JOURNAL FOR DERADICALIZATION

Fall 2022
No. 32
ISSN: 2363-9849

Sampling and the Curse of the Case Study

Michael J. Williams[a1]
[a]The Science of P/CVE

Abstract
Regardless of how the outcomes of a given P/CVE program are measured or evaluated, a fundamental, implicit (if not explicit) research question is: to what extent can obtained results apply to others within a given population. In short, to what extent can the results apply to others, in general; what is the so-called generalizability of the findings? In other words, the outcomes of a given P/CVE program are relatively useless unless they can be replicated, and the likelihood of replication is synonymous with generalizability. Therefore, it is virtually impossible to overstate the importance of generalizability with respect to P/CVE research and evaluation, and generalizability is fundamentally a function of how well sampling is performed. Therefore, it is also virtually impossible to overstate the importance of sampling with respect to an evidence-based approach to P/CVE. This research methods brief describes fundamental issues (including potential pitfalls and means to avoid them) with respect to sampling in the context of P/CVE program design and evaluation: including issues related to sampling online "Big Data," and "nested" (multi-level/hierarchical) program/research designs.

Article History
Received Mar 8, 2021
Accepted Mar 21, 2022
Published Sept 30, 2022

Keywords: Preventing Violent Extremism, PVE, Countering Violent Extremism, CVE, P/CVE, Design, Evaluation, Sampling

Sampling and the Curse of the Case Study

Regardless of what measures are selected to research or evaluate a given P/CVE program, a fundamental, implicit (if not explicit) research question is: to what extent can obtained results apply to others within a given population. In short, to what extent can the results apply to others, in general; what is the so-called generalizability of the findings? In other words, the outcomes of a given P/CVE program are relatively useless unless they can be replicated, and the likelihood of replication is synonymous with generalizability. Therefore, it is virtually impossible to overstate the importance of generalizability with respect to P/CVE research and evaluation. Although the topic of generalizability has multiple facets (and is the subject of its own branch of theory; see Shavelson & Webb, 1991), it is fundamentally a function of how

[1] Corresponding Author Contact: Michael J. Williams, Email: m.williams@thescienceofpcve.org, Twitter: @MickWilliamsPhD, ORCID iD https://orcid.org/0000-0001-5630-9814

well sampling is performed. Therefore, it is also virtually impossible to overstate the importance of sampling with respect to an evidence-based approach to P/CVE.

Sampling Fundamentals

Sampling is the process of selecting units (e.g., individuals, households, organizations, time periods) from a population of interest (Groves et al., 2009). As mentioned, the primary function of sampling is to enable inference about a larger population, based upon those selected into the sample. If sampling is performed well, research or evaluation results can apply to those who were not in the sample. If sampling is performed poorly, there is little-to-no hope of making inferences to the larger population of interest (ibid.). Even relatively small data collections, for relatively small-scope research or evaluation projects—for example, interviews to inform a process evaluation—need adequate sampling procedures to be both valid and reliable. Consider that, even for a small number of interviewees, appropriate sampling is still relevant toward obtaining data that might represent the diversity of perspectives present in the population.

To facilitate inference beyond a given sample, some form of random sampling (from the population of interest) must be employed: to reduce the likelihood that those selected into the sample differ from the population on factors that systematically vary with the selection procedure. In other words, a sample is only as good as it adequately represents a population of interest, and random selection strategies are the only ways to produce a sample that plausibly represents a cross-section of the population of interest (about whom the research or evaluation intends to make inferences). **However, even when sampling procedures are performed flawlessly—the inferences that can be drawn are strictly limited to the population from which the sample was drawn** (also referred to as the "sampling frame"). In other words, strictly speaking, one cannot make inferences from a study about those who were never candidates for inclusion in the study.

The burden of proof in asserting that a given finding is replicable beyond a given population rests with those who would assert such an overgeneralization (Sagan, 2011). In other words, those who would claim that a given finding should extend beyond a given sampled population also are responsible for demonstrating how those outside of the originally studied population are similar on *all* factors plausibly related to the outcome(s) of interest.

That task is theoretically impossible, because it is unknowable whether those outside a given population differ significantly (perhaps in ways that were not measured) from those of another population. If one thinks that a given research or evaluation finding should apply to a population different from the source of a previous sample, the only valid way to find out is replication with those sampled from the new population: to rerun the research or evaluation with participants from the new population of interest.

Three Types of Samples (But Only Two That Are Acceptable)

The following list of sampling strategies covers the two primary types of valid sampling procedures: probability sampling and purposive sampling. (See their respective footnotes for additional resources on each subtype). Additionally, a third type of sampling is mentioned—convenience sampling—that should be avoided. This list, and its accompanying references, are intended to help you to consider carefully, perhaps more broadly, the range of sampling strategies that might be available for a given P/CVE research or evaluation project: to assist you in selecting the strongest design that available time and treasure permit.

Probability Sampling

This form of sampling entails random selection to draw a sample from a known population of eligible sampling units (as mentioned, known as the sampling frame). Probability sampling plausibly results in a kind of cross-section of the population of interest. Thus, as mentioned, findings about the sample plausibly generalize to the broader population.

Simple random sampling. As its name implies, simple random sampling entails drawing a sample completely at random from a sampling frame.[2]

Stratified random sampling. This form of sampling also draws sampling units at random from a sampling frame, but it does so by drawing a given number of units from subcategories ("strata") of interest. For example, the procedure could select an equal number of participants from various demographic backgrounds (e.g., ethnicity, sex, age), even if those demographic attributes are not equally distributed in the population. Such oversampling of the smaller strata can be statistically accounted for (i.e., adjusted via sample weighting

[2] Simple random sampling: https://tinyurl.com/Simple-random-sampling

procedures), and the advantage of this procedure is that subcategories of interest can be guaranteed to be represented within the sample in sufficient quantities to afford statistical inferences.[3]

Multi-stage sampling. This form of sampling has a "nested" property (Groves et al., 2009). First, "primary" selections are made; for example, cities could be randomly selected from a province. Then, within those primary units, "secondary" units are selected (for example "neighborhoods," and perhaps, still further subsampled from tertiary units [e.g., households], etc.; ibid.). As another example, from said cities (primary units), schools (secondary units) could be randomly selected, and classrooms (tertiary units) could be randomly selected from those schools. The intent of such a sampling strategy is to increase the "efficiency" of the research design. In other words, it is intended to minimize the number of sampling units needed to afford both sufficiently reliable statistical inferences about the population, and sufficiently narrow margins of error.[4]

Purposive Sampling

Purposive sampling (or "purposeful sampling") is a type of nonprobability sampling that, therefore, *cannot* be used for drawing reliable statistical inferences about a population (Groves et al., 2009). Instead, however, purposive sampling remains a valid means to "reach the hard-to-reach," to facilitate theory development about such populations (see Palinkas et al., 2015). (More on this to follow, under § "The Curse of the Case Study.")

Criterion sampling. This type of sampling selects cases that meet a specified condition; for example, membership in a given violent extremist organization, or participation in a given P/CVE program.[5]

Snowball sampling. This type of sampling selects respondents based on referrals from previous respondents.[6]

[3] Stratified random sampling: https://tinyurl.com/Stratified-random-sampling
[4] Multi-stage sampling: https://tinyurl.com/Multi-stage-sampling
[5] Criterion sampling: https://tinyurl.com/Criterion-sampling
[6] Snowball sampling: https://tinyurl.com/Snowball-sampling

Convenience Sampling

This form of sampling draws units that are readily available, without the use of a sampling frame or other systematic sampling strategies (Sample, n.d.). Convenience sampling is contrary to the very purposes of sampling: the purposes either to generalize knowledge to others from a population, or to seek specialized knowledge from those well-informed on a topic of interest. It is mentioned here only as an example of what not to do.

Online Data and Nested Research Designs

"Big Data:" Random Sampling Still Applies

A big secret about "big data" is that it often is not as big as one might assume (Leetaru, 2019). Given the vastness of available online data, virtually all P/CVE research and evaluation projects that collect online data, rely upon samples of data: if only by virtue of the fact that the online sampling frame (e.g., of social media posts) is ever-expanding (Ruggiero, 2020). Consequently, given the enormity of available online data, "Big Data" is often a euphemism for comparatively small sample sizes (Leetaru, 2019). Therefore, it is even more important to exercise great skill in sampling online data, given the potential enormity of the inferential leap between a relatively small sample and its relatively large sampling frame (Kaplan et al., 2014).

To illustrate the importance of carefully sampling online data, consider a hypothetical study focused on assessing the prevalence of disinformation spread, via twitter, by members a given violent extremist organization. Assume that the project obtains an ostensibly "random sample" of tweets based on a) randomly selecting 100 known members of a given violent extremist organization, then b) selecting the first 50 tweets, from those persons, for a given year: resulting in a sample of 5000 tweets. The results of that analysis might be considerably different than if 5000 tweets had been randomly sampled from the pool of *all* available tweets from those persons throughout the same year: due, perhaps, to history effects[7] associated with the relatively narrow time frame of the first data collection method (Leetaru, 2019).

[7] History effects are events that are external to, yet co-occur with, a given longitudinal experiment, and that inadvertently affect an outcome variable of interest.

Sampling Strategy to Increase the Statistical Power of Nested Designs

As mentioned, multi-stage sampling (e.g., sampling at two levels; for example, randomly sampling schools from within randomly sampled cities) is a means of increasing statistical power of research designs: improving the likelihood of detecting significant effects. It does this not by virtue of increasing the number of sampling units, but by virtue of improving the so-called efficiency of the research design: a design which more effectively accounts for homogeneity within (vs. between) the levels of the sampling design (compared to a simple random sample, Kim et al., 2013).[8] Conceptually, multi-stage sampling is similar to the aforementioned stratified random sampling, whereby more reliable estimates of segments of the population can be obtained by randomly sampling within the designated strata (ibid.).

Likewise, the above concepts can be applied to improve the statistical power of nested research designs (also referred to as "cluster," or "hierarchical" designs). Nested designs are those where research participants are not wholly independent from one another, but who share a group or "cluster" factor in common. For example, the outcomes of participants in a multi-site prison rehabilitation program are not wholly independent from each other; instead, they are nested within a given prison. To increase the statistical power of nested research designs, it is more advantageous to increase the number of sampled clusters (e.g., the number of sampled sites) than to increase cluster size (i.e., the number of participants selected from within each cluster), although both improve power (Baldwin et al., 2011).

Therefore, if given the choice (for example) between adding one more research participant to each of 10 clusters, or increasing the design simply to include one more cluster of 10 participants (note that both options would require 10 additional participants) it would be more statistically advantageous (i.e., more powerful) to add one more cluster of 10. This assumes, however, that the clustering factor is meaningful: that it accounts for a statistically significant effect. Otherwise, cluster designs are statistically disadvantageous/inefficient: requiring a larger number of participants (than a simple random sample) merely to afford analysis/estimation of the (insignificant) clustering factor.[9]

[8] This assumes that the levels of a multi-stage sampling design account for significant variation in the outcome variable(s) of interest. Otherwise, multi-stage sampling designs are less efficient than simple random samples (see Whittemore & Halpern, 1997).

[9] In other words, by including a nesting factor in a given analysis, at least one degree of freedom is sacrificed, which unnecessarily weakens statistical power unless the nesting factor is significantly associated with the dependent variable(s).

Furthermore, given that nested/cluster/hierarchical sampling designs are conceptually similar to stratified sampling designs, the same is true of stratified random sampling designs; they are statistically efficient only to the extent that the strata are significantly associated with the outcome(s) of interest (Kim et al., 2013). The more distinct/homogenous the strata, the greater the gains in precision of a stratified random sample compared to a simple random sample, and vice versa.[10]

This is also why sample weighting—to adjust a sample so that it more closely represents a population on various known population parameters (e.g., employing raking techniques, also known as iterative proportional fitting)—is unnecessary, if one intends to make parameter estimates, or between-group comparisons, with respect to outcomes that are uncorrelated with the factors used to rake/calculate the sampling weights. For example, raking/weighting a sample along dimensions of race, sex, etc., is useful only to the extent that the outcomes of interest are correlated with such variables.

Participant Recruitment

Participants' data are the treasure of research and evaluations. Given the importance of obtaining a sampling large and diverse enough to represent the broader population of interest, recruitment materials, incentives, and procedures must be made appealing to the spectrum of prospective participants (Williams, 2022). Consider that even a well-planned random sampling strategy can be foiled if selection bias is high. In short, researchers and evaluators need to exercise creativity and technical savvy—to do all that is reasonably within their power—to encourage those who are invited to participate to accept that invitation and subsequently provide their precious data (ibid.).

To design effective recruitment procedures and attractive recruitment materials, there is an immense body of research on persuasion and social influence that can be drawn upon to enhance a project's appeal to would-be participants (see Cialdini, 2013; see Pratkanis, 2011). Additionally, cultural experts (who should be a part of the research team) can be invaluable both in developing and implementing participant recruitment plans designed to appeal to the

[10] Insofar as stratifying a sample produces relatively homogeneous groups of sampling units (i.e., within each stratum) sampling error is reduced. Therefore, estimates based on such strata have higher precision compared to a simple random sample.

population of interest. In designing recruitment procedures, take pains to ensure participants' privacy and other forms of personal safety, remove logistical barriers to participation (e.g., transportation barriers), and incentivize participants sufficiently (Williams, 2022). Recognize that some types of participants will be unable to accept financial incentives to participate: for example, police, other city officials, or prisoners (ibid.). Nevertheless, research involving such persons should conceive of other ways to incentivize their participation—*what is in it for them?*—so that the project's sample does not unduly suffer from selection bias (e.g., obtaining a sample skewed toward participants who happen to be idle/bored, or otherwise desperate enough to participate; ibid.).

Furthermore, it might be necessary to gain the assistance of others (e.g., program staff) to assist in the process of participant recruitment. Therefore, earn the buy-in of recruiters, in part, by ensuring that they are sufficiently informed of the project's purpose. Additionally, recruiters should be trained in their task, and provided with materials (e.g., wording for promotional announcements, participant information sheets, consent forms) to facilitate that task.

In designing recruitment procedures and associated materials, a balancing act—present in virtually any P/CVE research or evaluation projects (except those that exclusively involve archival data)—is to make sufficient disclosures to would-be participants regarding the nature of the research or evaluation, without being counterproductive. The competing interests, in this regard, are to provide enough information to participants to permit them to make reasonably informed decisions about whether to participate, without unnecessarily alarming, offending, boring, or otherwise prompting their disinterest. As mentioned, there is a vast repertoire of social influence principles that can be drawn upon to inform this effort (see Cialdini, 2013; see Pratkanis, 2011). Additionally, consistent with practices recommended by the Ankara Memorandum on Good Practices for a Multi-Sectoral Approach to Countering Violent Extremism, P/CVE programs (and, by extension, research and evaluations of those programs) should avoid associating P/CVE initiatives with any religion, culture, ethnic group, nationality, or race (Global Counter Terrorism Forum, 2013).

The Curse of the Case Study

Avoid generalizing too soon (if at all), from qualitative data. Recall that generalizability is primarily a function of how well sampling has been performed. Typically, the goal of qualitative work is not to generalize across populations; though, too often, qualitative data are abused this way: as though they represent generalizable "lessons learned." Such is the curse of the case study. A case study—a sample of one—is invalid for generalizing to a population of any size: not even a population of two. Consider the following thought experiment: even if a population consisted of only two individuals, and we randomly sampled one of them—a random sample of a staggering 50% of the population—it would be impossible to generalize to the population of two, because there is no way to verify the extent to which the selected case is typical, or atypical, of the larger group. In other words, was the person selected for our case study normal, or somehow strange, compared to the larger group (in this case, compared to the other [unknown] person). The answer is unknowable, because we have no means of triangulating our findings about the sampled person to other points of reference. Therefore, case studies can be used for theory development, but not for generalization.

Caution: Criterion Samples

Case studies are not the only type of research designs at risk of overgeneralization. To perform research or evaluation of a given P/CVE program—for example, to demonstrate participant outcomes—ideally, participants for the study would be randomly sampled from a frame of all current and/or former participants of the program. In practice, such a frame may be nonexistent. In such cases, research/evaluation participants are identified through other means: for example, a partial list of known participants are invited via email, or a call for participants is announced through the social media account(s) of the public-facing partner(s) of the P/CVE program. In either case, participants are included in the study of the P/CVE program if they meet a certain criterion: namely, that they have participated in said program.

However, recall that criterion sampling is a type of purposive sampling: a type of nonprobability sampling. Therefore, the results from such a study cannot be said to generalize to the program's participants overall, because it does not randomly sample from the pool of all of the program's participants. Even if such a study takes pains to randomly

sample from those who volunteer to participate—perhaps going so far as to take a stratified random sample of those volunteers, to ensure that key demographic categories are sufficiently represented in the sample—recognize that the study's generalizability still is limited to the sampling frame: in the previous example, those who were sent the broadcast invitations to participate in the study.

From this example, it is not hard to imagine that those who come forward to participate in the study might differ from the P/CVE program's participants overall. Those who come forward to participate might be, for example, the "overachievers," or "fans," of the program. Alternatively, the sample might be biased toward representing those who have "nothing better to do," than to participate in the study. In either case, it is plausible that their reported (or observed) outcomes might differ from the population of the program's participants overall.

Nevertheless, criterion samples are superior to case studies in at least two ways. First, a criterion sample still can serve as the basis for a longitudinal (i.e., within-participant / pre – post) research design. Therefore, even if the criterion sample is biased toward over (or under)-achievers, participants serve as their own control group, and changes in their outcomes, over time, still can be assessed (West et al., 2004). Second, criterion samples still can be employed in between-group research designs that compare the P/CVE program participants to one or more comparison groups: for example, to those who have never encountered the program.

In both of the above designs—whether longitudinal or between-groups—the results of the analyses are, nevertheless, still circumscribed to the sampling frame (in the previous example, those who were sent the broadcast invitation to participate in the study). Such limitations must be made explicit in discussions of the research or evaluation's results, so that readers may make better-informed judgements regarding how the sample's characteristics might have affected the outcome(s) of interest.

This also suggests one of the easiest, yet effective—and cost effective—steps that any P/CVE program can take to vastly improve its so-called "evaluability" (the extent to which it can be evaluated): maintain a comprehensive roster of the contact information of each and every program participant. If the program does not yet have such a roster, resolve to begin it as of "today" so-to-speak. From "today" forward, the likelihood of the program being able to

demonstrate—convincingly—its (hopefully beneficial) effects will have improved dramatically: providing a comprehensive sampling frame of program participants, from which a probability sample can be drawn.[11]

Conclusion

A fundamental requirement, of an evidence-based approach to P/CVE, requires that the methods of its research and evaluations be made sufficiently explicit: to provide a recipe for replication. Consequently, it can be considered scientific misconduct should research or evaluations fail to describe their sampling procedures. As discussed, sampling procedures are critically important regarding the extent to which findings could be expected to generalize. Failing to describe a study's research methods, including its sampling procedures—explicitly and comprehensively—is to conceal them in a "black box:" the researchers or evaluators stating, essentially, "Trust us; we're doctors." It is scientifically appropriate to call foul on such empirical omissions, and—given the potentially high stakes of P/CVE—it is also ethically appropriate.

Acknowledgement

Special thanks to Dr. Peter J. Martini III, PhD, for his pre-submission review of this article's discussion of multi-stage sampling.

[11] Of course, due consideration must be paid toward safeguarding participant rosters and that such data are both collected and stored in accord with applicable laws.

References

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. In *Psychological Methods* (Vol. 16, Issue 2, pp. 149–165). American Psychological Association. https://doi.org/10.1037/a0023464

Cialdini, R. B. (2013). *Influence: Science and practice* (5th ed.). Pearson.

Global Counter Terrorism Forum. (2013). *Ankara memorandum on good practices for a multisectoral approach to countering violent extremism*. https://www.thegctf.org/documents/10162/72352/13Sep19_Ankara+Memorandum.pdf

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science*, *7*(4), 342–346. https://doi.org/10.1111/cts.12178

Kim, Y. J., Oh, Y., Park, S., Cho, S., & Park, H. (2013). Stratified sampling design based on data mining. *Healthcare Informatics Research*, *19*(3), 186–195. https://doi.org/10.4258/hir.2013.19.3.186

Leetaru, K. (2019). *The Big Data revolution will be sampled: How "Big Data" has come to mean "small sampled data."* Forbes. https://www.forbes.com/sites/kalevleetaru/2019/02/17/the-big-data-revolution-will-be-sampled-how-big-data-has-come-to-mean-small-sampled-data/?sh=1ffc0051199e

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y

Pratkanis, A. R. (Ed.). (2011). *The science of social influence: Advances and future progress.* Psychology Press.

Ruggiero, G. (2020). *How Big Data killed sampling*. Altavia. https://medium.com/altavia/how-big-data-killed-sampling-72205a8e1b6a

Sagan, C. (2011). *The demon-haunted world: Science as a candle in the dark*. Ballantine Books.

Sample. (n.d.). Better Evaluation. Retrieved June 15, 2019, from https://www.betterevaluation.org/en/rainbow_framework/describe/sample

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Sage.

West, S. G., Biesanz, J. C., & Kwok, O. M. (2004). Within-subject and longitudinal experiments: Design and analysis issues. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The SAGE Handbook of Methods in Social Psychology* (pp. 287–312). https://doi.org/10.4135/9781412976190.n13

Whittemore, A. S., & Halpern, J. (1997). Multi-stage sampling in genetic epidemiology. *Statistics in Medicine*, *16*(1–3), 153–167. https://doi.org/10.1002/(sici)1097-0258(19970130)16:2<153::aid-sim477>3.0.co;2-7

Williams, M. J. (2022). Incentivizing Participants P/CVE Research, Evaluation, & Program Participants. *Journal for Deradicalization*, *31*, 164–174.

JOURNAL FOR DERADICALIZATION

Fall 2022
No. 32
ISSN: 2363-9849

About the JD Journal for Deradicalization

The JD Journal for Deradicalization is the world's only peer reviewed periodical for the theory and practice of deradicalization with a wide international audience. Named an "essential journal of our times" (Cheryl LaGuardia, Harvard University) the JD's editorial board of expert advisors includes some of the most renowned scholars in the field of deradicalization studies, such as Prof. Dr. John G. Horgan (Georgia State University); Prof. Dr. Tore Bjørgo (Norwegian Police University College); Prof. Dr. Mark Dechesne (Leiden University); Prof. Dr. Cynthia Miller-Idriss (American University Washington D.C.); Prof. Dr. Julie Chernov Hwang (Goucher College); Prof. Dr. Marco Lombardi, (Università Cattolica del Sacro Cuore Milano); Dr. Paul Jackson (University of Northampton); Professor Michael Freeden, (University of Nottingham); Professor Hamed El-Sa'id (Manchester Metropolitan University); Prof. Sadeq Rahimi (University of Saskatchewan, Harvard Medical School), Dr. Omar Ashour (University of Exeter), Prof. Neil Ferguson (Liverpool Hope University), Prof. Sarah Marsden (Lancaster University), Prof. Maura Conway (Dublin City University), Dr. Kurt Braddock (American University Washington D.C.), Dr. Michael J. Williams (The Science of P/CVE), Dr. Mary Beth Altier (New York University) and Dr. Aaron Y. Zelin (Washington Institute for Near East Policy), Prof. Dr. Adrian Cherney (University of Queensland), Dr. Håvard Haugstvedt (Center for Research on Extremism, University of Oslo).

For more information please see: www.journal-derad.com

Twitter: @JD_JournalDerad
Facebook: www.facebook.com/deradicalisation

The JD Journal for Deradicalization is a proud member of the Directory of Open Access Journals (DOAJ).

ISSN: 2363-9849

Editor in Chief: Daniel Koehler