Introduction to the Special Issue: Innovations and Current Challenges in Experimental Methods

Libby Jenke*

Political science has increasingly embraced the experimental method to establish causal relationships suggested by theories and observational studies – from experiments' traditional subdisciplinary home, political psychology, to international relations. The most frequently voiced concern with experiments, qualms about their external validity in terms of sample, has been well documented and addressed (McDermott, 2011; Krupnikov and Levine, 2014; Coppock et al., 2018; Lupton, 2019; Krupnikov et al., 2021; Mutz, 2021).[1] Experiments uniquely provide scholars with the internal validity necessary to confidently identify causal effects, and issues with specific experiments tend to arise through errors of application by individual scholars rather than through any broad problems with the methodology.

Yet while experiments allow scholars to identify causal effects, how and why treatment effects occur remain questions that challenge experimentalists. The issue is identifying which of many possible mechanisms transmit an effect, and weaknesses exist in current experimental techniques in parsing these mechanisms. But doing so is key to designing real-world policy interventions. For example, if citizens are reluctant to support an African American Supreme Court justice or presidential candidate, it matters whether they do so because of the color of the nominee's skin or because African Americans are assumed to be liberal (McDermott, 1998; Sen, 2017). The two mechanisms suggest different solutions to address the problem. Without knowledge of causal anatomies, it is possible to predict outcomes but not to prescribe policy changes.[2]

*Assistant Professor, Department of Political Science, University of Houston, 3551 Cullen Boulevard, Room 447, Houston, TX 77204-3011, United States. E-mail: ljenke@uh.edu

[1] External validity concerns may also pertain to settings or experimental treatments.

[2] For example, machine learning methods can produce black box models, which predict but do not allow an examination of causal mechanisms (de Marchi and Stewart, 2020).

This special issue has two aims. First, I highlight recently developed methodological tools that confront challenges involved in identifying causal mechanisms. Scholars often use composite treatments that allow for an efficient design but leave them unable to specify which mechanism is responsible for effects. Hainmueller et al. (2014) offered a solution that avoids this tradeoff with conjoint analysis. This work was extended by Jenke et al. (2021), who validated attribute importance as the mechanism behind average marginal component effects (AMCEs) in conjoint analysis using eye-tracking. Eye-tracking allows for the measurement of information accrual and processing, which gives insight into the cognitive models employed behind a choice. Another weakness in testing causal mechanisms is that average causal mediation effects rely on strong assumptions that are often known to be false. Bansak (2020) presented causal mediation estimands that trace the causal anatomies of multiple treatments without the assumption of no unobserved confounding of the mediator-outcome relationship. Acharya et al. (2018) also relaxed this assumption in their approach to mediation analysis, which experimentally manipulates the mediator rather than observing it.

Second, I include papers that underscore rarely-noted issues with experiments, some of which have solutions and others that remain unsolved. Miratrix et al. (2018) and Dafoe et al. (2018) focused on weighting and the information equivalence (IE) assumption, respectively, and provided needed recommendations for best practices. Gaines et al. (2007) concentrated on four issues: the frequency with which subjects are treated in an experiment versus in the real world and the endurance of treatment effects, the possibility of mutual causation, the lack of control groups in many studies, and spillover effects. These issues are straightforward yet vital to experimental validity. Last, Franco et al. (2015) brought needed attention to the under-reporting of experimental results. My goal in highlighting these concerns is not to question past work but instead to increase the future scholarly returns of survey experiments.

1 Innovations

Experimentalists have traditionally faced the issue that the often-composite nature of experiments' treatment conditions leaves scholars uncertain as to the specific component responsible for an experimental effect. Treatments are often multidimensional, consisting of the alteration of many aspects of a prompt rather than a single element, and those that are one-dimensional suffer from confounding effects of other, correlated components. Hainmueller et al. (2014) have offered a solution with conjoint analysis. Conjoint analysis permits the identification of component-specific effects while allowing scholars to test causal hypotheses about multidimensional preferences efficiently, in a single experiment. Conjoint designs are suited for use in a wide swath of substantive areas including voting, public opinion, climate change, and international relations (Bansak et al., 2021).

Conjoint experiments and most experimental studies consist of asking respondents to state their choices, leaving the cognitive processes behind these stated choices unspecified. Jenke et al. (2021) leveraged eye-tracking to provide data on the information-gathering processes of respondents who were completing a conjoint experiment. Previous papers had established AMCEs as indicating the impact of attributes' effects. However, the assumed mechanism behind the causal effect – attribute importance – had not been validated. Eye-tracking allowed the authors to validate the interpretation of AMCEs as the relative importance of components in the decision. They also measured the change in information processing that occurres as conjoint tables become increasingly complex, showing that while the AMCEs remained consistent, respondents transitioned from one processing strategy to another in a manner that is consistent with bounded rationality. Broadly speaking, eye-tracking can be used to inform scholars as to the mechanisms behind causal effects if the mechanisms are consistent with one information processing strategy and not another. With the advent of webcam eye-tracking technologies (Semmelmann and Weigelt, 2018; Xu et al., 2015; Yang and Krajbich, 2020), this method's accessibility is rapidly improving and its application in experimental political science should increase.

Another improvement was made by Bansak (2020). He presented comparative causal mediation (CCM) estimands that allow for the estimation of the effects of multiple treatments (and a single mediator) and comparisons of their magnitudes. The advantage of the approach over others is two-fold. CCM estimands allow the assumption of no mediator-outcome confounders to be relaxed. This is always a concern because even if some of the possible confounders are measured, some may be missed or are unable to be measured. Audience costs provide a good example. Possible mediator-outcome confounders include age, left-right ideological leanings, beliefs on policy issues, and personal connections to the military. Many datasets utilized by international relations scholars do not contain variables for all of these cofounders. Bansak's approach also permits unit-specific parameters rather than assuming constant effects. The assumption of constant effects across all people is an extremely rigid assumption that is known to be false in most areas of study. With these advantages, CCM estimands should be used whenever designs feature multiple treatments.

Acharya et al. (2018) likewise avoided the assumption of no mediator-outcome confounders in their strategy to analyze causal mechanisms. Imai et al. (2011)'s foundational approach provided a framework to identify indirect effects, but their approach requires this assumption. Acharya et al.'s strategy is to manipulate the information environment of a survey experiment such that respondents selectively receive information about the mediator of interest rather than observing it. The difference between the overall average treatment effect and the treatment effect when the mediator is set at a particular value is identified as the role of the mechanism in the treatment effect, either through indirect effects or interactions. Their approach allows scholars to tease apart causal mechanisms without the restrictive assumptions previously employed.

2 Under-Examined Procedures and Issues

In addition to new methodologies, scholars have provided useful recommendations regarding under-examined procedures involved in experiments. One such area is weighting. Scholars publishing survey experiments rarely state their weighting procedures and even more rarely provide a justification for weighting or not weighting (Franco et al., 2017), despite this being precisely what the Standards Committee of the Experimental Research Section of the American Political Science Association suggests (Gerber et al., 2014). Since characteristics that determine selection into survey samples may moderate treatments, it is problematic to treat nonprobability samples as if their sample average treatment effects (SATE) mimic population average treatment effects (PATE). Miratrix et al. (2018) analyzed this problem and provided recommendations for scholars. They recommended that the SATE be calculated without weights. Then, the PATE should be calculated with weights and the two estimates compared. A difference in the estimates should prompt consideration of the quality of the weights and covariates that could explain treatment effect heterogeneity. Regardless of the difference, their analyses suggest that for all survey experiments run on datasets that provide weights to researchers, both the SATE and the PATE should be reported.

Dafoe et al. (2018) provided the first in-depth analysis of the information equivalence (IE) assumption, which is another under-studied area of experimental methodology. This is a crucial assumption in regards to the legitimacy of causal claims in experiments. If the IE assumption does not hold then causal effects cannot be reliably attributed to the variable of interest and may operate through an unintended causal channel. The authors reviewed three solutions to IE violations – abstract encouragement, covariate control, and embedded natural experiments – and evaluated their effectiveness at reducing them using placebo tests. Across several studies, they found that embedded natural experiments best promote information equivalence. In addition to bringing attention to this underlying assumption of experiments and stressing the importance of considering alternative causal channels through which an effect might flow, they specify a useful recommendation of steps that should be taken to minimize IE violations.

Gaines et al. (2007) provided a list of rarely-noted issues with experiments. These include the mismatch of the number of real-world and experimental treatment incidences over time, the causal complexity that can lie behind the assumption of a one-way causal relationship, experiments that do not include a control group, and accidental spillover effects based on the placement of experiments within a survey. Despite the 14 years since the paper's publication, many of these challenges continue to be a concern in the literature, and they are important. For instance, the duration of experimental effects determines the implications of experiments for policies, political behavior, and candidate success in the real world. Some papers have measured how long their effects last and discussed the ensuing implications [see, e.g., Kalla and Broockman (2018); Chong and Druckman (2010); Levendusky (2013); Gerber et al. (2011)], but the number is far fewer than best practices would hope for. This paper deserves to be read by all experimentalists and its suggestions warrant common application.

Another under-noted issue is under-reporting of experimental results, which was shown by Franco et al. (2015) to be occurring at an alarming rate in political science. The authors analyzed experimental studies that won a Time-Sharing Experiments in the Social Sciences (TESS) award between 2002 and 2012 and made their way to publication. Based on the publicly available survey questionnaires (which were posted before data collection), they found that 30% of papers did not report all treatment conditions that were part of the experiment. Additionally, 60% of papers utilized fewer outcome variables than were listed in the questionnaire. This is problematic because if scholars are selectively reporting only results that confirm their hypotheses, then positive results will be overrepresented and lead to bias in the literature.

A solution to this problem has as of yet not been suggested in political science, beyond the rather optimistic hope of the authors that political science will change its culture and norms. Even if pre-registration becomes standard, reviewers will not necessarily compare the reported results to the pre-registration; doing so would significantly increase the time needed to review a piece. In other fields where pre-registration is required in order to publish, such as medicine, it has been found to be ineffective at preventing scholars from excluding negative results from their published studies (Chan et al., 2004; Chan and Altman, 2005). One potential solution is a requirement by journals that authors of pre-registered work explicitly state that they have reported all pre-registered measures, conditions, and hypotheses.[3] This would change under-reporting from an act of omission to active dishonesty about one's work and thus act as a more effective deterrent.

3 Concluding Remarks

Our goal as political scientists is to make discoveries about human behavior in real-world political contexts. To do this, it is necessary to explain outcomes – to establish both causal relationships and the mechanisms behind them – not merely predict outcomes. Predictive models that are not interpretable or do not identify a causal mechanism are of limited utility. Experiments provide the most rigorous way for social scientists to identify causation and, thanks to recent innovations in experimental methods, we have made inroads in creating feasible ways to analyze causal mechanisms.

The following areas of research would be helpful for future work to focus on. First, we need to continue to create ways to identify which causal mechanisms are at play in a causal effect without introducing overly restrictive assumptions. The papers included in this special issue are a start, but more work needs to be done. Second, research should strive to incorporate the recommendations of Miratrix et al. (2018); Dafoe et al. (2018); Gaines et al. (2007). Third, solutions need to be found for the problematic status quo regarding under reporting of experimental results. As an

[3] Thanks to my colleague, Scott Clifford, for this suggestion.

experimentalist, I view these challenges with excitement, as their solutions will allow us to identify causation and make policy recommendations with increasing confidence.

4 About the Author

Libby Jenke is an Assistant Professor of Political Science at the University of Houston. She is an experimental methodologist with expertise in eye-tracking and mouse-tracking. Her papers are available at `http://www.libbyjenke.com`.

References

Acharya, A., M. Blackwell, and M. Sen (2018). Analyzing causal mechanisms in survey experiments. *Political Analysis 26*(4), 357–378.

Bansak, K. (2020). Comparative causal mediation and relaxing the assumption of no mediator–outcome confounding: An application to international law and audience costs. *Political Analysis 28*(2), 222–243.

Bansak, K., J. Hainmueller, D. J. Hopkins, T. Yamamoto, J. N. Druckman, and D. P. Green (2021). Conjoint survey experiments. *Advances in Experimental Political Science*, 19.

Chan, A.-W. and D. G. Altman (2005). Identifying outcome reporting bias in randomised trials on pubmed: review of publications and survey of authors. *Bmj 330*(7494), 753.

Chan, A.-W., A. Hróbjartsson, M. T. Haahr, P. C. Gøtzsche, and D. G. Altman (2004). Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Jama 291*(20), 2457–2465.

Chong, D. and J. N. Druckman (2010). Dynamic public opinion: Communication effects over time. *American Political Science Review*, 663–680.

Coppock, A., T. J. Leeper, and K. J. Mullinix (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences 115*(49), 12441–12446.

Dafoe, A., B. Zhang, and D. Caughey (2018). Information equivalence in survey experiments. *Political Analysis 26*(4), 399–416.

de Marchi, S. and B. Stewart (2020). Wrestling with complexity in computational social science: Theory, estimation and representation. *The SAGE Handbook of Research Methods in Political Science and International Relations*, 289.

Franco, A., N. Malhotra, and G. Simonovits (2015). Underreporting in political science survey experiments: Comparing questionnaires to published results. *Political Analysis*, 306–312.

Franco, A., N. Malhotra, G. Simonovits, L. Zigerell, et al. (2017). Developing standards for post-hoc weighting in population-based survey experiments. *Journal of Experimental Political Science 4*(2), 161–172.

Gaines, B. J., J. H. Kuklinski, and P. J. Quirk (2007). The logic of the survey experiment reexamined. *Political Analysis*, 1–20.

Gerber, A. S., K. Arceneaux, C. Boudreau, C. Dowling, S. D. Hillygus, T. Palfrey, D. R. Biggers, and D. J. Hendry (2014). Reporting guidelines for experimental research: A report from the experimental research section standards committee. *Journal of Experimental Political Science*, 81–98.

Gerber, A. S., J. G. Gimpel, D. P. Green, and D. R. Shaw (2011). How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *American Political Science Review*, 135–150.

Hainmueller, J., D. J. Hopkins, and T. Yamamoto (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis 22*(1), 1–30.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 765–789.

Jenke, L., K. Bansak, J. Hainmueller, and D. Hangartner (2021). Using eye-tracking to understand decision-making in conjoint experiments. *Political Analysis 29*(1), 75–101.

Kalla, J. L. and D. E. Broockman (2018). The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review 112*(1), 148–166.

Krupnikov, Y. and A. S. Levine (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science 1*(1), 59.

Krupnikov, Y., H. H. Nam, and H. Style (2021). Convenience samples in political science experiments. *Advances in Experimental Political Science*, 165.

Levendusky, M. S. (2013). Why do partisan media polarize viewers? *American Journal of Political Science 57*(3), 611–623.

Lupton, D. L. (2019). The external validity of college student subject pools in experimental research: A cross-sample comparison of treatment effect heterogeneity. *Political Analysis 27*(1).

McDermott, M. L. (1998). Race and gender cues in low-information elections. *Political Research Quarterly 51*(4), 895–918.

McDermott, R. (2011). Internal and external validity. *Cambridge handbook of experimental political science*, 27–40.

Miratrix, L. W., J. S. Sekhon, A. G. Theodoridis, and L. F. Campos (2018). Worth weighting? how to think about and use weights in survey experiments. *Political Analysis 26*(3), 275–291.

Mutz, D. C. (2021). Improving experimental treatments in political science. *Advances in Experimental Political Science*, 219.

Semmelmann, K. and S. Weigelt (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods 50*(2), 451–465.

Sen, M. (2017). How political signals affect public support for judicial nominations: Evidence from a conjoint experiment. *Political Research Quarterly 70*(2), 374–393.

Xu, P., K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

Yang, X. and I. Krajbich (2020). Webcam-based online eye-tracking for behavioral research.