Paluck, Elizabeth Levy (2010). The promising integration of qualitative methods and field experiments. *The ANNALS of the American Academy of Political and Social Science 628* (1), 59–71.

Psillos, Stathis (2004). A glimpse of the secret connexion: Harmonizing mechanisms with counterfactuals. *Perspectives on science 12* (3), 288–319.

Rubin, Donald B (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66* (5), 688.

Seawright, Jason (2016). *Multi-method social science: Combining qualitative and quanti- tative tools*. Cambridge University Press.

Waldner, David (2016). Invariant causal mechanisms. *Qualitative & Multi-Method Research 14* (1-2), 28–33.

Qualitative and Multi-Method Research | Fall 2023, Volume 21.2 | https://doi.org/10.5281/zenodo.8418917

Measuring Costly Concepts: Validation Samples for Measuring Many-N Cases

Sarah Moore
Northwestern University

Costly concepts are concepts that are expensive or otherwise resource-intensive to obtain measurement for over many cases. Costly concepts are present across the social sciences, though particularly in the subnational study of comparative politics. Subnational democracy, local-level armed group presence, and municipal corruption are all costly concepts for which measurement requires fine-grained data that may be practically impossible to collect for many units where the data are not already available to researchers. In the absence of actual measures of costly concepts, scholars will often substitute measurement by using proxy variables in empirical analyses, which causes non-random measurement error where measurements of the costly concept and proxies are not identical. This non-random measurement error means we risk conducting biased analyses when we cannot overcome the structural challenges that preclude precise measurement of costly concepts.

For example, the quantitative literature on non-state armed actors and violent conflict has overwhelmingly relied on local violence data to measure the presence of armed groups throughout a territory (for more extensive reviews of this literature, see Arjona and Castilla 2022; Vela Barón 2021)for obvious reasons, on violence. Yet, civil war is about much more than violence. We argue that the focus on violence hinders our understanding of the most common type of armed conflict in the world today. In particular, equating civil war and violence leads to (i. However, measuring armed group presence through violence fails as a proxy in ways we would easily expect given existing theory on civil war violence (Arjona 2016; Kalyvas 2000). Alternative measures of armed group presence entail gathering extensive knowledge from local experts through fieldwork (e.g., Arjona 2016; Aponte-González, Hirschel-Burns, and Uribe 2023). However, fieldwork-based approaches to measuring local-level armed group presence are incredibly expensive and thus limited to a reduced number of cases.

How do we know the extent to which a proxy can reliably substitute measures of our costly concept? How do we improve proxies or other measures when the proxy alone is unreliable? In this work, I develop methodological tools to understand the performance of existing proxies for costly concepts and inform more sophisticated measurement strategies based on the direct measurement of a subset of cases where obtainable. Here, I focus on a summary of the former, in which I develop a framework for collecting and analyzing validation samples wherein the accurate measurement of the costly concept is obtained for a set of cases to discern the performance of a proxy over three dimensions: the extent of disagreement, the variation in the disagreement, and the predictive features of the disagreement. I further assess the type of sample required to best estimate proxy performance relative to three potential options: a random sample, a stratified random sample, or a theoretically informative sample.

My overarching argument is that having at least some information about the relative performance of a potential proxy is better than uninformed analysis with said proxy. Collecting validation samples of at least a subset of cases to obtain direct measurements of a costly concept allows researchers to understand the degree to which a proxy and concept of interest converge and provides insight into the circumstances where they do not. To illustrate the proposed methodological framework and discuss the trade-offs of some of the sampling approaches available for these validation samples in the larger paper, I rely on simulated data. I use the concept of armed group

presence relative to the oft-used violence proxy to motivate the data generation process for the simulation study and I present some of this illustration and my findings here.

Proxy Performance

Three dimensions of interest characterize proxy performance relative to proxy-costly concept disagreement: extent, variation, and predictive features. *Disagreement* is measured as any case where the proxy measure and the costly concept measure are not equal. The *extent* of disagreement is the proportion of cases where there is proxy-costly concept disagreement relative to the number of measured cases. The *variation* in disagreement is the degree to which *extent* of disagreement varies across all cases and is calculated given the sample variance of the *extent* of disagreement. Lastly, *predictive features* of disagreement are potential variables that contribute to additional knowledge about the cases where there is disagreement between the proxy and the costly concept of interest. Although this could be derived several ways, an efficient way is to estimate a feature selection model to determine which of a set of specified variables are meaningful in predicting proxy-costly concept disagreement.

In measuring armed group presence, the extent of disagreement is the proportion of cases where there was violence and no presence, or where there was no violence, but armed groups were present. The variance of disagreement between violence-presence is the dispersion of cases where violence inaccurately measures presence relative to the number of cases sampled (i.e., the sample size).[1] Lastly, the predictive features of measurement disagreement between armed group presence and violence may be variables like state capacity, historical local communist organization, or economic development. In my work, I use classification trees to perform feature selection and determine which variables are important in predicting measurement disagreement. However, any appropriate modeling scheme that highlights important predictive features of disagreement is suitable.

What type of sample is necessary?

While collecting at least some information about the performance of proxies relative to actual measures of costly concepts is helpful, this collection ought to be guided by a systematic sampling approach. So, what sampling design is best for uncovering and estimating proxy performance? I specifically test three different sampling strategies: random sampling, stratified random sampling, and theoretically informative sampling.[2] I find that in the case of the simulation study, the three different sampling strategies provide substantively similar information over the three proxy performance dimensions. Though these sampling approaches should also be tested using real world data, as I do in later work, these initial findings indicate that researchers should feel comfortable employing any systematic sampling approach among those explored here that most efficiently meets their additional data collection needs.

Contribution and Further Work

My larger project on difficult-to-measure concepts provides scholars with a unified framework related to the challenges and existing tools for concept measurement in the social sciences and beyond, as well as where there are gaps for continued methodological improvements. In the work I summarize here, I have focused on measurement of concepts that are directly observable, though costly to measure. I have further provided a framework to assess just how bad existing measurements are and how potential case insights can inform us of the location of bad measurement. Through the larger research project, I hope to show the ways that case-based research can help to refine large-N quantitative research toward the end of expanding the utility of the multi-method toolkit.

References

Aponte-González, Andrés F., Daniel Hirschel-Burns, and Andres Uribe. 2023. "Contestation, Governance, and the Production of Violence Against Civilians: Coercive Political Order in Rural Colombia." *Journal of Conflict Resolution*. https://doi.org/10.1177/00220027231177591

Arjona, Ana. 2016. *Rebelocracy*. New York: Cambridge University Press.

Arjona, Ana, and Juan Pablo Castilla. 2022. "The Violent Bias in the Study of Civil War." Working Paper.

Kalyvas, Stathis N. 2000. *Logic of Violence in Civil War*. Cambridge University Press.

Vela Barón, Mauricio. 2021. "Identifying Armed Group Presence Using Hidden Markov Models." Masters Thesis, Dept. of Mathematics and Statistics: Washington University in Saint Louis.

1 This is informative that *variance* dimension is not always useful and depends on the measurement properties of the underlying variables given that we are here presumably measuring violence and presence as binary variables.

2 This is a sampling approach I develop in the paper, wherein cases are stratified along primary strata of interest and then combined into secondary strata based on their theoretical likelihood. This secondary stratification helps to condense the strata allocations and eliminates the unnecessary allocation of some sampled units to primary strata combinations that are highly unlikely.