Public Opinion Quarterly, Vol. 00, No. 0, 2021, pp. 1–3

BOOK REVIEW

Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady T. West, eds. *Experimental Methods in Survey Research: Techniques That Combine Random Sampling with Random Assignment*. Hoboken, NJ: John Wiley and Sons. 2019. 544 pp. $120.00 (cloth).

ALEXANDER COPPOCK
Yale University

Broadly, this volume is concerned with research designs that combine the virtues of both random sampling and randomized experimentation. Randomized experiments license inferences about average causal effects on the grounds that treatments are randomly allocated to subjects with known probabilities. Randomized surveys license inferences about whole populations using data from just a sample, on the grounds that subjects are randomly chosen for the sample with known probabilities. The two techniques, randomized experimentation and random sampling, share deep parallels since, fundamentally, randomized experimentation *is* random sampling from alternative potential outcomes.

The methodological literatures in random sampling and random experimentation are often separated, but lessons in one often apply to the other. For example, blocking in experiments occurs when separate (completely randomized) experiments are conducted in pre-specified demographic cells. Stratification in surveys occurs when fixed numbers of people from pre-specified demographic cells are sampled into the study. The main purpose of both procedures is to decrease sampling variability. By learning about one technique or the other, we learn about both. This point is nicely explained in Judith Tanur's preface to the book, along with pointers to other parallels between sampling and experimental designs.

In an opening essay, the editors of this volume (Paul J. Lavrakas, Courtney Kennedy, Edith de Leeuw, Brady T. West, Allyson L. Holbrook, and Michael W. Traugott) make a strong case that probability surveys should—as a matter of routine—contain some randomized survey experimental element. They use the "validity" framework that contrasts internal with external validity. As the story goes, experiments are internally valid but possibly not externally valid because average causal effects estimated on a sample can't be generalized to a population without a randomized survey design or further assumptions.

Therein lies the pitch for randomized experiments embedded within probability surveys. Randomized experiments allow us to estimate causal effects, and randomized survey designs allow us to generalize those estimates from the sample to the relevant population. Both kinds of randomization offer researchers design-based assurances that their inferences can only be so far off from the truth and that the uncertainty statistics—the standard error estimates, the confidence intervals—correctly characterize how far off our conclusions might actually be.

I wholeheartedly agree that nearly any survey presents a valuable opportunity to conduct survey experiments. That said, I'd like to insert a plea that we move on from "validity" framework when discussing research designs. In my view, the classic internal versus external validity distinction for survey experiments is mostly not helpful. We should focus on *estimands*, not alternative flavors of validity. For survey experiments conducted with online convenience samples, a common estimand is the Sample Average Treatment Effect, or SATE. If the survey experiment is designed and analyzed well, then estimates of the SATE will, on average, be close to the SATE. This is "internal validity," or a claim about the unbiasedness or consistency of the full set of procedures that lead to an estimate. Of course, a SATE on an online convenience sample might not be the same as a SATE in a probability sample of Minnesotans. A SATE estimate is said to lack external validity when it is not equal to a particular Population Average Treatment Effect (PATE). There are many possible populations, so there are many possible PATEs: the ATE among all Americans in 2020 or the ATE among world citizens in 1983. Since these PATEs could easily be different, *any* SATE will always be "externally invalid" for some PATE or other, regardless of whether the subjects are drawn at random from a well-defined population or not. By this logic, no design can ever achieve external validity because there may always be some new population to which the results will fail to generalize. Any claim about a study's external validity would be sharper if the target estimand were made explicit.

If I had a criticism to offer of this volume as a whole, it's that most of the articles are not explicit about their estimands. The clear majority of the articles are investigations of the impact of survey design choices on inferences. These include within-household selection procedures, interview mode, whether or not to send advance letters, whether to use simple or complex survey questions, scale direction and alignment decisions, and how to control race- and ethnicity-of-interviewer effects. Of course, since this volume is about the intersection of probability sampling and randomized experimentation, the impacts of the survey design choices on inferences are assessed using RCTs. These articles will be useful for survey technicians who are concerned about each of these subtle design choices. For example, we learn from Susanne Vogl, Jennifer A. Parsons, Linda K. Owens, and Paul J. Lavrakas that sending advance letters substantially increases survey response rates (by approximately 15 percentage points in their application!). This effect is large and very important to designers of surveys. The result raises interesting questions about the estimand for any experiments embedded in such surveys. Do we learn about the PATE from this design? Or is it the conditional average treatment effect (CATE) among those who would respond without an advance letter, the CATE among those who would only respond if sent an advance letter, or a weighted average of the two? Survey researchers and experimenters should be explicit about their estimands and consider the effect of their survey design choices on their ability to estimate them.

Two chapters in this volume stood out to me as being especially good at keeping the focus on estimating estimands: the chapter by Samara Klar and Thomas Leeper on purposive samples and the piece by Elizabeth Tipton, David S. Yeager, Ronaldo Iachan, and Barbara Schneider on effect heterogeneity. Tipton et al. explain that SATE estimates might be biased for the PATE if different units respond to treatments differently and if different units have different probabilities of responding to the survey. Without effect heterogeneity, SATEs and PATEs are equal. Without selection into the sample, SATEs and PATEs are equal, too. In general, the distance between a PATE and a SATE estimate is a complex function of treatment effect heterogeneity and sample selection. Some statistical fixes for this problem are available, but I appreciated how these authors focused on developing theories of effect heterogeneity that inform survey design choices. The Klar and Leeper article is in this same spirit. They focus on the question of estimating effects among tougher-to-reach populations (in particular, those with rare intersections of overlapping group identities). They point out that the typical design goal of estimating a PATE is in tension with the goal of estimating a CATE among a rare group. In this case, it's the PATE that isn't "externally valid" to the CATE under consideration.

doi:10.1093/poq/nfaa053