Educ Psychol Rev (2013) 25:331–337
DOI 10.1007/s10648-013-9239-x

COMMENTARY

Recommendations for Practice: Justifying Claims of Generalizability

Larry V. Hedges

Published online: 15 August 2013
© Springer Science+Business Media New York 2013

Abstract Recommendations for practice are routinely included in articles that report educational research. Robinson et al. suggest that reports of primary research should not routinely do so. They argue that single primary research studies seldom have sufficient external validity to support claims about practice policy. In this article, I draw on recent statistical research that has formalized subjective notions about generalizability from experiments. I show that even rather large experiments often do not support generalizations to policy-relevant inference populations. This suggests that single primary studies are unlikely to be sufficiently generalizable to support recommendations for practice.

Keywords Replication · Practice policy · Generalization

The process of basic research, development of interventions, and evaluation of their effects in education involves distinctively different activities and skills. Each phase of this process is a serious intellectual endeavor, but the activities, settings that support them, and skills required to carry out the work are themselves quite different in each phase. Basic research (e.g., on learning) might well be carried out via laboratory studies. The development of interventions based on research findings would typically require small-scale studies in schools or other field settings. Evaluation of the impact of an intervention would typically require much larger-scale experiments or quasi-experiments conducted in field settings. Moreover, each phase requires personnel with specialized knowledge and skills. It would be unusual to find a researcher who is expert in all three of them and rarer still to find a single study that incorporated all of these aspects of the research, development, and evaluation process.

Robinson et al. (2013) propose a healthy separation between scientific research in education and prescriptions for practice (practice policies). I believe that this separation would reflect a healthy (and appropriate) respect for the distinct methods and skills required for each phase of research. For example, it is well known in the implementation research

This paper is based in part on work supported by the US National Science Foundation (NSF) under grants #0815295 and #1118978. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily represent the views of the NSF.

L. V. Hedges (✉)
Institute for Policy Research, Northwestern University, 2040 N. Sheridan Road, Evanston, IL 60208, USA
e-mail: l-hedges@northwestern.edu

community that not all interventions can be implemented well, and this limitation may doom an intervention whose impact would seem highly plausible given laboratory research. Moreover, it is useful to remember that in education as well as other fields such as labor research, medicine, and even business, most large-scale evaluations fail to confirm that promising interventions have the expected impacts (Coalition for Evidenced Based Policy, 2013). Therefore, it does not necessarily follow that small-scale research findings would be confirmed in larger-scale evaluation studies, let alone have the intended effects when implemented in practice.

The authors also make a broader point about the kinds of evidence that are required to provide a strong warrant for claims about generalizability. I am sympathetic to their claim that strong evidence is required to justify broad generalizability of causal effects. However, I would like to focus on one aspect of generalization, namely generalization of treatment effects from a study sample to a policy-relevant inference population. Generalization from a study sample to a policy-relevant population is somewhat less demanding than causal generalization of the form discussed by Robinson et al. (2013), but it is more tractable. Generalization in this more limited sense is a necessary condition for valid evidence-based policy recommendations, and emerging evidence supports the contention that few single studies meet even this limited sense of generalizability to policy-relevant populations.

Generalization based on probability sampling (which is sometimes called random sampling or representative sampling) is supported by a strong technical rationale. It can support unbiased estimates of population quantities and defensible estimates of their sampling uncertainties. Unfortunately, except for research based on secondary analyses of large-scale surveys, virtually no research in education uses probability samples. Instead, the research is conducted with convenience samples of schools and students. While some attempts may be made to assure "representativeness," probability sampling is not used, and the concept of representativeness is more a rhetorical claim than a concept with any technical meaning in statistics (see Kruskal and Mosteller 1979).

Moreover, I believe that probability sampling is actually ill suited to most education policy research for three reasons. First, there are often many targets of inference. A single research study may be intended to inform policy choices in a number of diverse settings, leading to several policy-relevant inference populations. Second, the inference populations are often not known in advance; they are discovered after the study has been conducted. Third, the study sample may not even be part of the inference population, because a study in one state is used to draw inferences in another state. For example, the Tennessee class size experiment was conducted to draw policy inferences in Tennessee (albeit without a probability sample of schools from the state). However, it has since been applied to draw inferences about class size effects in several other states, which was not planned in advance, and the study sample (drawn in Tennessee) was not part of the inference populations in other states.

Recent research has begun to formalize subjective notions about generalization in ways that shed light on the requirements for external validity and causal generalization when probability sampling has not been used (e.g., Hedges and O'Muircheartaigh 2010; O'Muircheartaigh and Hedges 2013; Stuart et al. 2011; Tipton 2013; Tipton and Hedges 2013). This work is all in the same spirit in that it formalizes generalization as a problem of estimating an average treatment effect (and its sampling uncertainty) in a well-defined inference population. Thus, generalizability is framed as a statistical estimation problem: How well can the population average treatment effect be estimated from a given study sample?

This makes it possible to apply well understood technical methods of statistical estimation to study the generalization problem. For example, because the properties of bias and variance (or standard error) are used to characterize the performance of statistical estimators, we can employ these concepts to characterize generalizability as the performance of

estimators of the population average treatment effect. Note that either large bias or large variance (or both) might undermine a generalizability claim. Moreover, framing generalizability as a statistical estimation problem exposes the fact that more than one estimator might be derived from a single study sample and there is a problem of selecting the best or at least a good estimator. For example, the conventional estimate of treatment effect is typically not the most generalizable estimate of the population average treatment effect.

This approach starts with the assumption that it makes no sense to say that a treatment effect is generalizable without defining the inference population to which generalizations are supposed to apply. In other words, to say a research result is generalizable makes no sense unless you say generalizable *to what*. Usually the target of generalization is the average treatment effect in some inference population of interest. By referring to the average treatment effect, we acknowledge that the treatment effect may not be identical for every unit in the population, a point to which we will return later.

While this framework is straightforward, it leads to several immediate implications. For example, if a generalizability claim must be accompanied with a specific target of generalization, then it follows that the results of a study might be quite generalizable to some policy-relevant inference populations, less so to others and not at all to yet others. The factor limiting generalizability might be bias of the estimate in the population average treatment effect, the variance of that estimate, or both. To evaluate generalizability claims, it is essential to have some evidence about both bias and variance (or a functionally related quantity such as the standard error of the estimate).

If the study sample was a simple random probability sample from the inference population, generalization (that is estimation of the population average treatment effect) would be reasonably straightforward. However, it is more likely that the probability sample would be a multistage cluster sample (e.g., obtained by first sampling schools, then students within schools). It is worth noting that when two-stage probability sampling is used, the conventional analysis of a balanced cluster randomized or randomized block experimental design does not yield an unbiased estimate of the population average treatment effect when cluster (e.g., school) sizes are not identical in the population (see Hedges and O'Muircheartaigh 2010). An unbiased estimate can easily be computed by weighting, but this example illustrates that even under ideal conditions (e.g., probability sampling), more than one estimator (of the population average treatment effect) is possible and attention to issues of generalization is necessary in the analysis.

If treatment effects do not vary in the inference population (and if the study sample is a subset of it), then an internally valid study's estimate of the treatment effect will be an unbiased estimate of the population average treatment effect, and generalizability is straightforward. Unfortunately, surprisingly little is known about the variation in treatment effects. Most of what we know comes from the handful of interactions that happen to have been estimated (usually with designs that are underpowered to test interactions, because they are not the primary focus of the study). Thus, we do not have convincing bodies of evidence about the nature of variation in treatment effects and their correlates.

This is important because if treatment effects do vary, then this variation has implications for the estimation of the population average treatment effect. All of the methods that have been proposed to make population inferences from nonprobability samples require that there is a set of covariates that are measured on both the units in the study sample and the units in the inference population (from either a census or a probability sample) that explain the variation in the treatment effects. This requirement has been called coherence (e.g., by Stuart et al. 2011) or ignorability of sampling of treatment effects given the covariates (e.g., by Hedges and O'Muircheartaigh 2010). Note that this requirement is parallel to those used in estimation with

missing data (see, e.g., Little and Rubin 2002) or in using covariates to reduce bias in observational studies and quasi-experiments (see, e.g., Rosenbaum and Rubin 1983).

In education experiments, the units of analysis are often schools, so census data can be obtained from collections such as the National Center for Education Statistics (NCES) common core of data, US Census data, or from state data systems which often have quite extensive public-use data on schools. The surveys conducted by NCES provide even richer data on national probability samples, and if these data have been collected within study samples, they can provide a rich covariate set with which to explain plausible variation in treatment effects.

The strategy used by these methods is to match (explicitly or implicitly) the study sample to the inference population on a set of covariates that make the sampling of treatment effects ignorable. Usually a large set of covariates will be used to increase the likelihood that the ignorability assumption will be met, but they will be incorporated into a single variable that summarizes the information in all the covariates (e.g., a propensity score reflecting propensity of units with a given ensemble of covariate values to be in the study sample) (see, e.g., Hedges and O'Muircheartaigh 2010).

Once a covariate set has been summarized, estimates of population average treatment effects can be obtained in several ways. The simplest is to stratify the distribution of the covariate (e.g., the propensity score). Five to seven strata have been shown to be adequate to control at least 90 % of the bias arising from variation in the covariate (see Cochran 1968 or Rosenbaum and Rubin 1983). Then the portion of the study sample in each stratum is used to estimate the treatment effect within that stratum. The estimate of the population average treatment effect is the weighted average of the treatment effects in each stratum, and its variance is a simple function of the weights and the variances of the treatment effects in each stratum. If the strata are chosen to have equal proportions of the population distribution (e.g., if five strata are divided at the quintiles of the distribution of the covariate), then the weights will be equal (e.g., 1/5 for five strata defined by population quintiles), the estimate of the population average treatment effect will be the simple (unweighted) average of the treatment effects in each stratum, and the variance will be the simple average of the variances divided by the number of strata.

Note that in this methodology, the treatment effect depends on the covariate. The function of the study sample is to provide estimates of the treatment effect for each stratum of the covariate distribution. If the ignorability assumption is met, the treatment effect is the same for units that have the same covariate value, whether they are in the inference population or not. Therefore, it is irrelevant whether the study sample is a subset of the inference population. The method estimates the average treatment effect in a population that has the same composition (on the covariate) as the inference population. A similar approach is used in demography where it is called standardization (that is, standardizing a population composition so that comparisons between desired quantities in two different populations are not confounded with difference in population composition) (see, e.g., Kitagawa 1964). It is also used in economics to form index numbers and to decompose effects to isolate the impact of changes in population composition (see, e.g., Oaxaca 1973). Finally, it is used in survey research (see, e.g., Rosenberg 1962 or Kalton 1968), and it is a basic tool in the analysis of missing data (see Little and Rubin 2002).

If the study sample were a probability sample of the inference population, the proportion of the study sample in each stratum would be the same as the proportion of the inference population in that stratum. In a nonprobability sample, the proportion of the study sample in a stratum may be quite different from the proportion of the inference population in that stratum. The degree to which the proportions match in each stratum is a reflection of how representative the covariate distribution (and therefore the treatment effect distribution) in the study sample is

of the inference population. The degree to which the proportions are different (that is, the study sample and the inference population do not match) inflates the variance of the estimate of the population average treatment effect.

If it is possible to obtain an estimate of the treatment effect in each stratum (and if the ignorability conditions are met), then the estimate of the population average treatment effect will have reasonably small bias. If the number of study sample units within a stratum is small, the variance of the treatment effect estimate in that stratum will be large, and the variance of the combined estimate will also be large. If there are strata in which treatment effects cannot be estimated (e.g., strata that do not include some units assigned to the treatment group and some units assigned to the control group in the study sample), then the study sample simply does not support any estimate of the average treatment effect in the inference population. Another way to characterize the latter situation is to say that the variance of the estimate of the population average treatment effect is infinite.

The problem of strata in which treatment effects cannot be estimated is analogous to undercoverage in sample surveys and has the same consequences. One way this can be handled is to change the definition of the inference population, which is equivalent to acknowledging that the study is not able to provide inferences to the desired population, but can to some related populations. For some policy purposes, this may be acceptable if the newly defined inference population can be defined in a meaningful way (e.g., all of a state except the five most rural counties). Another approach is to impute treatment effects to the strata where they are missing. Of course, this approach involves substituting assumptions for data, but it too may be acceptable in some situations. However, neither of these responses is what is typically done. Rather, the conventional treatment effect estimate (and its standard error) is taken at face value as applying to the inference population.

Applications of these methods to a few large-scale experiments in education are instructive and provide some empirical evidence about generalizability. One application is to the Tennessee class size experiment (Hedges and O'Muircheartaigh 2010). The Tennessee class size experiment has been called "one of the great education experiments of the century" by Mosteller, Light, and Sachs (1996). The experiment was conducted to assess the effects of substantial class size reduction (from an average of 22 to 15) on academic achievement of elementary school children (see Nye et al. 2000). Initially, every school district in the state of Tennessee was invited to participate in the experiment. In order to participate, schools had to agree to (1) the random assignment of both teachers and students to classes, (2) keep each student's class size assignment the same for the 4 years of the experiment, and (3) give the research team access to the students in order to verify actual class sizes and administer tests. A total of 79 elementary schools throughout the state eventually participated in the experiment. Several analyses of the data suggested interactions that implied that treatment effects were heterogeneous. The set of covariates measured in the experiment was limited (we used a set of ten covariates), and they (or reasonable proxies of them) were measured in the US census so that the values for various potential inference populations could be constructed using the 1 % sample of the 1990 census microdata (available at IPUMS). The inference populations (we considered several) were subsets of the national population of children who were 9 to 10 years old in April 1990, the same age as the children in the Tennessee class size experiment. The data permitted us to develop estimates of the average treatment effect for children in the state of Tennessee. The estimate for the state of Tennessee was almost identical to that given by the conventional analysis of the experiment (it differed by less than 1 %), but the standard error was only about 15 % larger, which suggests that the results might generalize fairly well to the state of Tennessee. However, generalizability was considerably worse for other potential policy-relevant inference populations. For example, the point estimate for the population average

treatment effect in Los Angeles was about half as large as the conventional estimate, and the standard error was six times as large so that the standard error was twice the size of the point estimate in Los Angeles. We interpreted these results as indicating that the results generalized poorly to Los Angeles.

We agree with Robinson et al. (2013) that a series of replicated studies would generally provide a better basis for generalization. The methods described above can be applied to collections of replicated studies. These methods were applied to the data from three experiments involving a total of 90 middle schools in Texas (see O'Muircheartaigh and Hedges 2013). In this case, the treatment was a software-based intervention that uses dynamic representations to help students learn how to solve rates and proportions problems (Roschelle et al. 2010). The initial inference population was the set of 1,713 public middle schools in Texas with seventh grade classrooms in 2009 that were not charter schools. The covariate set was slightly larger involving 26 covariates chosen from the state's academic excellence indicators system, which provides data on all public schools in Texas. In spite of relatively wide dispersal of the study sample throughout the state, it was not possible to generalize the results of the experiment with any degree of confidence to the entire state. It was, however, possible to characterize the quality of the generalization (that is, estimate the mean and variance of the average treatment effect) for different school districts in the state.

These examples illustrate that even large single experiments may not be able to provide generalizable estimates to the populations that are obvious targets of inference. It also illustrates that even if a study is generalizable to some inference populations, it may not be very generalizable to others. One might criticize these particular empirical demonstrations because the covariate sets they chose are not large enough to guarantee that the crucial ignorability condition is met. However, if anything, these examples illustrate how difficult it is to support generalizability from a single study. A more comprehensive list of covariates would likely make population coverage even more difficult and generalizability less likely. In other words, these examples are likely to have overestimated generalizability, not underestimated it.

These empirical demonstrations suggest that we should be skeptical of any single study claiming to provide evidence that is generalizable enough to provide a basis for policy and practice recommendations. Such a study should be required to present substantial evidence to support the claim because even large and well-designed studies may fail that test. It would appear that collections of studies assembled in research syntheses would seem to be more likely to provide evidence that could support broad generalizations. However, the example of the studies in Texas shows that even collections of replicated studies may not be generalizable to the obvious policy-relevant populations.

References

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295–313.

Hedges, L. V., & O'Muircheartaigh, C. (2010). Improving inference for population level treatment effects in social experiments. Working paper, Northwestern University Institute for Policy Research.

Kalton, G. (1968). Standardization: a technique to control for extraneous variables. *Applied Statistics, 17*, 118–136.

Kitagawa, E. M. (1964). Standardized comparisons in population research. *Demography, 1*, 296–315.

Kruskal, W., & Mosteller, F. (1979). Representative sampling III: the current statistical literature. *International Statistical Review, 47*, 245–265.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: lessons learned from skill grouping and class size. *Harvard Educational Review, 66*, 797–842.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on achievement: the results of the Tennessee class size experiment. *American Educational Research Journal, 37*, 123–151.

O'Muircheartaigh, C., & Hedges, L. V. (2013). Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society*, *Series C* (in press)

Oaxaca, R. (1973). Male–female wage differentials in urban labor markets. *International Economic Review, 14*, 693–709.

Robinson, D. H., Levin, J. R., Schraw, G., Patal, E. A., & Hunt, E. B. (2013). On going (way) beyond one's data: a proposal to restrict recommendations for practice in primary educational research journals. *Educational Psychology Review* (in press).

Roschelle, J., Shechtman, N., Tatar, D., & Hegedus, S. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics. *American Educational Research Journal, 47*, 833–878.

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

Rosenberg, P. (1962). Test factor standardization as a method of interpretation. *Social Forces, 41*, 53–61.

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*. (in press)

Tipton, E., & Hedges, L. V. (2013). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness* (in press)