrollment	0.271	0.173	0.172	0.168
Pupil Teacher Ratio	0.219	0.184	0.05	0.062
Couny Population	**0.439**	**0.374**	0.252	0.251
Title I (proportion)	0.146	0.219	0.216	0.29
Student Title I (proportion)	0.073	0.166	0.07	0.056
Male	0.013	0.225	**0.357**	**0.361**
White	**0.292**	0.175	0.19	0.209
Special Education	0.181	0.179	**0.381**	**0.347**
Free/Reduced Lunch	0.027	0.109	0.144	0.174
ELL/LEP	**0.32**	**0.459**	**0.424**	**0.463**
logits	**0.753**	--	0.156	--
RE-logits	**0.762**	--	--	0.173

Note: The experiment included 54 schools out of 1514 in the inference population. In both the logistic and RE-logistic regression subclassification, there were three strata (each with 1/3 of the population) with 38, 14, and 2 schools respectively (though the particular schools in each stratum differed). Bolded values are those greater than the 95% critical value (0.28), based on the simulations study.RE-logits result from a rare-events logistic regression model