PA

The External Validity of College Student Subject Pools in Experimental Research: A Cross-Sample Comparison of Treatment Effect Heterogeneity

Danielle L. Lupton

Assistant Professor of Political Science, Colgate University, 13 Oak Drive, Hamilton, NY 13346, USA.
Email: dlupton@colgate.edu

Abstract

Are student subject experiment pools comparable across institutions? Despite repeated concerns over the "college sophomore problem," many experiment-based studies still rely on student subject pools due to their convenience and accessibility. In this paper, I investigate whether student subject pools are comparable across universities by examining how respondents across three student subject pools at distinct educational institutions perform on the same survey experiment about crisis bargaining between states. I argue that, due to selection biases inherent in university matriculation and the self-selection of students into experimental protocols, respondents across these subject pools will exhibit key demographic differences. I also examine whether respondents across these subject pools think similarly about international politics and respond comparably to experimental treatments. I find that, while there are significant demographic differences across subject pools, subjects across institutions respond similarly to experimental treatments—with the key exception of information regarding the regime type of a state. Furthermore, there is little evidence that these demographic differences impact conditional average treatment effects across subgroups. These findings carry critical implications for the use of student samples across political science and within international relations more specifically, particularly regarding the current replication crisis in the discipline.

Keywords: survey experiments, subject pools, generalizability, student subjects, treatment effects

1 Introduction

One of the largest concerns with the proliferation of experiments across political science and international relations (IR) is the external validity and generalizability of results derived from nonrepresentative samples to broader populations of interest (McDermott 2002, 2011; Barabas and Jerit 2010; Mullinix *et al.* 2015; see Mintz, Yi, and McDermott (2011) for discussion of experiments in IR). While scholars increasingly employ large online subject pools, many studies rely on college student samples due to their convenience and accessibility (see Kam, Wilking, and Zechmeister 2007). Yet, despite concerns over the external and internal validity of student populations, commonly known as the "college sophomore problem" (see Sears 1986), scholars of political science have not thoroughly examined either (1) whether college students across universities are comparable to each other or (2) treatment effect heterogeneity across subject pools within IR specifically.

This paper addresses both these issues. To the former, I test the assumption that college students are a homogeneous group when it comes to studies within political science. To the latter, I examine these questions of generalizability and treatment effect heterogeneity within the context of the IR subfield. While scholars of American politics have examined subject pool comparability

Author's note: Thank you to Matt Luttig and Aila Mattanock for their comments on earlier drafts of this paper. Thank you also to Chris Gelpi, Tim Büthe, and Bill Boettcher for their comments on the survey instrument. This work was supported by funding from Duke University and Colgate University. This research was approved by Institutional Review Boards at Colgate University (#ER-S15-33), Duke University (#B0170), and North Carolina State University (#2999). Replication files are available at Lupton (2018b).

Political Analysis (2019)
vol. 27:90–97
DOI: 10.1017/pan.2018.42

Published
19 October 2018

Corresponding author
Danielle L. Lupton

Edited by
R. Michael Alvarez

and treatment effect heterogeneity (Imai and Strauss 2011; Green and Kern 2012; Krupnikov and Levine 2014; Coppock, Leeper, and Mullinix 2017; Grimmer, Messing, and Westwood 2017), there has been little explicit research on this issue in IR. Thus, my study contributes to the broader methodological literature on the generalizability of sample populations and the heterogeneity of treatment effects across subject groups, and it expands our understanding of these issues to IR more specifically.

To examine these issues, I employ the same scenario survey experiment with random assignment at student subject populations across three distinct educational institutions as well as on Amazon mechanical turk (MTurk). To examine the comparability of these subject pools, I compare the demographics and worldviews of respondents across samples, as well as consider how subjects in each research pool respond to experimental stimuli. My results show there are key differences in the demographics of student subject populations, as well as how individuals think about international politics. While the effects of treatments are largely comparable across these groups, I do find heterogeneous treatment effects across student subject pools for one specific treatment: regime type. This result is important, as regime type (e.g., whether a state is democratic or nondemocratic) is a common experimental treatment in IR. My results also caution that student samples derived from educational institutions without dedicated experimental research pools may experience particularly high dropout rates, undermining the reliability of results derived from these populations. Scholars, therefore, need to consider how the internal recruitment mechanisms for experiments at their institution may affect survey response rates or sample diversity before relying on college students to take their experiments.

2 Questioning The External Validity of Student Subject Populations

Scholars across political science have repeatedly challenged the acceptability of college students as sample populations, due primarily to concerns about the homogeneity and broader generalizability of undergraduate populations (e.g., Mintz, Steven, and Vedlitz 2006; Krupnikov and Levine 2014; see also Hyde 2015).[1] In response, there has been a push toward recruiting participants with actual policy experience, particularly for studies of foreign policy decision making (e.g., Mintz 2004; Mintz, Steven, and Vedlitz 2006; Renshon 2015), or for using broader subject populations, such as MTurk. Yet, the generalizability of results derived from MTurk samples is also hotly debated (Paolacci, Chandler, and Ipeirotis 2010; Berinsky, Huber, and Lenz 2012; Clifford, Jewell, and Waggoner 2015; Coppock, Leeper, and Mullinix 2017; Coppock forthcoming; see also Kosinski *et al.* 2015), and not all researchers have access to current or former elite decision makers. Furthermore, other evidence indicates students behave comparably to nonstudent populations on experimental surveys, suggesting concerns about the generalizability of student samples may be overstated (Druckman and Kam 2011; Falk, Meier, and Zehnder 2013; Krupnikov and Levine 2014; Mullinix *et al.* 2015; Kees *et al.* 2017).

Across these studies, the current assumption, even among critics of the use of undergraduate samples, is that results derived from student samples at one institution are generalizable to other student populations. This assumption, however, has not been explicitly tested. I believe there is good reason to suspect that experimental results derived from college student samples may vary across institutions for three reasons. First, there are selection effects in the admittance and attendance of students at different educational institutions, including based on prestige, selectivity, and geographic location. Second, these self-selection effects may lead to important demographic differences across student populations as well as which students participate in experimental protocols. Third, different institutions may have distinct recruitment mechanisms to incentivize students to take on-campus experiments, affecting which students choose to

1 Current studies only compare a single student subject population to other samples.

participate in studies as well as dropout rates for protocols across institutions (see Kam, Wilking, and Zechmeister 2007). Each of these problems calls into question the external and internal validity of results derived from student subject pools at any one institution.

3 Research Design

I address these issues by examining whether student samples across differing educational institutions are indeed comparable to each other in terms of their demographics, worldviews, and the effect of experimental treatments embedded in the survey. To do so, I administer the same scenario survey experiment to student subject pools across three distinct universities, as well as to an online convenience sample using MTurk consisting of participants over the age of 18.[2] Student sample A is taken from a selective liberal arts college in the Northeast. Student sample B is derived from a private research university in the Mid-Atlantic, and student sample C is taken from a public research university in the Mid-Atlantic.

Across each of these sample populations, participants engage in a scenario survey experiment in which they are asked to make predictions about how the leader of a state would react to a militarized threat during a hypothetical foreign policy crisis. More specifically, they are asked whether they believe the target leader, against whom the threat is directed, would back down (1), stand firm (2), or escalate the crisis (3) in the face of a threat to send troops to a disputed international border. In the survey, I manipulate the information given to participants about the characteristics of the target leader and/or the target leader's state in the scenario prompt. Participants are randomly assigned to different treatment and control groups. In the control group, participants are told nothing about the target leader or state. In the treatment groups, participants can receive information about a leader's behavior in past foreign policy crises (stood firm or backed down during past disputes), information about the state's behavior in past foreign policy crises (stood firm or backed down during past disputes), information about the state's regime type (democracy or nondemocracy), or information about the state's strategic interest in the crisis (high or low).[3]

These treatments were chosen for several reasons. First, past research has shown these factors to be influential to the onset of international conflict and during crisis bargaining.[4] Second, these treatments are used in experiments in IR to explain a variety of outcomes (e.g., Tomz 2007; Lupton 2018c). The impact of regime type, in particular, has been repeatedly studied with experimental methods in IR (e.g., Tomz 2007; Tomz and Weeks 2013). For the intent of this study, however, the contents of the treatments themselves are not of central importance.[5] Rather, my purpose is to determine whether participants across different subject populations respond similarly to treatment stimuli. The survey also includes pretreatment measures to gauge how respondents think about world politics, including their views on the importance of international leaders and on the acceptability of the use of militarized force (Tomz 2007). Both of these may affect participants' responses to the primary dependent variable (i.e., the leader's response to a threat) as well as the effect of individual treatments. The survey also asks respondents a variety of demographic questions, including their political affiliation, gender, interest in and attention to

2 The MTurk sample was restricted to users within the United States with a 95% or higher prior approval rating within the platform (see Paolacci, Chandler, and Ipeirotis 2010; Berinsky, Huber, and Lenz 2012). The experiment was fielded at Institution A in 2016 and at Institutions B and C and on MTurk in 2013. These institutions were included as they were accessible to the researcher. The scenario text does not mention a specific international conflict, nor does it mimic a particular on-going conflict between any two states. Replication files are available at Lupton (2018b).

3 The full text of the experiment is available in the supplementary files. Additional treatment groups receive information about a leader's past behavior and one additional factor (that is, past state behavior, regime type, or state interest), allowing the researcher to parse out the distinction between leader traits and state traits if desired (see Lupton 2018a,c).

4 The literature tying these variables to the initiation and conduct of international conflict is vast.

5 I encourage future research to use employ other experiments across multiple student subject pools and engage in a similar comparative analysis.

politics, educational level, and age. Finally, the experimental design itself is representative of many experiments in IR, where participants are given hypothetical foreign policy scenarios. Thus, the survey experiment is useful for answering questions about cross-campus comparability of subject pools as it represents a more typical, rather than highly specialized, experiment in IR.

It is important to note that there were key differences in recruitment protocols across each of the three student subject pools. While institutions B and C had dedicated political science research pools, institution A did not. Accordingly, students at institutions B and C received class credit for completing the experiment, while students at institution A took the survey out of personal interest without any direct benefit. Furthermore, students in subject pool B received extra credit for their participation, while students in subject pool C were institutionally required to take a minimum number of protocols during the semester. The recruitment pools at institutions A and B consisted of students across a variety of introductory political science courses, while the pool at institution C consisted of students enrolled in a single large introductory political science course. In the next section, I consider how these recruitment mechanisms may explain differences in dropout rates or the number of participants from each institution.[6]

4 Results: Differences Across Student Subject Pools

I find there are key differences across student subject populations regarding the demographics of participants as well as how they think about international politics.[7] Most notably, students from institution A are significantly more interested in international politics and pay more attention to international events when compared to all other subject populations. This makes intuitive sense given that students at institution A self-selected into the experimental protocol without receiving any course credit and were drawn from political science and IR majors at the institution. This sample also had the largest dropout rate (over 34%), suggesting that students who fully completed the experimental protocol were most personally interested in doing so. Students across subject populations also vary in their views on the role of leaders in world politics and the acceptability of the use of militarized force, with participants from institution A being most opposed to the use of force.

I next consider whether treatments have a similar effect on participants' decision making across each student subject pool. In particular, I examine how information regarding state regime type, strategic interest in the dispute, past state behavior, and past leader behavior influence participants' predictions about how the opposing leader would react to a militarized threat and whether these effects are comparable across subject populations. Again, my focus is not on the causal impact such treatments may have on participants' predictions regarding the leader's threat response. Rather, my intention is to determine whether these treatment effects are similar (or homogeneous) across different student samples. Figure 1 presents the average effect of different treatments broken down by subject pool.

As Figure 1 shows, the treatment effects of regime type do appear to exert divergent substantive effects across student subject populations at different institutions.[8] Here, information regarding the regime type of a leader's state has heterogeneous effects on subjects at institution B versus institution C. While students from institution B think democratic leaders will be significantly more likely to "back down to the threat" ($p = 0.036$), students from institution C predict democratic leaders will be significantly more likely to "escalate" the crisis ($p = 0.006$).[9] In terms of the dependent variable, 25.0% of students from institution B assigned to the democracy condition predict democratic leaders will back down and none believe democratic leaders will escalate

6 Participants in the MTurk pool were paid $0.85 for their participation in the survey.
7 See Tables A.1 and A.5 in the supplementary files.
8 Regression analyses are available in Tables A.2 and A.3 in the supplementary files.
9 See Table A.2 in the supplementary files.

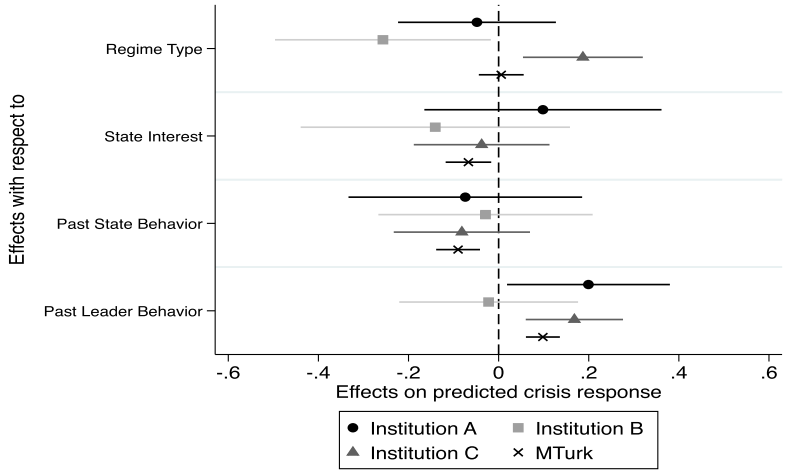

Figure 1. Average treatment effects by sample population.

the crisis. In contrast, no students from institution C assigned to the democracy condition think democratic leaders will back down and 55.6% predict democratic leaders will escalate the conflict. Furthermore, the difference between the means across these subject populations is statistically significant ($p = 0.000$).[10] However, the average treatment effect of regime type is not statistically significant for subjects at institution A or for participants in the MTurk sample.

In contrast, participants from institution A ($p = 0.031$), institution C ($p = 0.002$), and the MTurk sample ($p = 0.000$) are significantly more likely to predict that leaders with a history of responding firmly to past disputes will "escalate" the current crisis; yet, this condition is not statistically significant across participants at institution B. There is also evidence to suggest that the levels of significance of treatment effects are smallest (that is, most statistically significant) for results drawn from the MTurk pool. Here, the data reveals that the treatments for state strategic interest and past state behavior are highly statistically significant for MTurk respondents ($p < 0.01$ and $p < 0.001$ respectively), but not across any of the student subject populations. The level of significance here is likely driven by the larger sample size of the MTurk pool compared to the student subject pools.

Cross-sample comparisons of difference by treatment group, however, reveal that the substantive effects of state strategic interest and past state behavior, as well as that of past leader behavior, remain statistically comparable across subject groups. Thus, it appears that student sample populations react similarly to these different experimental stimuli. In fact, I find there are only significant variations in cross-sample treatment effects for the impact of regime type. Here, there is a statistically significant difference in the effect of this treatment when comparing the responses of student subjects at institution A to institution C ($p = 0.012$) and when comparing the responses of students at institution B to institution C ($p = 0.000$) or to the MTurk sample ($p = 0.002$). Overall, however, the results suggest that students from different institutional subject pools generally respond similarly to treatment stimuli. In other words, I find little evidence for heterogeneous treatment effects across student samples, with the exception of regime type.

I also consider whether the effect size of these treatments is comparable across subject groups (see Krupnikov and Levine 2014). Here, I use Cohen's d as a standardized measure of relative effect size (Cumming 2012), with $|d| > 0.80$ indicating a large substantive difference in effect size across two groups. I find that, across all treatment groups and all subject group comparisons, there is only one instance in which treatment effect size across groups is substantively different—when comparing the impact of the regime type treatment on subject populations from institution B to

10 See Tables A.4 and A.8 in the supplementary files.

institution C ($|d| = 0.904$).[11] Overall, therefore, I find little evidence to support the notion that treatments have substantively different effects across student subject populations drawn from distinct educational institutions.

Finally, I directly test for the heterogeneity of treatment effects within subpopulations across samples by examining conditional average treatment effects (CATEs) across subgroups. I test for CATEs based on key factors that may moderate these treatment effects across individuals, including political affiliation, partisanship, gender, interest in international politics, age, and views on the acceptability of the use of militarized force, using a standard regression analysis with robust standard errors in which each factor is interacted with each treatment (Freedman 2008).[12] I find evidence that two factors may significantly condition the effects of the leader past behavior treatment: political ideology and participants' views on the acceptability of the use of force ($p < 0.05$). Here, the treatment effects of a leader's past behavior are reduced for participants who identify as having a liberal political ideology and for participants who disapprove of the use of force. However, the substantive impact of this reduction is small for both factors (a reduction of 0.037 points on a three point scale or, alternatively, a 1.76% reduction in the dependent variable in each case). Furthermore, I find little evidence to indicate there are heterogeneous treatment effects across other subgroups, including those based on age, interest in politics, political affiliation, or gender.[13] Overall, therefore, these demographic factors do not appear to consistently condition the impact of treatments in the experiment.

5 Conclusions

This study reveals that participants across student subject pools have different demographic distributions and views on international politics, but that students at different institutions largely respond similarly to treatment stimuli. Thus, the issue of whether student subject pools constitute acceptable research samples may vary based on the broader questions scholars want to answer with their experiments. My study provides evidence that student subject populations may not be comparable in their demographic compositions across institutions, particularly regarding political affiliation.[14] While this may not be a concern for many studies in political science (see Mutz and Pemantle 2015, 8), there are some studies where this may be problematic. Researchers who have reason to believe that their theoretical causal mechanisms may be affected by factors such as political affiliation (for example) may want to be especially careful about using student subject pools or at least consider how the demographics or culture of their educational institution may influence their empirical findings (see Druckman and Kam 2011).

Yet, despite these differences in the demographics of participants and the views respondents hold about international politics across student subject populations, I find little evidence that such factors condition treatment effects within the experiment. Furthermore, my results show that students from different institutions respond to treatment stimuli in comparable ways—with the key exception of the effects of regime type. This may suggest that researchers who use experiments to study questions related to regime type, such as audience costs, need to be cautious in using college students in their studies. This result, however, warrants further research, such as replicating past experiments on audience costs across multiple student subject

11 See Table A.6 in the supplementary files.
12 See Table A.7 in the supplementary files.
13 This finding regarding gender is particularly interesting as past research suggests that women may view conflict differently than men. My result may be due to the consistent gender of the opponent in the scenario prompt (male). Future work should examine this issue further.
14 IRB restrictions at one institution prohibited the collection of information regarding race, ethnicity, or income. As a result, these factors were not gathered at other institutions and are not included in this study. Enrollment statistics report the following percentage of students that identify as Black or African American: A(5.3%), B(11%), C(20%). Future research should examine the influence of race, ethnicity, and income, as well as other factors such as military service, across subject populations.

populations. Future research should also consider whether different student subject populations respond similarly to other treatments or with alternative protocols like lab-based experiments. Furthermore, and with regards to the on-going replication crisis across political science, my results may suggest such replication issues may be due less to experimental design and more to the sample populations researchers use. Thus, scholars need to further consider how not only the type of subject population they use may influence their results in comparison to other sample populations (e.g., students vs. MTurk), but also how generalizable their sample is compared to other samples of the same type (e.g., students at one institution vs. students at another institution).

There is one additional consideration that scholars should also take into account before deciding whether the use of college student samples are appropriate for their research: the recruitment mechanisms for experiments at their educational institution. My study finds that the one student sample that was not derived from a dedicated political science research pool had a high dropout rate, which may pose a threat to the external and internal validity of results derived from this sample. Conversely, the two student subject populations from institutions with dedicated research pools for political science had dropout rates comparable to the MTurk sample. Researchers who do not have access to dedicated experimental research pools need to be especially mindful about the dropout rates in their experimental protocols and consider whether students who actually complete the experiment differ in nonrandom ways from those who begin the experiment but fail to finish it. This may also be a critical concern for researchers employing panel wave studies, who may need to be particularly cautious about using college students if they do not have a dedicated experimental research pool at their institution.

Thus, my results do not suggest that scholars should abandon the use of college students as subject populations, nor do my findings condone their unconditional use for all experimental protocols. Scholars at educational institutions with dedicated research pools from which they can recruit participants and at institutions with diverse student bodies may have fewer concerns about the internal and external validity of their experimental results when using student subject pools. In contrast, scholars whose theoretical causal mechanisms are linked to factors that may vary in nonrandom ways based on the composition of their institution's student body or who do not have access to dedicated research pools need to be particularly conscientious about how these factors may influence the generalizability of their results. In short, scholars need to think carefully about how the demographics of their student bodies and the self-selection of students into experimental protocols at their institution may affect their experimental results if they choose to use college students as their primary sample population.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2018.42.

References

Barabas, Jason, and Jennifer Jerit. 2010. Are survey experiments externally valid? *American Political Science Review* 104(2):226–242.

Berinsky, Adam, Gregory Huber, and Gabriel Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis* 20(3):351–368.

Clifford, Scot, Ryan Jewell, and Philip Waggoner. 2015. Are samples drawn from mechanical turk valid for research on political ideology? *Research and Politics* 2(4): 2053168015622072.

Coppock, Alexander. Forthcoming. Generalizing from survey experiments conducted on Amazon mechanical turk: A replication approach. *Political Science Research and Methods*. Available at https://doi.org/10.1017/psrm.2018.10.

Coppock, Alexander, Thomas Leeper, and Kevin Mullinix. 2017. The generalizability of heterogeneous treatment effect estimates across samples. Working Paper. Available at https://acoppock.github.io/projectpages_CLM_heterogeneity.html.

Cumming, Geoff. 2012. *Understanding the New Statistics*. New York: Routledge.
Druckman, James, and Cindy Kam. 2011. Students as experimental participants: A defense of the narrow base. In *Cambridge Handbook of Experimental Political Science*, ed. James Druckman, Donald Green, James Kuklinski, and Arthur Lupia. Cambridge: Cambridge University Press, pp. 41–57.
Falk, Armin, Stephan Meier, and Christian Zehnder. 2013. Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association* 11(4):839–852.
Freedman, David. 2008. On regression adjustments to experimental data. *Advances in Applied Mathematics* 40:180–193.
Green, Donald, and Holger Kern. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* 76(3):491–511.
Grimmer, Justin, Solomon Messing, and Sean Westwood. 2017. Estimating heterogeneous treatment effects and the effects of heterogeneous treatment effects with ensemble methods. *Political Analysis* 25:413–434.
Hyde, Susan. 2015. Experiments in international relations: Lab, survey, and field. *Annual Review of Political Science* 18:403–424.
Imai, Kosuke, and Aaron Strauss. 2011. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis* 19:1–19.
Kam, Cindy, Jennifer Wilking, and Elizabeth Zechmeister. 2007. Beyond the 'narrow data base': Another convenience sample for experimental research. *Political Behavior* 29(4):415–440.
Kees, Jeremy, Christopher Berry, Scot Burton, and Kim Sheehan. 2017. An analysis of data quality: Professional panels, student subject pools, and Amazon's mechanical turk. *Journal of Advertising* 46(1):141–155.
Kosinski, Michal, Sandra Matz, Samuel Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70(6):543–556.
Krupnikov, Yanna, and Adam Seth Levine. 2014. Cross-sample comparisons and external validity. *Journal of Experimental Political Science* 1(1):59–80.
Lupton, Danielle. 2018a. Reexamining reputation for resolve: Leaders, states, and the onset of international crises. *Journal of Global Security Studies* 3(2):198–216.
Lupton, Danielle. 2018b. Replication data for: The external validity of college student subject pools in experimental research: A cross-sample comparison of treatment effect heterogeneity, https://doi.org/10.7910/DVN/U0TYY3, Harvard Dataverse, V1, UNF:6:w4bmPwNI9Q8PJmgXvWMVJQ==.
Lupton, Danielle. 2018c. Signaling resolve: Leaders, reputations, and the importance of early interactions. *International Interactions* 44(1):59–87.
McDermott, Rose. 2002. Experimental methods in political science. *Annual Review of Political Science* 5:31–61.
McDermott, Rose. 2011. Internal and external validity. In *Cambridge Handbook of Experimental Political Science*, ed. James Druckman, Donald Green, James Kuklinski, and Arthur Lupia. Cambridge: Cambridge University Press, pp. 27–40.
Mintz, Alex. 2004. Foreign policy decision making in familiar and unfamiliar settings: An experimental study of high-ranking military officers. *Journal of Conflict Resolution* 48(1):91–104.
Mintz, Alex, Steven Redd, and Arnold Vedlitz. 2006. Can we generalize from student experiments to the real world in political science, military affairs, and international relations? *Journal of Conflict Resolution* 50(5):757–776.
Mintz, Alex, Yi Yang, and Rose McDermott. 2011. Experimental approaches to international relations. *International Studies Quarterly* 55(2):493–501.
Mullinix, Kevin, Thomas Leeper, James Druckman, and Jeremy Freese. 2015. The generalizability of survey experiments. *Journal of Experimental Political Science* 2(2):109–138.
Mutz, Diana, and Robin Pemantle. 2015. Standards for experimental research: Encouraging a better understanding of experimental methods. *Journal of Experimental Political Science* 2(2):192–215.
Paolacci, Gabriele, Jesse Chandler, and Panagiotis Ipeirotis. 2010. Running experiments on Amazon mechanical turk. *Judgment and Decision Making* 5(5):411–419.
Renshon, Jonathan. 2015. Losing face and sinking costs: Experimental evidence on the judgment of political and military leaders. *International Organization* 69(3):659–695.
Sears, David. 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(3):515–530.
Tomz, Michael. 2007. Domestic audience costs in international relations: An experimental approach. *International Organization* 61(4):821–840.
Tomz, Michael, and Jessica Weeks. 2013. Public opinion and the democratic peace. *American Political Science Review* 107(4):849–865.