Improving the Replicability and Generalizability of Inferences in Quantitative Communication Research

Jacob A. Long

School of Journalism and Mass Communications

University of South Carolina

This is a pre-copyedited, author-produced version of an article accepted for publication in Annals of the International Communication Association. *The accepted manuscript (AM) is the final draft author manuscript, as accepted for publication, including modifications based on referees' suggestions, before it has undergone copyediting, typesetting and proof correction. This is sometimes referred to as the post-print version. The version of record,*

Long, J. A. (2021). Improving the replicability and generalizability of inferences in quantitative communication research. *Annals of the International Communication Association*, 1–14. doi: 10.1080/23808985.2021.1979421

is available online at:

https://www.tandfonline.com/doi/full/10.1080/23808985.2021.1979421

Author Note

Jacob A. Long (Ph.D., The Ohio State University) is Assistant Professor in the School of Journalism and Mass Communications at the University of South Carolina. Correspondence concerning this article should be addressed to Jacob A. Long, School of Journalism and Mass Communications, University of South Carolina, 29208. Email: jacob.long@sc.edu ORCiD: 0000-0002-1582-6214

Abstract

This paper discusses the assessment of quality of quantitative communication research in light of the so-called “replicability crisis” that has affected neighboring disciplines. For social scientific research, it is useful to think of research results as estimates which include error. I propose a framework suited to a variable field like communication, factoring in all sources of error, for assessing the quality of research. In communication research, greater consideration of generalizability is essential, which is at once both a higher standard than replicability but also a goal that should increase it. Furthermore, more explicit discussion of generalizability may help to further internationalize the discipline by clarifying the limitations of the large portion of research conducted within a narrow subset of world cultures.

Keywords: inference, replication, total survey error, generalizability

Improving the Replicability and Generalizability of Inferences in Quantitative Communication Research

As the social sciences grapple with renewed concerns over the credibility of quantitative research findings, it is important for communication researchers to do the same. Although communication has avoided most of the public embarrassment suffered by other disciplines due to unflattering replication efforts, this may not last. Even when not in times of crisis, self-reflection is essential for cumulative science. The goal of this paper is to engage in some of this reflection and offer suggestions for improving research designs and their assessment. In a discipline as methodologically heterogeneous as communication, it is important to adopt explicit standards and norms for the evaluation of research quality. Of course, communication research is also epistemologically heterogeneous; this paper focuses on the part of the field that makes inferences using quantitative methods from a social scientific perspective. Methodological pluralism in communication research can be a great strength for answering the field's many and difficult questions, but the extent to which different approaches can be synthesized efficiently depends on common understanding of how to evaluate them. I suggest a framework for evaluating research and designing studies, taking inspiration from an approach widely used by survey methodologists. First, I give a brief description of this framework and describe how it follows logically from this framework that generalizability and replicability are both important goals for quality research. I then give some background on the so-called "replication crisis" and discuss its relationship to communication research. I conclude by giving suggestions for how researchers can evaluate and plan their work accordingly.

Research Design as Error Reduction

Communication research, when conducted within the social science paradigm, aims to make data-based inferences. Researchers and consumers of research understand that, according to prevailing philosophies of science, even well-designed studies cannot remove all doubt about the inferences one may make based on the data collected (see Kelly & Westerman, 2020 for a thorough treatment of this topic). In an overall body of work, evidence may accumulate to progressively increase the confidence one has in a particular scientific hypothesis, but it is rare for an inference or theory to be taken as certainly accurate in all its particulars. In other words, "science is about confidence rather than Truth with a capital t" (Kelly & Westerman, 2020, p. 183). My suggestion is to think of scientific claims as akin to statistical estimates; we cannot create a perfect study, but we can use theory and design to strategically reduce the plausibility of errors. Much like the process of statistical estimation, one may produce estimates that are known to reduce the expected size of errors, but one cannot definitively remove the possibility of errors. This way of thinking about research and applying widely-known principles draws upon the practices and philosophy adopted in survey methodology in recent decades.

In research on survey methodology, the Total Survey Error framework (TSE; see, e.g., Groves, 2004) has become "the dominant paradigm" (Groves & Lyberg, 2010, p. 849) for planning and evaluating research. Put succinctly, the TSE approach is "optimally allocating the available survey resources to minimize [error] for key estimates" (Biemer, 2010, p. 818). Specifically, TSE is about formal consideration of sources of error beyond those induced by random error inherent in random sampling. There is no single, canonical enumeration of these additional elements of error, but several appear with frequency (see Lavrakas & Kosicki, 2018 for one such listing). The errors that most often come to mind in this context have to do with

representation of the target population: random sampling error, coverage error, and nonresponse error. Another set relate to conceptual issues: construct misspecification, measurement errors, and statistical errors (i.e., using the wrong statistical techniques). TSE also makes a distinction between two types of error: variance and bias (see Biemer, 2010 for a figure trying to capture this distinction). Some error contributes only to variability — sample size (which affects random sampling error) and random measurement error chief among them — while other sources of error contribute bias. Some errors related to sampling and survey design can also cause bias, such as when cell phone users are omitted from telephone surveys. Certain types of measurement error can cause the resulting statistics to become biased as well, especially but not only when there is a specification error. The TSE approach also encourages explicit consideration of available resources (principally money but also infrastructure, time, and other factors). Finding potential flaws in one's work is a skill well-developed in most graduate programs, but dealing more openly with justifiable tradeoffs in the face of reasonable constraints is just as important and more challenging.

Empirical communication research consists of much more than surveys, of course, so there are limits to the plug-and-play applicability of the TSE. That said, the discipline can benefit from thinking in terms of error reduction in the vein of the TSE approach. Indeed, this should be nothing entirely new to most trained quantitative researchers but instead a reframing of the fundamentals we already know. This kind of framework is useful for emphasizing the identification of threats to the overall validity of a study, doing so at the design stage, and factoring them in within a fixed cost structure. What I propose, in other words, is that one can think in terms of Total *Study* Error. When a study makes a claim, rather than making an up or

down assessment of its credibility, this approach calls for considering the amount of uncertainty and bias in its design within a general model.

Implicit in this idea is that one can conceive of research findings as parameter estimates (see Lundberg, Johnson, & Stewart, 2021 for a full treatment of this idea). Just as one can estimate that smokers comprise a certain proportion of the population of Germany, researchers can estimate (for example) the causal effect of having peers who smoke on whether adolescents take up smoking. Often, the latter sort of finding is ultimately communicated as the presence or absence of effect, but the commonly-accepted best practice across the social sciences is to both plan studies with an effect size in mind (Cohen, 1992) and emphasize the size of observed effects when reporting findings (Kelley & Preacher, 2012). It is also common to provide confidence intervals for such estimates to reflect underlying uncertainty. But just as the margin of error in a public poll understates the true level of uncertainty, scientific inferences typically have more uncertainty than what is conveyed by *p* values, confidence intervals, and so on. An error-reduction approach encourages thinking about the size of effects and the extent to which the study design allows for accurate estimation of such effects. This includes considerations that are typically conceived of as non-statistical, like confusing correlation and causation; a study design that cannot satisfactorily separate cause from effect can be said to contribute error in this regard. And, as will soon be discussed, this mode of thinking can help researchers consider how to balance concerns about a study's replicability and generalizability.

To be clear, what is being proposed here is to simply give more structure to what is often a more ad hoc process of research evaluation and justification. Researchers need to be able to talk about contributions in terms of how findings avoid and are susceptible to error. In characterizing literatures, studies can be treated as individual pieces of evidence, each addressing

another potential source of error. This is an example of what Shapiro (2002) calls an "evolving scientific discourse" and "complex analysis of the evidence supporting a claim," which is "our most powerful tool for generalization" (p. 492). A compulsion to talk explicitly about target populations for inference is also valuable given that social scientists often treat Western, educated samples and contexts as default (Cheon, Melani, & Hong, 2020) despite such people making up a small portion of the world population (Henrich, Heine, & Norenzayan, 2010). In communication, quantitative research appears to be the most likely to decontextualize findings (Walter, Cody, & Ball-Rokeach, 2018), perhaps contributing to the relatively poor representation of non-Western authors in the research record (Chakravartty, Kuo, Grubbs, & McIlwain, 2018). If one's goal is to generate findings broadly applicable to humanity, then findings from or about the Global South, for instance, should not be treated as either unjustifiably non-generalizable nor duplicative of research conducted in or focused on places like the United States.

Replicability and Generalizability

It may not be obvious how the idea of taking an error-reduction approach to research design relates to replication, a topic of growing importance in the social sciences. In this discussion, replication refers to efforts to exactly reproduce the procedures of a study in order to compare the results of the replication with the original. When it comes to this kind of replication, making the procedures, measurements, and context as similar as possible to the original is a virtue. Consistent with sampling theory, replications give some insight into the uncertainty of an observed result. Each time a replication gives a similar result to previous studies, it suggests those procedures produce estimates that are consistent. Low variability of results does not guarantee a lack of error in the underlying scientific inference one makes, but it does provide evidence against the source of error being randomness. Of course, there are ways to anticipate

the replicability of a study without performing a replication, such as by considering the design's statistical power. Well-executed replications are always informative, however, since replicability-enhancing design decisions are based on assumptions — such as how statistical power is based on the expected effect size. Other issues, such as the quality of measurement, also affect replicability without being explicitly factored into most power calculations. Replicable findings are not necessarily correct; the design being reproduced may still be vulnerable to important errors. For instance, an experimental stimulus may lack construct validity despite reliably producing an empirically observed effect. Regardless, designing studies to give confidence that the results would be approximately equal if the procedures were performed again is an exercise in error reduction.

A group known as the Open Science Collaboration (OSC; 2015) attempted to replicate 100 studies sampled from 3 top psychology journals published in 2008. OSC researchers worked to directly replicate the materials and methods of the originals, involving the original authors for advice whenever they were willing. Ultimately, only 36% of studies had the replication also obtain a result with p less than .05. A more generous threshold yielded a 47% replication rate. No effort comparable to the OSC's has (yet) been attempted for communication research, but problems with replicability have not been isolated to psychology. A more recent replication effort (Camerer et al., 2018) selected all social science experiments published in *Science* and *Nature* meeting certain criteria between 2010 and 2015. Depending on the standard applied, estimates of replication range from 57% to 67% for the 21 studies replicated. A similar effort in experimental economics (Camerer et al., 2016) yielded a replication rate between 61% and 78%, depending on the metric. A field that likely carries more institutional prestige than any of the aforementioned, cancer biology, had a mere 11% success rate in an attempt to replicate 53 "landmark" studies

(Begley & Ellis, 2012). Likewise, after large heart disease clinical trials were required to pre-register their studies, the proportion with a statistically significant benefit on the primary outcome dropped to 8% from 57% prior to the requirement (Kaplan & Irvin, 2015).

There is little reason to believe that research in communication is much better situated than neighboring disciplines when it comes to replicability (Dienlin et al., 2021), although it should be said that large-scale replication efforts in other fields have been largely restricted to experimental research and nonprobability samples, a set of design features which only characterizes around half of quantitative communication studies (Erba, Ternes, Bobkowski, Logan, & Liu, 2018). Keating and Totzkay (2019) recently studied the prevalence of replication attempts in the field of communication; they found that although studies claiming to be replications were common (comprising 1 in 7 manuscripts across several journals), they were typically "conceptual" replications which do not reproduce the procedures of the original. Although conceptual replications have evidential value, they are not as helpful for assessing the replicability of the field. One means to assess a research literature's replicability from a bird's-eye view is to examine the distribution of reported *p* values (Simonsohn, Nelson, & Simmons, 2014). Vermeulen and collaborators (2015) recently took this approach to examine research in 25 communication journals in the years 2010 through 2012. Their findings suggest that approximately 35% of published findings in communication are false positives due to a mixture of publication bias and low statistical power — and this ignores inferential errors from other sources like construct validity, external validity, and so on. The estimate is quite consistent with previous replication efforts in neighboring disciplines, making it a reasonable guess that replicability in communication research is (or was) similar to those fields.

An underappreciated factor in psychology's so-called "replication crisis" is generalizability. To briefly define terms, for the purposes of this paper "generalizability" refers to the extent to which a scientific finding is likely to reappear in another setting. The new setting may be a close replication, an extension into a different sample population, a conceptual replication with new measurements, stimuli, or analytic method, or some combination of these. This is distinct from a narrower conception of generalizability as a statistical concept focused only on sampling (see Lee & Baskerville, 2003 for a more thorough explication of different kinds of generalizability). As pointed out by McEwan, Carpenter, and Westerman (2018), one can think of replicability as the weakest form of generalizability; can a study generalize to nearly identical circumstances? Replicability — that is, when attempts to recreate the circumstances of the original research generate the same or similar findings — is a necessary but not sufficient condition for high generalizability. To be clear, generalizability is treated here as continuous; a finding could generalize reliably only within a finite universe of conditions, such as within a country, age group, media system, and so on. Such a finding is not a universal law of human behavior but may nonetheless be a valid scientific finding. A more generalizable finding is one that is expected to reappear across more of these kinds of settings.

One criticism of the OSC replication project claimed that it may have underestimated the replicability of psychology research due to inconsistencies in the procedures of the replications compared to the originals (Gilbert, King, Pettigrew, & Wilson, 2016), such as by using non-equivalent sample populations or modes of data collection.. Another replication project, which replicated a number of studies in moral psychology in 25 separate labs, found substantial variation depending on sample characteristics (Schweinsberg et al., 2016). For example, US samples consistently provided larger effect sizes for the effects that were largely researched in

the US prior the replication efforts. General population samples differed in effect size compared to student samples as well. It is difficult to say how many failures to replicate have something to do with sample variability, given the existence of several other research practices that are known to reduce the likelihood of replication[1]. Once it can become accepted that a given finding will replicate using the same or similar procedures with the same or similar population under study, the remaining work to do is establishing that the essential finding holds as the procedures change operationally (but not conceptually) and the populations under study differ.

A response among communication researchers and beyond to concerns about replicability is the desire to implement a set of practices known as open science (Lewis, 2020). A full review of such practices are beyond the scope of this article, but they are focused on transparency about the research process to help others understand what was done to better evaluate findings. These are laudable goals and advance the cause of reducing errors in scientific inferences. A fairly modest proposal coming from open science advocates that is not well known but merits serious consideration for communication research is the establishment of a new norm of reporting "constraints on generality" (COG; Simons, Shoda, & Lindsay, 2017). Simons and collaborators propose including a section in each article in their field (psychology) in which authors describe the target population to whom they believe their results should generalize along with a justification for that reasoning. This is also a space in which authors can make clear which population(s) to whom generalization is doubtful or more uncertain. A COG statement makes

[1] To be clear, the available evidence suggests the largest causes of the replication crisis in psychology were a combination of publication bias, questionable research practices, and research designs with poor statistical power (Vermeulen et al., 2015). But when better statistical and disclosure practices are established, issues of generalizability will become one of the most important remaining obstacles for valid inferences.

generalizability an explicit consideration in the review process, making it more likely design choices that enhance generalizability are rewarded. Researchers are also more able to guide potential replicators to choose appropriate sample populations or direct follow-on studies to test specific boundary conditions. As a variable level field, communication researchers have even more work to do as inferences are often either about or constrained by social and structural factors that vary across and within countries. For example, some political communication research conducted in the United States may have limited conceptual or operational generalizability to multi-party systems, authoritarian regimes, and democracies with more influential public media.

Sampling, Causality, and Generalizability in Communication Research

An area in which this mode of thinking can be applied is to the problem of how researchers can balance desires for generalizable samples with the desire to make solid causal claims. Although communication researchers may not be great at fielding representative samples for their research, they are excellent at doing the kind of navel-gazing necessary to establish these facts. The most useful recent example of this is a comprehensive analysis of the samples used in mass communication studies in 6 journals[2] (Erba et al., 2018). Over the years 2000 through 2014, 51% of all the included studies of human subjects used student samples. When not using student samples, authors typically opted for other non-representative samples; 83% of studies overall used nonprobability samples.

---

[2] *Communication Research, Journal of Broadcasting & Electronic Media, Journal of Communication, Journal of Computer-Mediated Communication, Journalism and Mass Communication Quarterly*, and *Mass Communication and Society*.

Few would argue with the claim that it is better, all else being equal, to have a sample that is chosen using some form of random selection from a suitable frame representing the target population of scientific inferences. The norm in human subjects research about communication — at least those studying mass media — is to sacrifice generalizability at the sample selection stage. When taking an error-reduction perspective, one can conclude this is sometimes justifiable. For most populations of interest to social scientists, probability sampling is one of the most expensive design choices available. With resources limited, as they always are, there are some sound justifications for prioritizing other aspects of the design over trying to acquire the most representative sample possible. As a clear example of this kind of tradeoff, if measures must be taken via fMRI, it is virtually impossible without massive-scale cooperation to do a study of a representative sample of, say, the United States. The more researchers need from each participant to minimize specification or measurement error — time or proximity chief among those needs — the better the arguments in favor of lower quality samples (or more sampling error of various kinds). As Shapiro (2002) wisely noted, a narrow focus on sample representativeness can plausibly harm generalizability. As research programs and areas that require these tradeoffs mature, large-scale cooperation to address sample problems becomes more appropriate and justified. Developmental neuroscientists, who rely on fMRI on samples of children and adolescents, teamed together with the National Institutes of Health to launch the Adolescent Brain Cognitive Development Study, which involves imaging the brains of over 10,000 children who were recruited in a probability sample of the United States (Garavan et al., 2018). Communication scientists in general would do well to initiate ambitious projects like these to address research problems that cannot be tackled by lone researchers or labs. In the social sciences, the American National Election Study is a useful model as well.

Ruling out the best choice of sample does not necessarily mean giving up all consideration of sample quality. One way to think about this in terms of reducing inferential error is to say one may accept sampling error but still seek to minimize coverage error, errors induced by not including enough of the population of interest in the sampling frame. The difference in coverage between a student sample and samples available cheaply via vendors like mTurk is considerable. There are options of varying quality at many price points for acquiring samples, especially if the study design can be embedded in an online format. For example, opt-in panels like Qualtrics are not representative but can at least draw a diverse sample. YouGov, on the other hand, uses opt-in recruitment but uses statistical methods to draw samples that rather closely approximate probability samples by a number of benchmarks (Pew Research Center, 2016; Rivers, 2016). This does come at a cost that is usually several times higher to the researcher, however, compared to panels like Qualtrics that may allow for quota sampling to roughly resemble some number of target demographics but cannot rightly be described as representative of any socially meaningful population. In order to reduce inferential errors, explicit specification of a target population is necessary. This is also a useful cue for researchers as more attention is paid to the extent to which social scientific findings tested on samples from wealthy, Western countries may not generalize to the rest of the world (Rad, Martingano, & Ginges, 2018). And when findings do not generalize, that often means the scientific community learned this fact due to a failure to replicate.

The point of all this is maximizing the ability to generalize research findings beyond the specific people and context under analysis. The representativeness of the sample is just one part of generalizability. If the goal is inference about brain activity, generalizability is less jeopardized by a non-representative sample in an fMRI machine than it is by asking a representative sample

to self-report the physiological details of their brain activity. Put simply, if valid measurement or causal inference is not possible with a given sample, then there is nothing to generalize. Nevertheless, it still must be said that in either case the generalizability is compromised by the design. Although on one hand researchers must be pragmatic and accept some tradeoffs in order to advance knowledge, they still must accept that some types of research will be very difficult to conduct in a generalizable way. That does not mean one must suspend the normal principles of scientific judgment because of the difficulty of the topic; advances will simply be costlier, more incremental, and more tentative. With that in mind, I proceed to discuss some reasons why it behooves researchers to consider sample quality as a major design consideration.

First, probability sampling satisfies a statistical assumption at the heart of inferential statistics. Tests of significance are interpretable only as measures of the population from which the units are sampled. For nonprobability samples, the *design* offers no quantitative support for extrapolating even to the larger population from which the units were non-randomly sampled. Inferences afforded by study design have higher evidential value than those that depend on statistical models. By way of analogy, consider randomized experiments[3]. Although researchers apply a statistical test to compare the experimental groups on the dependent variable, the meaningfulness of this test is a function of the design. The same pattern of results might occur if assignment to groups was decided on some non-random basis — after all, the intervention(s) will still have the same effect — but the non-random assignment introduces uncertainty about how to interpret the statistical results. The typical experimental researcher would likely find assurances about the similarity of non-randomly assigned groups on demographic or attitudinal attributes to

[3] This comparison is explained in more technical detail by Mercer, Kreuter, Keeter, and Stuart (2017).

be a cold comfort when considering the overall evidential value of an experiment. The threat to the internal validity of statistical inferences from quasi-experiments is conceptually equivalent to the threat to generalizability of statistical inferences from nonprobability samples. Threats, however, are just that: a potential problem, not a complete invalidation.

Concerns about low-quality samples are not new. The most obvious concern with student samples in particular is that they may be quite different from the true population of interest on a number of relevant attributes (Henry, 2008; Sears, 1986). Average differences between the nonprobability sample and the target population are not necessarily damning. This is especially true if one focuses on the direction of effects (a pathology of the way significance testing dichotomizes results) rather than the magnitude, a parameter estimate that is likely to be inaccurate though not necessarily biased. More pernicious are unmeasured and unanticipated interactions, especially interactions between variables of interest with characteristics of the sample that are greatly underrepresented, overrepresented, and/or do not vary within the sample.

One might assume that problems relating to sample quality can be anticipated at the design stage, or at worst after the fact. This type of scientific reasoning — anticipating potential interactions and other issues related to the sample — is the pragmatic solution for learning from such studies. Sometimes researchers are correct to assume their nonprobability samples are not meaningfully different from the target population *in terms of direction and/or magnitude of effect*, even if the sample is quite different in other ways. This is difficult to know without empirical testing, however (see Exadaktylos, Espín, & Brañas-Garza, 2013 for an example of such a test). Caution is advised in assuming that reasoning about the results alone, especially *post hoc*, will reliably identify the cases in which it matters. Paul Lazarsfeld's trick in his review of *The American Soldier* is instructive here (Lazarsfeld, 1949). He initially reports several topline

findings of the study as a demonstration of their obviousness: better-educated soldiers were more neurotic (intellectuals being well-known for their instability), soldiers from the American South better tolerated the tropical island climates at which they were stationed (owing to their upbringing in warmer weather), and soldiers preferred being sent back to their homes over fighting until the German surrender (they logically want to avoid the danger)[4]. Lazarsfeld asks, rhetorically, "why, since they are so obvious, is so much money and energy given to establish such findings" (p. 380)? He then reveals that it is because those findings are all false; he deliberately reported each of them in a way exactly opposite of the actual results to show how easy it is for something to seem obvious and expected once one has seen the data. In other words, it is important in programmatic research to establish that assumptions about less-than-ideal samples do, in fact, hold up.

Finally, a common way of thinking about these issues is to say the goal of some studies, especially experiments, is to test (usually causal) relationships between variables and that other studies, especially probability-sampled surveys, are for estimating population parameters. As an example of this type of argument, Sparks (1995) notes — with qualification — that sometimes researchers assume "the operative processes producing an experimental result are taken to be so fundamental that the issue of generalizability appears to be trivial" (p. 277). In a scientific discipline, the claim that a process is essentially universal should be established on empirical grounds even if we are permitted to engage in informed speculation about the likelihood that it is the case. In the realm of clinical medical research, in which it is much more difficult to acquire representative samples but the fundamental biological processes would seem to be more

[4] I omit some references to the differences between White and Black soldiers.

obviously consistent across people, there is growing recognition of the perilous consequences of homogeneous samples (Denny & Grady, 2007; Johnson, Fitzgerald, Salganicoff, Wood, & Goldstein, 2014; Mosenifar, 2007). Realizing that women, racial minorities, or people with psychiatric disorders respond differently to medical treatments is the exact type of inference that seems obvious in retrospect but is only just beginning to be appreciated after decades of clinical trials. A far more modest argument for the usefulness of experiments on unrepresentative samples is that they "are useful for demonstrating *that* the human mind *can* work a certain way" (emphasis in original; Potter, Cooper, & Dupagne, 1995, p. 282). This is true, but with a caveat: A null result may be uninterpretable. If something happened in the lab, then of course it shows that it can happen. But if "nothing" happens, one has little leverage for saying that it cannot or does not ever happen.

Pitting the establishment of a relationship between variables against parameter estimation creates a false dichotomy. It is intuitive and maps neatly onto expectations about the difference between experimental and survey research, but just as doing a survey and doing an experiment are not mutually exclusive, parameter estimation and relationship detection are not either. There is growing recognition that science benefits when researchers focus less on whether *p* values are one or another side of a threshold (like .05), instead interpreting effect sizes and their plausibility (Gelman & Carlin, 2014). The statistical models used to analyze experiments are estimating a parameter: the effect size. Assuming the size of an effect matters — it almost always does, whether the threshold for substantive significance is small or large — it is essential to estimate the size accurately. Given the threat of hidden moderators, it is not known *a priori* whether the nonprobability sample will differ just in magnitude or also in direction of effect. By way of analogy, consider whether a political poll ought to be based more on whether it identifies the

winner correctly or estimates the results most closely, irrespective of the getting the winner right: If a political poll reports Candidate A leads Candidate B 60% – 40% and the election outcome is 50.1% – 49.9% in favor of A, was the poll accurate? The same logic applies to estimates of causal effects.

Design-based and Model-based Inference

Insisting upon the benefits of using probability samples does not address all the reasons for their rarity in communication research. Although part of the reason they are not often used is likely because their benefits are not appreciated by researchers and therefore not sufficiently awarded with better odds of publication, there are other barriers. Sometimes a research problem — or a research budget — calls for focusing on minimizing errors elsewhere and allocating resources accordingly. Furthermore, as noted earlier, there are degrees of sample quality. To think about the options, consider the ideas of design-based and model-based inference (e.g., Koch & Gillings, 2006; Sterba, 2009). The randomization of experimental participants to conditions and the random selection of respondents to a survey are examples of design features that *in and of themselves* provide a means for valid statistical inference. That is, the validity of the statistical inferences in those cases is based on those key design features. Model-based inference, on the other hand, involves the use of statistical models and their corresponding assumptions for statistical inference. This is not to say there are no assumptions in design-based inference; the assumptions are just about the design (e.g., that the procedure for random assignment is indeed random) rather than the statistical model. One does model-based inference, for example, when controlling for confounding variables to try to assess causation in non-experimental data.

Model-based inference is another route to generalizability. In other words, it can sometimes be possible to use statistical adjustment to correct for sampling errors. The most

obvious case of using adjustment to correct sampling errors is the use of weighting methods with survey data[5]. It would be very unusual for a communication researcher to generate weights when not provided by a vendor, but the same goal can sometimes be achieved by using covariates in regression models (Bollen, Biemer, Karr, Tueller, & Berzofsky, 2016). Effective use of covariates facilitates model-based inference in which sampling error may be reduced thanks to the statistical model. This works best when the sample is relatively similar to the target population, but it is not absolutely essential. Sample matching and related methods are forms of model-based inference for causal analysis in non-experimental data (Austin, 2011). Longitudinal studies, like panel surveys, often are analyzed using methods that blend design-based inference (from the temporal ordering of observations) and model-based inference with covariate adjustments and the like.

What is the lesson, then? In a debate over whether the rarity of representative sampling in communication research made the discipline pre-scientific, Lang (1996) called for a distinction between statistical inference and logical inference. In Lang's explanation, statistical inference is equivalent to what I have just called design-based inference. Logical inference is basically equivalent to the scientific reasoning discussed earlier: Using content knowledge and common sense to make decisions about the generalizability of samples. These judgments may not be as easily made or supported as one would hope, however. Model-based inference is another approach that may not measure up to design-based inference, but has real advantages over what Lang calls logical inference. Researchers do need diverse samples for model-based inference,

[5] This is not always an example of model-based inference. Sometimes, probability of selection is known due to the sampling design and in this case the weighting is still part of design-based statistical inference. Other cases, like the use of population parameters to generate weights, are model-based inferences targeted to reduction of sampling error.

and sometimes very careful and technical thinking about proper model specification, but it comes with substantial benefits at a relatively low cost. In this way, researchers can continue to guard aspects of internal validity closely but can conceptualize the generalizability of the sample as something other than all (representative) or nothing (nonprobability with minimal diversity).

Beyond Humans and Samples

Although the focus has been on studies of human subjects, the principles apply just as well for other units of analysis. For instance, in the case of content analysis it is becoming increasingly feasible to generate sampling frames that include most or all the population of media content of interest. It will sometimes be the case that the entire population can be analyzed thanks to automated methods; technical or human constraints may require that a smaller, perhaps randomly-sampled subset of content is subjected to analysis. Platforms like Twitter generally restrict access such that not all Tweets can be downloaded; researchers can ask and pay for access to all of them, but whether this is necessary will depend on the context. Given the computational tools now available, content analyses of news coverage that rely on a single source (in the United States, often *The New York Times*) are harder to justify on the basis of data availability. Even human coding techniques can be scalable beyond small numbers of trained raters via crowdsourcing (Budak, Goel, & Rao, 2016). Put succinctly, a potential strength of the growth of computational communication research is that it can greatly reduce coverage error and sometimes sampling error in situations in which doing so was previously intractable.

The uses of computational methodology extend beyond reducing coverage error. Another common feature of this varied research area is that data are often behavioral in nature (Shah, Cappella, & Neuman, 2015). Generating accurate self-reports of communication behavior is among the most difficult and longest-running challenges in the discipline (e.g., Hovland, 1959;

Hutchens, Eveland, Morey, & Sokhey, 2018; Scharkow, 2016). This means many studies that rely on self-reported communication variables are vulnerable to measurement error that may be subject to unknown biases. Some sources of data used in computational research, like social media or log data, include actual communications of interest for many research questions and therefore justify themselves on the basis of measurement quality. This is not to say big data and computational methods are a panacea; especially for inferences about people, there are often major tradeoffs in terms of representativeness, measurement of psychological variables that are typically well-measured by questionnaires, and (usually) the inability to experimentally manipulate variables of interest.

Another area for consideration is in experimental design. Beyond the randomization of treatments and the recruitment of samples is the experimental stimuli. On one hand, one must consider mundane realism, something fairly well-appreciated (e.g., Iyengar, 2011). Furthermore, in designs in which media are the stimulus, distortions are possible due to the forced exposure that rarely occurs in everyday life (Hovland, 1959; Stroud, Feldman, Wojcieszak, & Bimber, 2019). If a finding occurs due to exposure being forced rather than the content of the communication, the error is in the interpreted meaning of the stimulus. A more general concern about stimuli, however, is whether they are representative of the theoretically relevant stimuli that exist in the "real world." To exert the most control over the manipulation and enhance statistical power, researchers tend to lean on a small set of stimuli in media research. The risks of doing so, however, are two-fold: one is that the stimuli are not sufficiently similar to the communications to which researchers want to generalize and the other is that the messages may be incidentally confounded with one or more other variables (Slater, 1991). In some cases, it may be possible to construct a sampling frame and literally sample message stimuli randomly from

the real population of messages of interest. In others, messages may need sufficiently substantial researcher manipulation that the best that can be done is an effort to generate a diverse set of stimuli.

Methodological Pluralism in Communication Research

An advantage the communication discipline has over some of its peers is its methodological pluralism. Any randomly sampled group of quantitative communication researchers is likely to have the collective skillset for impressive methodological triangulation. This can be achieved in several ways. If developing a research program that is largely built on well-executed experiments on student samples, a logical next step is to try to use a higher-quality sample. Doing so might require some sacrifices to internal validity, but as part of an overall body of work one can sacrifice internal validity in one study to improve the sample quality in another. Although everyone would like each study to have as little error of all kinds as possible, doing so may not always be practical. A smart alternative is building on weaknesses in a piecemeal way as needed. A researcher might have one study with a strong experimental intervention, another that improves measurement, and another that at least conceptually replicates on a more representative sample. Needless to say, when resources are available and it is technically feasible, it is better to perform the best version of the experiment with the best measures on a representative sample. A logical way to preserve resources and learn from mistakes is to first do a study in a cost-efficient way and then add the more expensive components when expectations of success and payoff for reducing a specific type of error are highest. Adopting publication formats like Registered Reports (Nosek & Lakens, 2014) can reduce the risk for researchers who worry about null findings from expensive designs as well.

An impressive example of this kind of methodological synthesis in a single study comes from Goodall, Slater, and Myers (2013). First, the study uses a high-quality, probability sample of U.S. Americans with sufficient sample size for the amount of distinct experimental conditions. Participants were exposed to a news article as the experimental intervention embedded within the survey. The design manipulated the topic of the story (auto accident, non-auto accident, or violent crime) and whether alcohol was mentioned as a cause of the incident in the story. The stories, within conditions, were randomly selected from a representative sample of U.S. newspapers that had been content-analyzed to select relevant stories. Other design features included measures for which self-reports are likely to be valid (emotion, policy attitudes) and manipulation checks. Overall, the study could boast of multi-pronged design-based inference: The sample was a probability sample of a theoretically meaningful target population, the stimuli were randomly sampled from the population of media available to the human population, and the stimuli were administered in the context of a randomized experiment. The research team was rewarded with several strong statistical results — many would pass muster even under the most stringent redefinitions of statistical significance — and are further bolstered by the design which does not require the subjective hedging one must engage in when the design is of lower quality. Thanks to the sample size, the null results are informative as well because they are estimated with precision.

Not all research programs will be amenable, either due to technical or financial constraints, to such thorough methodological triangulation in single or perhaps even several studies. Nevertheless, it shows the power of combining the several methodological foci of communication research. Other possibilities include linkage analysis (de Vreese et al., 2017), in which inferences are made about media effects based on measurements of media

exposure/attention in surveys and content analysis of those media. Generally, experimentation and survey research are also no longer mutually exclusive options thanks to the ability to embed many types of stimuli in web surveys. Although not within the scope of this essay, many communication scientists regularly traverse the quantitative-qualitative divide as well, where deeper and different insights are available.

These various areas of expertise are particularly useful for creating a cumulative body of evidence that can address many potential sources of error. Of course, that requires the field to appreciate and understand work coming from multiple methodologies in order to build a coherent literature. As interest grows in creating the conditions for more replicable studies, researchers should consider how efforts to make findings generalize at the design stage can enhance replicability as well. When some sources of error cannot be prevented, transparency is essential to help the field understand how to integrate new evidence. Overall, prospects are bright for communication research if we capitalize on our disciplinary strengths.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424. doi: 10.1080/00273171.2011.568786

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533. doi: 10.1038/483531a

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*, 817–848. doi: 10.1093/poq/nfq058

Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., & Berzofsky, M. E. (2016). Are survey weights needed? A review of diagnostic tests in regression analysis. *Annual Review of Statistics and Its Application*, *3*, 375–392. doi: 10.1146/annurev-statistics-011516-012958

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*, 250–271. doi: 10.1093/poq/nfw007

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*, 1433–1436. doi: 10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637. doi: 10.1038/s41562-018-0399-z

Chakravartty, P., Kuo, R., Grubbs, V., & McIlwain, C. (2018). #CommunicationSoWhite. *Journal of Communication*, *68*, 254–266. doi: 10.1093/joc/jqy003

Cheon, B. K., Melani, I., & Hong, Y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study samples. *Social Psychological and Personality Science*, 1948550620927269. doi: 10.1177/1948550620927269

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi: 10.1037/0033-2909.112.1.155

de Vreese, C. H., Boukes, M., Schuck, A., Vliegenthart, R., Bos, L., & Lelkes, Y. (2017). Linking survey and media content data: Opportunities, considerations, and pitfalls. *Communication Methods and Measures*, *11*, 221–244. doi: 10.1080/19312458.2017.1380175

Denny, C. C., & Grady, C. (2007). Clinical research with economically disadvantaged populations. *Journal of Medical Ethics*, *33*, 382–385. doi: 10.1136/jme.2006.017681

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., … de Vreese, C. (2021). An agenda for open science in communication. *Journal of Communication*, *71*, 1–26. doi: 10.1093/joc/jqz052

Erba, J., Ternes, B., Bobkowski, P., Logan, T., & Liu, Y. (2018). Sampling methods and sample populations in quantitative mass communication research studies: A 15-year census of six journals. *Communication Research Reports*, *35*, 42–47. doi: 10.1080/08824096.2017.1362632

Exadaktylos, F., Espín, A. M., & Brañas-Garza, P. (2013). Experimental subjects are not different. *Scientific Reports*, *3*, 1213. doi: 10.1038/srep01213

Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R. Z., Heeringa, S., … Zahs, D. (2018). Recruiting the ABCD sample: Design considerations and procedures. *Developmental Cognitive Neuroscience*, *32*, 16–22. doi: 10.1016/j.dcn.2018.04.004

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651. doi: 10.1177/1745691614551642

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*, 1037–1037. doi: 10.1126/science.aad7243

Goodall, C. E., Slater, M. D., & Myers, T. A. (2013). Fear and anger responses to local news coverage of alcohol-related crimes, accidents, and injuries: Explaining news effects on policy support using a representative sample of messages and people. *Journal of Communication*, *63*, 373–392. doi: 10.1111/jcom.12020

Groves, R. M. (2004). *Survey errors and survey costs* (2nd ed.). Hoboken, N.J: Wiley.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*, 849–879. doi: 10.1093/poq/nfq065

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*, 29–29. doi: 10.1038/466029a

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, *19*, 49–71. doi: 10.1080/10478400802049936

Hovland, C. I. (1959). Reconciling conflicting results derived from experimental and survey studies of attitude-change. *American Psychologist*, *14*, 8–17. doi: 10.1037/h0042210

Hutchens, M. J., Eveland, W. P., Jr., Morey, A. C., & Sokhey, A. E. (2018). Evaluating summary measures of heterogeneous political discussion: The critical roles of excluded cases and discussion with people holding extreme views. *Communication Methods and Measures*, *12*, 276–294. doi: 10.1080/19312458.2018.1479844

Iyengar, S. (2011). Experimental designs for political communication research. In *Sourcebook for political communication research: Methods, measures, and analytical techniques* (pp. 129–148).

Johnson, P. A., Fitzgerald, T., Salganicoff, A., Wood, S. F., & Goldstein, J. M. (2014). *Sex-specific medical research: Why women's health can't wait* (p. 32) [A Report of the Mary Horrigan Connors Center for Women's Health & Gender Biology at Brigham and Women's Hospital]. Brigham and Women's Hospital. Retrieved from Brigham and Women's Hospital website: https://www.brighamandwomens.org/assets/bwh/womens-health/pdfs/connorsreportfinal.pdf

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLOS ONE*, *10*, e0132382. doi: 10.1371/journal.pone.0132382

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*, 225–239. doi: 10.1080/23808985.2019.1632218

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152. doi: 10.1037/a0028086

Kelly, S., & Westerman, D. (2020). Doing communication science: Thoughts on making more valid claims. *Annals of the International Communication Association*, *44*, 177–184. doi: 10.1080/23808985.2020.1792789

Koch, G. G., & Gillings, D. B. (2006). Inference, design-based vs. Model-based. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/0471667196.ess1235.pub2

Lang, A. (1996). Standpoint: The logic of using inferential statistics with experimental data from nonprobability samples: Inspired by cooper, Dupagne, potter, and sparks. *Journal of Broadcasting & Electronic Media*, *40*, 422–430. doi: 10.1080/08838159609364363

Lavrakas, P. J., & Kosicki, G. M. (2018). Survey research in mediated communication. In P. M. Napoli (Ed.), *Mediated communication* (Vol. 7, pp. 225–260). Berlin: De Gruyter Mouton.

Lazarsfeld, P. F. (1949). The American soldier—An expository review. *The Public Opinion Quarterly*, *13*, 377–404.

Lee, A. S., & Baskerville, R. L. (2003). Generalizing generalizability in information systems research. *Information Systems Research*, *14*, 221–243. doi: 10.1287/isre.14.3.221.16560

Lewis, N. A., Jr. (2020). Open communication science: A primer on why and some recommendations for how. *Communication Methods and Measures*, *14*, 71–82. doi: 10.1080/19312458.2019.1685660

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, *86*, 532–565. doi: 10.1177/00031224211004187

McEwan, B., Carpenter, C. J., & Westerman, D. (2018). On Replication in Communication Science. *Communication Studies*, *69*, 235–241. doi: 10.1080/10510974.2018.1464938

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*, 250–271. doi: 10.1093/poq/nfw060

Mosenifar, Z. (2007). Population issues in clinical trials. *Proceedings of the American Thoracic Society*, *4*, 185–188. doi: 10.1513/pats.200701-009GC

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. doi: 10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi: 10.1126/science.aac4716

Pew Research Center. (2016). *Evaluating online nonprobability surveys*. Retrieved from http://www.pewresearch.org/2016/05/02/evaluating-online-nonprobability-surveys/

Potter, W. J., Cooper, R., & Dupagne, M. (1995). Reply to Sparks's critique. *Communication Theory*, *5*, 280–286. doi: 10.1111/j.1468-2885.1995.tb00110.x

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*, 11401–11405. doi: 10.1073/pnas.1721165115

Rivers, D. (2016, May 13). Pew Research: YouGov consistently outperforms competitors on accuracy. Retrieved October 30, 2018, from YouGov website: https://today.yougov.com/topics/finance/articles-reports/2016/05/13/pew-research-yougov

Scharkow, M. (2016). The accuracy of self-reported internet use—A validation study using client log data. *Communication Methods and Measures*, *10*, 13–27. doi: 10.1080/19312458.2015.1118446

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. doi: 10.1016/j.jesp.2015.10.001

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515–530. doi: 10.1037/0022-3514.51.3.515

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*, 6–13. doi: 10.1177/0002716215572084

Shapiro, M. A. (2002). Generalizability in communication research. *Human Communication Research*, *28*, 491–500. doi: 10.1111/j.1468-2958.2002.tb00819.x

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

Slater, M. D. (1991). Use of message stimuli in mass communication experiments: A methodological assessment and discussion. *Journalism Quarterly*, *68*, 412–421. doi: 10.1177/107769909106800312

Sparks, G. G. (1995). Comments concerning the claim that mass media research is "prescientific": A response to Potter, Cooper, and Dupagne. *Communication Theory*, *5*, 273–280. doi: 10.1111/j.1468-2885.1995.tb00109.x

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research*, *44*, 711–740. doi: 10.1080/00273170903333574

Stroud, N. J., Feldman, L., Wojcieszak, M., & Bimber, B. (2019). The consequences of forced versus selected political media exposure. *Human Communication Research*, *45*, 27–51. doi: 10.1093/hcr/hqy012

Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., van de Velde, B., & Oegema, D. (2015). Blinded by the light: How a focus on statistical "significance" may cause p-value misreporting and an excess of p-values just below .05 in communication science. *Communication Methods and Measures*, *9*, 253–279. doi: 10.1080/19312458.2015.1096333

Walter, N., Cody, M. J., & Ball-Rokeach, S. J. (2018). The ebb and flow of communication research: Seven decades of publication trends and research priorities. *Journal of Communication*, *68*, 424–440. doi: 10.1093/joc/jqx015