Building Disciplinary Infrastructure for Generalization in the Social Sciences

Rachel Strohm[1]
University of California, Berkeley

6 August 2019

Abstract

Generalizability is widely agreed to be a desirable characteristic of social science research. Many discussions of the topic present it as a tradeoff between a study's internal validity, and its generalizability, which is best achieved by increasing its sample size. At present, individual researchers usually bear all the costs of expanding the sample size, which means that generalizable single studies are undersupplied. I argue that disciplines should subsidize and coordinate generalizable research by building infrastructure for systemic reviews and coordinated multi-site studies. Both of these techniques expand sample sizes by aggregating data across studies, which lowers the cost to individual researchers. The biomedical sciences provide a model of infrastructure for generalization within a mature research ecosystem. The social sciences have been slower to build such infrastructure, although it has been expanding more rapidly in the last decade. The substantive implication of this argument is that researchers should focus on their preferred type of internally valid research, and disciplines as a whole should take responsibility for assessing the generalizability of research findings.

[1] PhD candidate, political science (rstrohm@berkeley.edu). With thanks to Leonardo Arriola, Melissa Griffith, and Tom Wein for their helpful comments.

1. Introduction

The social world is complex, and attempts to find regular social "laws" often fall short. Of course, this difficulty is precisely why many social scientists highly value the endeavor. This is captured by the fact that generalizability (or alternatively "generality") is listed as a goal of social science research in all introductory methods texts. However, compared to many of the other primary desiderata of social science, generalizability is relatively undertheorized. Take Gerring's discussion of the principles of social science: "inquiry of a scientific nature, I stipulated, aims to be cumulative, evidence-based (empirical), falsifiable, *generalizing*, nonsubjective, replicable, rigorous, skeptical, systematic, transparent, and grounded in rational argument" (2012, p. 11).

Many of these other principles have generated intense debate, and led to shifts in the ways that scholars conduct their research and share their data. For example, over the last half-century the fields of political science and economics in particular have shifted from featuring largely descriptive work to prioritizing causal inference and the testing of falsifiable hypotheses (Shadish, Cook & Campbell 2001). In psychology, the replication crisis of the 2010s has led to major changes in the way that quantitative research is handled, including pre-registration of studies, controls against p-hacking, and a push to make data freely available for replication purposes (Shrout & Rodgers 2018). The same concerns have led to discussions of transparency in the social sciences more broadly, producing initiatives such as the Data Access and Research Transparency statement (DA-RT),[2] and the Berkeley Initiative for

[2] https://www.dartstatement.org. Accessed 15 January 2019.

Transparency in the Social Sciences (BITSS).[3]

By contrast, the discussion around how generalizability should be operationalized is much briefer. The standard recommendation remains King, Keohane & Verba's (1994) exhortation to simply expand the study's sample size, despite critiques about the feasibility of this approach by a range of qualitatively oriented scholars (Brady & Collier 2004). Gerring summarizes the situation well: "With respect to the scope of an argument bigger is better – but evidently only to a point. Indeed, as an argument is extended its veracity, precision, or coherence often declines. This is the point at which the criterion of generality begins to conflict with other scientific criteria" (2012, p. 64). In other words, individual researchers are expected to bear the costs of expanding their samples, and must do so with little clear guidance on how to balance the tradeoffs between greater example size and other goals of their research, such as in-depth case knowledge. This means that generalizable single studies are relatively undersupplied.

In this paper, I argue that the social sciences can increase the supply of generalizable research findings by building discipline-level infrastructure for aggregating data across studies. This includes journals, organizations which implement systematic reviews, and research initiatives which coordinate multi-site studies. Both systematic reviews and coordinated multi-site studies expand sample sizes by aggregating data across studies, which lowers the cost to individual researchers. The biomedical sciences provide a model of infrastructure for generalization within a mature research ecosystem. The social sciences have been slower to build such infrastructure, although it has been expanding more rapidly in

[3] https://www.bitss.org. Accessed 15 January 2019.

the last decade. The substantive implication of this argument is that researchers should focus on their preferred type of internally valid research, and disciplines as a whole should take responsibility for assessing the generalizability of research findings.

The paper proceeds as follows. Section 2 discusses the assumptions needed to make claims about a study's generalizability. Section 3 discusses the limitations of expanding sample sizes within individual studies. Section 4 summarizes the biomedical research cycle, which has a fully developed system of disciplinary organizations that assess the generalizability of research findings. Section 5 discusses the state of similar generalizability infrastructure in the social sciences, which is less well-developed but showing exciting growth. Section 6 concludes.

2. Generalizing from Assumptions

Generalizability is about making an inference from a sample to a population. As Pepinsky (2018) points out, there are only two ways to get from the sample to the population: assumptions or empirics. In other words, one can make a theoretically-informed argument about why one believes that the data from the sample accurately represents the population, or one can collect additional data to get more empirical evidence about the population.

It's important to first define the population to which the generalization is expected to apply. If the population is narrowly defined ("white, middle-class American adults from 1980 - 2000"), this increases the accuracy of generalization but reduces its scope. For a broadly defined population ("all states which existed between 3000 BCE and the present"), generalizations are typically less detailed but more widely applicable. In this paper, I assume that social scientists typically wish to

find social laws that apply across geographic contexts and across time. In other words, I am speaking about generalization to a broad population. This comes from my epistemic position as a comparative political scientist, as our whole *raison d'être* is doing cross-national work. For readers whose work is more narrowly focused, I believe this paper will still be useful, but some of the data constraints I address below may not be so binding.

In statistical terms, a large, random, unbiased sample from a defined population should produce a precise and unbiased estimate of the population parameters. If we can confirm that a sample meets these criteria, it is reasonable assume that it generalizes to the population.[4] Of course, for many research questions, it may be difficult to produce this type of sample. There are three main challenges here. I'll discuss some representative examples from various social sciences, without meaning to imply that the challenges are specific to the field discussed – they are all issues found widely within the social sciences.

First, there are constraints on the types of data available to social scientists. In terms of random selection and representativeness, a great deal of psychology research involves samples selected from Western, educated, industrialized, rich, and democratic (WEIRD) countries, which may not generalize very well elsewhere in the world (Heinrich, Heine & Norenzayan 2010). In terms of bias, historians and other historically-oriented social scientists are more likely to rely on elite written records than popular oral histories.

Second, the precision of even a large, unbiased random sample may be

[4] Although see Aronow & Samii (2016) for limitations on the accuracy of representative samples in multiple regression analysis.

reduced by the fact that there is a great deal of unpredictable variation in many outcomes in the social world. For example, the economists Rosenzweig & Udry (2019) note that rainfall shocks cause massive variation in annual agricultural yields in Ghana and India, and that macroeconomic shocks which affect urban wages lead to significant variation in the annual returns to schooling in Indonesia. This implies that even an accurate sample drawn from one of these populations in any given year is likely to be quite far from the long-run population average – or, in other words, that the estimate is quite imprecise.

Third, the nature of time itself may limit our ability to make durable statements about the social world. This is because many outcomes of interest to social scientists are path dependent. As Pierson notes, "Specific patterns of timing and sequence matter; … large consequences may result from relatively small or contingent events; [and] particular courses of action, once introduced, can be almost impossible to reverse" (2000, p. 251). If one takes a strict view that a generalizable statement should be true of a population at any point in time, the existence of path dependence means that such generalization is likely not possible. A more clement view is that the scope conditions of generalizing statements need to include temporal scope as well as geographic or institutional.

Thus, the conditions necessary to assume that a result can be generalized from a sample to the population are frequently not met, particularly if the intended generalization is a broad one. We are then left with Pepinsky's (2018) other pillar of generalizability: empirics. The best way to know if some outcome of interest to social scientists can transfer from one context to another is to study the second context.

3. Generalizing from Empirics

We are thus back to King, Keohane & Verba's (1994) advice to expand one's sample. In general, individual researchers are expected to bear the costs of expanding their samples – and must do so with little clear guidance on how to balance the tradeoffs between additional data collection and other goals of their research.

There are three main reasons why expanding the sample size of a single study may not always be feasible or desirable. First, data collection is costly in terms of time, money, and developing familiarity with new research contexts. These constraints are particularly acute for scholars who study rare or historical events, or who work in locations where off-the-shelf data on the topic of interest is not easily available (c.f. Skocpol 1979 on the study of social revolutions). This poses significant challenges for scholars who must publish regularly in order to succeed on the academic job market and get tenure.

Second, there are real tradeoffs between the amount of detail one can gather about a case and the number of cases one can sample. Using process tracing to explore relationships of causality and contingency within a small number of cases has produced some classic works of political science, such as Tannenwald's work on normative stigma attached to the use of nuclear weapons (1999) or Scott's work on peasant resistance to state formation in Southeast Asia (1985, 1998, 2009). Our understanding of the world would be poorly served by advising researchers to eschew these time-intensive, small-N methods entirely. However, this does limit the number of cases one can include in one's study.

Third, when researchers are individually responsible for finding low cost

ways to maximize their sample size, they may choose to study well-developed topics with more data available. This runs the risk of discouraging theoretical innovation, and can contribute to academia's disproportionate focus on high income countries[5], paradoxically limiting their work's generalizability across geographic contexts.

When researchers must bear high costs to produce socially beneficial outcomes, such as generalizable research, it is likely to be undersupplied. There is a clear need for discipline-level interventions to subsidize generalizability.

4. Generalization Infrastructure in the Biomedical Sciences

The biomedical sciences provide a useful example of disciplinary infrastructure designed to assess the generalizability of research findings. In general, the fundamental process of research in these fields is quite similar to the social sciences. Research cycles begin with the observation of a novel or puzzling phenomenon, followed by the collection of observational data, and then (frequently) experimental research. However, within these research cycles, the roles of individual researchers are structured quite differently (Lieberman 2016) – as are their approaches to generalizability.

Lieberman (2016) makes the insightful point that the biomedical sciences typically publish studies which are disaggregated by step within the research cycle. As he notes, a single issue of the field-leading *New England Journal of Medicine* from 30 July 2015 contained articles on a single clinical observation of a childhood inflammatory illness, two descriptive articles about the incidence of cancer and pneumonia in American cities, two laboratory studies of promising new drugs, and a

[5] Das et al. (2013) note that from 1985 – 2004, the top five American economics journals published 2383 papers on the US, 65 papers on China, and 34 papers on all 54 countries in Africa.

large-scale randomized controlled trial (RCT) on improving organ donation outcomes (pp. 1054 – 1055). Conversely, in political science (and often economics as well), "only the last set of studies—those that test causal relationships … —are the ones that get published in top [journals]. Many of the other steps are described in a cursory manner and barely find their way into the appendices of published work" (p. 1057). The biomedical approach thus makes the knowledge production process explicit, and allows researchers to specialize in and be professionally rewarded for working at different stages along the research cycle.

I observe that this differentiation also extends to the process of generalization. Rather than pushing individual researchers to continually expand their sample sizes, the biomedical sciences have discipline-level infrastructure in place to carry out the process of aggregating data across studies and assessing the generalizability of their results. The most prominent organization in this space is Cochrane, which produces systematic reviews on healthcare practice and policy. Established in 1994, it is supported by prominent health funders such as the National Institutes of Health (US) and National Institute for Health Research (UK), with a 2017 budget of over US$10 million.[6] Cochrane manages a network of 11,000 volunteers who have produced, edited, and translated over 7000 systematic reviews.[7] With a large, well-funded organization dedicated to assessing the generalizability of research results, biomedical researchers are free to focus on their preferred type of internally valid research, further supporting the role differentiation within their field.

[6] https://www.cochrane.org/about-us/our-funders-and-partners. Accessed 6 August 2019.
[7] https://www.cochrane.org/join-cochrane/cochrane-membership-thresholds. Accessed 6 August 2019.

5. Generalization Infrastructure in the Social Sciences

The social sciences have been slower to develop similar discipline-level infrastructure for generalization. However, in recent years there has been rapid growth in organizations carrying out systematic reviews, and also conceptual innovations in the form of coordinated multi-site studies. Both have important roles to play in assessing the generalizability of existing research findings, and designing new research with the intention of greater generalizability.

5.1. Systematic Reviews

Systematic reviews are the most common way of approaching generalizability. I use this phrase to refer to any study which systematically collects all of the internally valid studies on a topic, and then assesses their results. However, systematic reviews may differ in terms of what they consider internal validity, and how they aggregate and analyze data. Many systematic reviews only include studies with experimental or quasi-experimental research designs, since these are seen as having the highest internal validity. Quantitative data can be pooled for meta-analysis (Higgins & Green 2011). If a study includes qualitative data, researchers may carry out a metasummary, in which the data is extracted and recoded in a manner analogous to pooling quantitative data (Sandelowski, Barroso & Voils 2007). These strategies have the advantage of rigor, but can also exclude topics which aren't amenable to being experimentally studied. Narrative reviews tend to include a wider range of studies, and summarize their findings without reanalyzing the underlying data (Green, Johnson & Adams 2006). Evidence gap maps can be used to assess where data is weak or absent on a particular topic, which is a useful way of incorporating missing data into the systematic review process (Snilstveit et al. 2017) –

as well as a convenient shortcut for researchers looking for new topics.

The clear strength of systematic reviews is their ability to comprehensively summarize the literature on a topic, draw out generalizable findings, and highlight areas for future research. Because they draw on existing work and do not require novel data collection, they can be done relatively rapidly and inexpensively. Depending on the inclusion criteria for a particular review, they can cover a wide range of topics using both experimental and non-experimental methods, and qualitative and quantitative data. The weaknesses of systematic reviews are precisely the inverse – they are limited to topics which have already been studied. If the evidence base on a topic is weak, by reason of either limited studies on the topic or poor research design within those studies, they are unable to expand it. Furthermore, the potential for meta-analysis and metasummary is limited by social scientists' willingness to share their data.

There are a number of existing systematic review initiatives within the social sciences. These are summarized in table 1. These organizations differ from the biomedical infrastructure for generalizability in several ways. They are on average founded more recently than Cochrane, and their combined budgets and total research output are far short of Cochrane's. The landscape is more fragmented, with multiple organizations performing similar types of reviews, particularly on international development. Finally, only the Annual Review journals map neatly onto disciplinary boundaries, while the other organizations are instead arranged around thematic focuses. The Annual Reviews also list their reviews by year of publication (in journal format) rather than in a database with topic menus, which makes it more difficult to fully capture the range of topics their reviews cover.

Table 1: Systematic Review Initiatives in the Social Sciences

Organization	Description	Date Founded	Annual 2017 Budget (USD)	Reviews
AidGrade[8]	A small non-profit doing meta-analyses of interventions funded by foreign aid	2012	$10,000 (est.)[9]	10
Annual Reviews[10]	Journals publishing narrative reviews in a variety of social science fields	1950 - 2015	Unknown	4000 (est.)[11]
Campbell Collaboration[12]	Research center producing systematic reviews on crime and justice, disability, education, international development, nutrition and food security, and social welfare	2000	$1.7 million	186
International Initiative for Impact Evaluation[13]	Research center producing impact evaluations and systematic reviews on international development	2008	$6.3 million	600

This less mature generalizability ecosystem may be both cause and consequence of the fact that social scientists engage less often with systematic reviews than their colleagues in biomedical sciences. Ozier (2019) finds that American medical journals mention the terms "meta-analysis" or "systematic review" fully 30 times more often than do economics journals, with other social

[8] http://www.aidgrade.org. Accessed 6 August 2019.
[9] https://www.kickstarter.com/projects/972584134/what-works-in-development-10-meta-analyses-of-aid. Accessed 6 August 2019.
[10] https://www.annualreviews.org. Accessed 6 August 2019.
[11] The social science journals were founded in the following years: 1950 (psychology), 1972 (anthropology), 1975 (sociology), 1998 (political science), 2005 (clinical psychology), 2009 (economics), 2015 (linguistics). The most recent editions of the political science, linguistics and psychology journals all contained 20 or more articles. Assuming 20 articles per year for each year of publication, this produces an estimate of 4180 articles in the social sciences.
[12] https://campbellcollaboration.org/. Accessed 6 August 2019.
[13] https://www.3ieimpact.org. Accessed 6 August 2019.

sciences falling somewhere in between. Anecdotally, an informal poll of 163 social scientists on Twitter in May 2019 suggested that only 34% regularly seek out systematic reviews, although 45% report that they do find the reviews useful if they happen to come across them.[14]

5.2. Coordinated Multi-Site Studies

The second way of approaching generalizable research in the social sciences is by carrying out coordinated multi-site studies on specific topics. This strategy involves implementing research projects simultaneously in multiple different settings. To date, most of the initiatives working on this type of study have focused on experimental research, although there is no inherent reason why the same design couldn't be applied for non-experimental work. The advantage of this method is that it generates new evidence, and does so in a way which by design produces comparable estimates of effects across country contexts. The disadvantage is that it's often quite expensive and logistically challenging, and isn't feasible for studying historical or rare events.

Coordinated multi-site studies have been taking off within political science, economics, and psychology. In political science, the Evidence in Governance and Politics initiative was founded in 2009, and is currently coordinating five parallel studies across up to 17 countries through its Metaketa initiative.[15] Study topics include information interventions, taxation, natural resource governance, community policing, and political participation in hybrid regimes. In economics, Banerjee et al.

[14] I polled the academics among my 13,000 Twitter followers on 1 May 2019 about their use of systematic reviews. Most of the academics who follow me are also political scientists, although there are presumably some social scientists from other fields as well.
https://twitter.com/RachelStrohm/status/1123599539758350338
[15] http://egap.org/metaketa

(2015) recently tested the Ultra Poor Graduation Pilots poverty reduction program across six low income countries, and found that it was broadly successful at increasing consumption and improving health. In psychology, the Many Labs 2 project at the Center for Open Science brought together 186 researchers from more than 36 countries to try to replicate the findings of 28 major studies in the field. They ultimately found that only half of the studies could be reproduced (Klein et al. 2018).

Coordinated multi-site studies are an exciting frontier in the design of generalizable research. They can also be integrated with existing systematic review infrastructure through the use of evidence gap maps to identify topics in need of multi-site research. However, there is also a clear need for additional investment in staffed organizations to manage these initiatives beyond the scope of a single project. EGAP and the Center for Open Science both represent promising starts to this. (The Ultra Poor Graduation Pilots project was run by Innovations for Poverty Action and the Jameel Poverty Action Lab, which are research implementers rather than organizations specifically focused on improving generalizability in economics.)

6. Conclusion

This discussion of generalizability has four main implications for the practice of social science research. First, in terms of research design, internal validity should be valued over increasing a study's generalizability by continually adding more cases or observations. There are two reasons for this. Most obviously, internal validity by definition has to precede the generalizability of results. We should not be attempting to generalize about a study which doesn't actually produce accurate knowledge of the world. In addition, this allows researchers to specialize in the type of internally valid research which they prefer. This should increase the overall

quality of research produced, which will also increase the quality of the subsequent generalizability assessments.

Second, one thing that stands out from this discussion is the degree to which generalizability can only be assessed via systematic review when researchers make their data publicly available. This is true for both qualitative and quantitative data. Because synthesis methods are available for both types of data, studies using either of these methods can contribute to the cumulation of knowledge.

Third, the fragmented landscape of systematic review organizations in the social sciences may be due partly to low demand from researchers. It's important for social scientists to engage with systematic reviews as both consumers and producers, and to advocate for disciplinary investments in the organizations which produce them.

Fourth, researchers should continue to collaborate on multi-site studies. Developing the institutional infrastructure necessary to manage this type of research is a good role for established academics with high social capital within their disciplines. Bringing younger researchers on board can help them to launch careers with a strong focus on generalizability from the outset.

Bibliography

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects? Regression and Representative Causal Effects." *American Journal of Political Science* 60(1): 250–67.

Banerjee, Abhijit et al. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348(6236): 772–88.

Brady, Henry, and David Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers.

Das, Jishnu, Quy-Toan Do, Karen Shaines, and Sowmya Srikant. 2013. "U.S. and Them: The Geography of Academic Research." *Journal of Development Economics* 105: 112–30.

Gerring, John. 2012. *Social Science Methodology: A Unified Framework*. 2nd ed. Cambridge: Cambridge University Press.

Green, Bart, Claire Johnson, and Alan Adams. 2006. "Writing Narrative Literature Reviews for Peer-Reviewed Journals: Secrets of the Trade." *Journal of Chiropractice Medicine* 5(3): 101–17.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33(2–3): 61–83.

Higgins, Julian, and Sally Green. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane Collaboration.

King, Gary, Robert Keohane, and Sidney Verba. 1994. *Designing Social Inquiry Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Klein, Richard A. et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1(4): 443–90.

Lieberman, Evan S. 2016. "Can the Biomedical Research Cycle Be a Model for Political Science?" *Perspectives on Politics* 14(04): 1054–66.

Ozier, Owen. 2019. "The Reproducibility Crisis and the Case of Deworming." The World Bank. Policy Research Working Paper no. 8835.

Pepinsky, Thomas B. 2019. "The Return of the Single-Country Study." *Annual Review of Political Science* 22: 187–203.

Rosenzweig, Mark, and Christopher Udry. 2019. "External Validity in a Stochastic World." *Review of Economic Studies* 0:1-39.

Sandelowski, Margarete, Julie Barroso, and Corrine I. Voils. 2007. "Using Qualitative Metasummary to Synthesize Qualitative and Quantitative Descriptive Findings." *Research in Nursing & Health* 30(1): 99–111.

Scott, James. 1985. *Weapons of the Weak: Everyday Forms of Peasant Resistance*. New Haven: Yale University Press.

———. 1998. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.

———. 2009. *The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia*. New Haven: Yale University Press.

Shadish, William, Thomas Cook, and Donald Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2nd ed. Boston: Cengage Learning.

Shrout, Patrick, and Joseph Rodgers. "Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis." *Annual Review of Psychology* 69: 487–510.

Snilstveit, Birte, Raag Bhatia, Kristen Rankin, and Beryl Leach. 2017. *3ie Evidence Gap Maps: A Starting Point for Strategic Evidence Production and Use*. International Initiative for Impact Evaluation. Working Paper. http://3ieimpact.org/evidence-hub/publications/working-papers/3ie-evidence-gap-maps-starting-point-strategic-evidence.

Tannenwald, Nina. 1999. "The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use." *International Organization* 53(3): 433–68.