Journal of Experimental Political Science **1** (2014) 59–80
doi:10.1017/xps.2014.7

Cross-Sample Comparisons and External Validity

Yanna Krupnikov* **and Adam Seth Levine†**

Abstract

Experimentation is an increasingly popular method among political scientists. While experiments are highly advantageous for creating internally valid conclusions, they are often criticized for being low on external validity. Critical to questions of external validity are the types of subjects who participate in a given experiment, with scholars typically arguing that samples of adults are more externally valid then student samples. Despite the vociferousness of such arguments, these claims have received little empirical treatment. In this paper we empirically test for key differences between student and adult samples by conducting four parallel experiments on each of the three samples commonly used by political scientists. We find that our student and diverse, national adult sample behave consistently and in line with theoretical predictions once relevant moderators are taken into account. The same is not true for our adult convenience sample.

Keywords: Experiments, external validity, subjects, sample comparisons.

INTRODUCTION

Experiments have become increasingly popular in political science (Druckman et al. 2006; Mutz 2011). Although they provide unparalleled ability to identify whether one factor causes another, scholars have frequently voiced concerns about the potential limitations of experimental research. Foremost among these concerns are questions about external validity, or the "extent to which conclusions [of a given study] can be applied across different populations or situations" (McDermott 2011, 34). This point is critical, as scholars aim to use experimental work to answer

Many thanks for advice and feedback from seminar participants at the Vanderbilt Conference on Lab Experiments, Center for Political Studies at the University of Michigan, and the Vincent and Elinor Ostrom Workshop at Indiana University. We also thank Adam Berinsky, Jamie Druckman, Stanley Feldman, Cindy Kam, Spencer Piston, Brendan Nyhan, and our anonymous reviewers for extremely helpful advice. Financial support was generously provided by the Political Science Department at Indiana University and the Center for the Study of Democratic Institutions at Vanderbilt University.
*Department of Political Science, Stony Brook University, Stony Brook, NY 11794, USA;
email: yanna.krupnikov@stonybrook.edu
†Department of Government, Cornell University, Ithaca, NY 14850, USA;
email: ASL22@cornell.edu

questions of import to political science as a discipline and politics at large (Kinder and Palfrey 1993).

Although the idea of external validity broadly applies to the replicability of an experimental finding across various contexts, the characteristics of the subjects participating in a given experiment have emerged as especially critical factors (McDermott 2011). Discussion about subject characteristics has largely focused on implicit or explicit comparisons between three types of samples: undergraduate student samples, adult convenience samples, and national adult samples. Indeed, McDermott (2002:334) has gone so far as to label these concerns a "near obsession" among critics of experiments. Long-term concerns about student samples stem from the idea that students may be so sufficiently different from "everyday citizens" that generalizations are impossible in most cases (Brady 2000; Benz and Meier 2008; Gerber and Green 2008; Sears 1986).

In response to these worries, many scholars have shifted focus to samples of adults (McGraw 2011), with a more recent turn toward adult convenience samples obtained via Amazon's Mechanical Turk (MTurk). MTurk's popularity rests in its relatively low cost and high accessibility for most scholars. Despite the fact that recent research suggests that MTurk produces results that replicate canonical experiments in both psychology and political science (Berinsky et al. 2012), much of the recent discussion about the use of MTurk samples has cautioned against excessive use of this medium for subject recruitment. Concerns have focused on the fact that, despite countless studies, scholars know very little about these individuals. It is unclear *why* they have opted in to MTurk, why they are willing to participate in countless studies for – in some cases – mere pennies, and how their motivation affects the quality of their work.[1]

Underlying concerns about experimental subjects are thus two related questions. First, under what conditions can the results of studies with undergraduate samples and adult convenience samples such as MTurk be replicated with samples drawn from different populations? Second, from an external validity point-of-view, is there an inherent benefit to relying upon samples of "everyday citizens" who are not students, even if they are convenience samples?

To answer these questions with the highest degree of internal validity, it is necessary to compare the results from identical experiments conducted with different samples at approximately the same time. Our paper is one of the first to do just that, and in particular the first political science paper to compare results from common types of undergraduate, adult convenience, and adult national samples.

Although there are an unlimited number of different studies that we could have conducted across these three samples, we narrowed our focus based on a theoretically-relevant dimension: the fact that in some cases experimenters have reason to expect a heterogeneous treatment effect whereas in others cases they

[1] Many recent criticisms of MTurk have appeared not in academic journals, but in various blogs covering topics related to political science and public opinion.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

do not. We thus chose four experiments that allow us to capture common and theoretically-meaningful potential moderators. These moderators reflect factors that might affect how different types of people process political information in general (such as their party identification) as well as how they might process political information when they know they are in an experiment (such as their previous experience taking experiments). Both are common attributes of political science experiments, and thus good places to start for investigating key questions about external validity.

We conducted each of the four experiments with an undergraduate sample recruited via a university subject pool, an adult convenience sample recruited via Amazon's Mechanical Turk, and a diverse national sample,[2] for a total of 12 separate studies. Hewing to the questions above, our key goal in each study is to compare the results across all samples as well as to long-held theoretical predictions. While previous work has offered arguments about various samples (Sears 1986), analyzed the role of sample characteristics in already-existing research (Morton and Williams 2010), relied on simulations to trace the effects of sample differences (Druckman and Kam 2011), considered mode effects (Barabas and Jerit 2010), or focused on the characteristics of one sample type (Berinsky et al. 2012), our paper is the first to use original experiments deliberately designed to compare these three different samples in parallel.

Our investigation is timely for several reasons. First, scholars have begun to think more critically about how the experimental process affects the types of conclusions drawn from experiments (Barabas and Jerit 2010; Druckman and Leeper 2012). Our paper contributes to this growing body of work. Second, our research comes at a time when concerns about student samples continue to fuel enthusiasm for experimental research that relies on "adult" (i.e. non-student) subjects.[3] Although experimentalists are careful to note that a sample's usefulness depends upon having variance on relevant moderators (Druckman and Kam 2011), it remains the case that the large majority of experimental studies published in the discipline's top three journals do not rely on student samples (see Web Appendix A). Third, we have seen the recent expansion of local convenience samples of adults (Kam et al. 2007) as well as adult convenience samples – such as MTurk – recruited over the Internet (Berinsky et al. 2012; Iyengar 2011), but scholars have only begun to identify the conditions under which such samples are substitutable for more nationally-diverse ones.

CONCERNS ABOUT SAMPLE CHARACTERISTICS

The goal of an experiment is to compare subjects who have been randomly assigned to different stimuli. In its most basic form, one group receives one version of

[2] As we make clear later on, we do not have strict probability samples, as we rely on weighted-convenience samples collected by YouGov.

[3] Although we recognize that most undergraduates are legally adults, we use the terms "students" and "adults" for presentational purposes.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

a stimulus and another group receives a different version. Post exposure, group differences in response are used as evidence of a stimulus effect (Druckman and Leeper 2012). Worries about samples, then, hinge on the possibility that these differences might reflect attributes of the subjects included in a particular study rather than the power of the treatment to leave a broader political footprint across time and place.

When might this outcome arise? One way in which a sample might be considered "narrow" is if the subjects differ in theoretically-relevant ways from populations to which experimenters may wish to generalize. Writing about undergraduates, for example, Sears (1986:522) argues that they "are quite uncertain about many of their values, preferences, abilities, and emotions, and for good reason. Many of these dispositions are still developing." Indeed, these considerations have frequently motivated how scholars choose their samples (see Web Appendix B for examples). With adult convenience samples like MTurk, there are also several possibilities. Because MTurk relies on people who are often willing to take political experiments for little compensation, it is possible that they have levels of knowledge, experience, or partisanship that vary from the general citizenry (although, notably, Berinsky et al. 2012 find few such differences between their MTurk and nationally-representative samples). As a result, forming inferences based on these convenience samples may lead scholars to overstate or understate the power of a given stimulus despite the fact that experiment participants are technically adults (Shadish et al. 2002).

A sample's narrowness might also arise due to attributes of the subject pool. Given their varying rules and structure, in some subject pools it is likely that participants will have participated in numerous previous studies. As a result, they would have gained experience being debriefed and learning about the experimental process (including the types of manipulations that experimenters frequently use). In short, they may have become *savvier*.[4] Such savviness can affect subsequent experimental behavior. Kam (2007), for example, notes that subjects who are savvy can deliberately control their responses to such an extent that they undermine measures designed to test implicit beliefs about various topics. More generally, savviness can increase vigilance, suspicion of experimental studies, and a desire to search for the "twist" rather than taking the experimenter's word about the goals of the study (Cook et al. 1970). When subjects are savvy, it seems reasonable to suspect that considerations will be brought to mind that would not arise in the midst of the hustle-and-bustle of everyday life. Put another way, these savvier subjects heighten the artificiality of the experimental setting (McDermott 2011).

At its core, savviness reflects the number of studies that a subject has taken and any requirements of the subject pool to which they belong (Dalen et al. 2001;

[4]There are many other reasons why subjects might differ as well, such as the artificiality of the lab or survey setting (Barabas and Jerit 2010), contamination from previous real-world experiences (Gaines et al. 2007, Transue et al. 2009), or the college environment, which has been shown to affect measures of racial prejudice (Henry 2008).

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Steffens 2004). While we may assume that undergraduate students are more likely to have taken numerous studies as part of subject pools and that members of opt-in platforms such as MTurk may also engage in unlimited participation, increasingly adult subjects who are part of platforms which purport to offer representative samples are also participating in many studies. Indeed, political scientists commonly rely on YouGov and the GfK Group – services that maintain large national subject pools of adults (e.g. Banks and Valentino 2012; Brooks and Geer 2007; Hopkins and King 2010). At the same time, these adult subject pools do not have the same requirements as undergraduate pools. Undergraduate subject pools have a stronger educational norm (Brody et al. 2000) and thus require that researchers conclude studies by offering educational materials that explain the goals of each experiment, a requirement that speeds up learning about the experimental process. Thus, we expect that all participants in subject pools will become savvier as they take more studies, but also that the savviness gradient should be much steeper for those in undergraduate pools.

Typically concerns about experiments arise due to undergraduates, yet we have presented arguments as to why they could also apply to adult samples. We do not mean to suggest that our four experiments capture every possible source of narrowness. Rather, our goal in this paper is to investigate dimensions which, arguably, have formed some of the most ardent criticisms of the experimental method in political science.

Scope of analysis

As with the degrees of variation related to experimental subjects, from a substantive point of view there is also an infinite number of experimental designs that we could use to examine sample differences. After considering the full range, we chose to focus on framing studies (Chong and Druckman 2007; Kinder 2003; Nelson et al. 2011). Framing is ideal for our purposes for three reasons: it is a prominent line of inquiry in political behavior research, it has long-standing theoretical expectations, and it has frequently been the subject of experimentation (Chong and Druckman 2010). Our four framing studies reflect different ways in which results may or may not reflect the narrowness of the subject pool. Our first study has no predicted heterogeneous treatment effect. It just includes a single question that asks people to report on their political news consumption and varies the set of response options. Our second study requires subjects to read a news article and report their opinion, and we expect heterogeneity based on respondents' partisanship. Our third and fourth studies do not require much reading, but might be affected by a subject's degree of savviness.

RESEARCH DESIGN

The undergraduate sample was recruited through a subject pool based in a political science department at a large public institution in the Midwest. The adult

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Table 1
Overview of Studies

Studies	Subject tasks	Samples	N
Studies 1-3: Question-Wording	Answer question	Student	218
		MTurk	301
		YouGov	196
Studies 4-6: Airline Ownership Policy	Read article	Student	96
	Answer question	MTurk	201
		YouGov	181
Studies 7-9: Death Penalty	Follow precise instructions	Student	212
	Read message	MTurk	309
	Answer question	YouGov	292
Studies 10-12: Electoral College	Bogus pipeline	Student	107
	Read message	MTurk	155
	Answer question	YouGov	151

convenience sample was recruited through MTurk following procedures similar to Berinsky et al. (2012). We focus on MTurk as the source of our adult convenience sample because, as Berinsky et al. note, it is becoming a popular method for researchers who wish to use adults that (a) presumably do not have the pitfalls of undergraduates that Sears (1986) highlighted, and (b) are not as expensive as nationally-representative samples.[5] Lastly, our diverse, adult national sample was recruited via YouGov, and our analyses of this sample are weighted to be nationally-representative. YouGov is widely used, including some high-profile cases in which a nationally-representative sample is desired (such as the Cooperative Congressional Election Study and the Cooperative Campaign Analysis Project).[6] Table 1 contains a full summary of our experiments. For each set of experiments, we will compare the results from each sample to existing theoretical expectations and also compare estimates across samples.

As much as possible, we strove to have studies with similar sample sizes to ensure approximately equivalent statistical power. Yet, we acknowledge that in some cases study logistics dictated otherwise. To ensure that sample differences do not affect our ability to discern group differences, we rely on *a priori* power analyses (Levine and Ensom 2001).[7] As three of four experiments are replications of previous studies, we use the effect sizes observed in these previous studies as a baseline and then calculate

[5] As Berinsky et al. note, respondent pools in MTurk are more representative of the adult population than typical in-person convenience samples. They also do not wildly diverge from nationally-representative samples on most measures. Several published, peer-reviewed articles leverage MTurk for all or part of their analysis (e.g. Arceneaux 2012; Gerber et al. 2011; Huber and Paris 2013).

[6] YouGov maintains a panel of over one million participants and uses matching and sampling techniques to approximate a nationally-representative sample (see Vavreck and Rivers 2008 for more detailed information on these techniques). The response rate (RR3) for our particular YouGov study was 41.2%.

[7] Following Barabas and Jerit (2010), however, we also conduct an auxiliary power analysis in cases where we obtain null results.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

the sample size necessary to observe significant group differences at that effect size with power between 0.8 and 0.9 and $0.05 \leq \alpha \leq 0.10$.[8] Our samples meet these thresholds in all but one case, even when we analyze the results by theoretically-important covariates. To the extent that in some samples we do not see significant group differences, this result is thus not a function of sample size differences.

Finally, in all cases we conducted randomization checks using factors measured prior to treatment. Results show that randomization was successful in all studies.

Demographic comparison

We begin by comparing the demographic characteristics of our samples (Table 2). As a benchmark for comparison we also present data from the 2008 American National Election Study (ANES). In addition, we compare our undergraduate and MTurk samples to other experiments that relied on the same subject types. We compare our undergraduates to the student subjects in Taber and Lodge (2006)[9] and we compare our MTurk participants to Berinsky et al. (2012). We do so to ensure that there is nothing unusual about the particular samples we have recruited and that results are based on "typical" samples.

Table 2 points to a couple of notable comparisons. First, as expected, the YouGov sample comes closest to the ANES, though the MTurk sample performs better than the undergraduate sample on a number of demographic factors. Particularly notable are the income comparisons, which show that our undergraduate sample is much wealthier than either MTurk or YouGov. The MTurk sample, however, is distinctly younger and better educated than YouGov and ANES and, notably, nearly 20% of the sample report that they are current undergraduate students – an important figure for scholars who opt to rely on MTurk as an adult sample.[10] Even more importantly, however, the MTurk sample is distinctly more Democratic – in fact, less than 15% of these subjects report that they identify as Republicans. The distribution of partisanship in our undergraduate sample actually comes closer to

[8]The calculations for the full samples are as follows. In the case of our question-wording experiment (studies 1–3), which replicates Schwarz et al. (1985), a N≥90 would provide power of 0.8 and N≥124 would provide a power of 0.9 (Effect size calculated using χ^2 values and sample size in Schwarz et al. 1985). Our death penalty (studies 7–9) and Electoral College studies (studies 10-12) follow from treatments in Mutz (1992), and using the original results to calculate effect size shows that N≥120 would produce a statistically significant effect with power of 0.8 and N≥164 would give us power of 0.9 (Effect size calculated using overall sample size, group difference, and F statistic; this is equivalent for both the death penalty and Electoral College cues, and our samples exceed it with the exception of our undergraduate Electoral College sample that is just shy of N=120). Our airline treatment (studies 4–6) is the only study which, although its structure is similar to canonical framing studies, does not replicate any particular previous experiment. Relying upon standard effect size thresholds (Maxwell and Delaney 2004; see Web Appendix C for details on effect size calculation), we calculated that N≥96 would allow us to observe even small differences with power of 0.8 and N≥210 with power of 0.9.

[9]Patterns are a function of combining Study 1 and Study 2. Statistics not shown in Table 2 were not reported in Taber and Lodge (2006).

[10]In our results we do not omit this 20% of respondents because, at least so far, it is not standard practice to do so when using MTurk. Our results remain the same if we do omit them.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Table 2
Demographics and Experimental Experience Across Samples

	Student Krupnikov & Levine	Student Taber & Lodge (2006)	MTurk Krupnikov & Levine	MTurk Berinsky et al. (2012)	YouGov	2008 ANES
Partisanship						
% Dem	46.1%	47.7%	53.6%	40.8%	35.1%	36.11%
% Rep	31.2%	20.9%	13.1%	16.9%	26.5%	24.04%
Age (avg)	20.0	—	33.5	32.3	49.5	48.7
Race						
% White	83.6%	55.5%	76.5%	83.5%	70.1%	75.3%
% Black	2.1%	—	8.9%	4.4%	11.3%	15.8%
% Female	44.3%	51.5%	55.0%	60.1%	51.4%	56.0%
Income						
% $60,000+	77.7%	—	39.02%	*	36.3%	47.2%
High Deg.						
% H.S.	—	—	9.9%	**	41.37%	31.4%
% B.A.	—	—	31.2%	**	15.88%	13.63%
% Current College Stdnt.	100%	100%	18.8%	—	—	—
Avg. Num Studies	5.98	—	37.2	***	38.8	—

** Berinsky et al. (2012) report the mean income ($55,332) and the median income ($45,000).*
*** Berinsky et al. (2012) report the average years of education (14.9), which suggests their sample is consistent with our sample.*
**** Berinsky et al. (2012) report the average number of studies* about politics *that MTurk participants have taken, which makes it difficult to compare our measure to their results.*

the distribution of partisanship in our YouGov sample and the ANES sample than the MTurk sample.

Comparing our student and MTurk samples to samples of similar groups in existing work points to some differences. While our student sample is equally as Democratic as the sample in Taber and Lodge (2006), the racial make-up differs markedly. This difference is likely due to the differences in student populations at the institutions where these studies were conducted.[11] Our MTurk sample is quite similar to Berinsky et al. (2012). In particular, the two samples are nearly equivalent in age and are largely similar on gender, race and the percentage identifying as Republican.

In addition to the traditional demographic characteristics we also present information about the number of studies these subjects have taken, which will be crucial for our savviness investigation. For now, we note a few striking comparisons. As the bottom row of Table 2 shows, our undergraduate students are the *least*

[11]At the institution where our studies were conducted 76% of students identify as White. At Stony Brook, where Taber and Lodge conducted their study, only 38% identify as White. (Based on: http://www.stonybrook.edu/offires/students/fall2010/ethnic10.html.)

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

experienced of our subject groups. Indeed, our MTurk and YouGov participants average over 30 studies.[12] While these differences may be surprising given conventional wisdom about undergraduate subject pools, they are unsurprising given sample construction: undergraduate students typically spend only a limited period in a subject pool and subject pools typically limit the number of studies conducted in a given semester. In contrast, MTurk participants can take as many studies as available (and numerous studies are often available). Similarly, YouGov panel members can also remain on the panel for as long as they wish, leading to potentially higher rates of study accumulation. We return to the implications of these numbers later in the paper.[13]

STUDIES 1, 2, 3: QUESTION-WORDING

Our first experiment is based on Schwarz et al.'s (1985) canonical study of response option effects. In this experiment, subjects were asked how much time on a typical day they spend following the news and were randomly assigned to receive one of two response option scales: a "low scale" that ranged from "up to 30 minutes" to "more than 2.5 hours" or a "high scale" that ranged from "up to 2.5 hours" to "more than 4.5 hours." Following Schwarz et al.'s explanation, people do not typically store a precise answer to questions like this one in long-term memory that are ready to be accessed. Instead, they use the response scale to help formulate their answer, as it provides an indication of the amount for the average respondent.

Beginning with this study is beneficial for several reasons. First, this study is the "simplest" in the sense that we do not expect a heterogeneous treatment effect (while still acknowledging the possibility that the overall size of treatment effects may be larger for those in one sample versus another). Subjects are not required to read an article or even a particularly long question. There are no detailed instructions and they are not presented with any novel information on any topic. They simply answer a question about their news habits. Second, Schwarz et al. obtain clear results that have since been replicated in a variety of contexts and using a variety of samples.

[12]The number in the paper for YouGov includes the total amount of studies that they have taken. If we restrict our attention just to research studies, the number is 21.2, still well above that of the students.

[13]Berinsky et al. (2012) report a substantially lower average participation rate among their MTurk subjects than we do among ours. There are several possible reasons for this discrepancy. First, they ask their subjects specifically about *political* surveys, whereas we ask about surveys in general. We opted for the broader definition because scholars from other disciplines (e.g. psychology, sociology) also rely on MTurk. Since much experimental research is interdisciplinary, a subject can gain experience after taking a psychology study that could later apply to a political study. Moreover, not all political science studies are clearly political and not all MTurk subjects can easily classify the studies they take. Second, they ask about the number of studies taken in the last month, while we ask about studies in general, which fits our interest in savviness. Finally, they conducted most of their studies largely in the first six months of 2010, whereas ours were in late 2011. There was a notable increase in the use of MTurk for political science research during that time gap. For example, a simple search shows that in 2010 only one paper presented at the Midwest Political Science Association Annual Meeting relied on MTurk, compared to more than 25 papers in 2012.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Table 3
Question Wording Study: Comparison of Response Option Effects by Sample

	Low scale	High scale	Difference	Cross-sample Comparison of difference
Student	13.6%	80.0%	$p < 0.01$	Students/MTurk: $p = 0.00$ Students/YouGov: $p = 0.00$
MTurk	7.3%	27.8%	$p < 0.01$	MTurk/YouGov: $p = 0.89$
YouGov	4.9%	26.7%	$p < 0.01$	

Following Schwarz et al., we look for the effect of differing response scales by comparing the percentage of people who reported following the news for more than 2.5 hours. Table 3 displays the results. In each of our three samples, the results are *conceptually* equivalent to each other and to what Schwarz et al. observed: subjects assigned to the high scale were significantly more likely to report that they follow news more than 2.5 hours per week (based on two-tailed *t*-tests; *p*-values reported in Table 3). Notably, when we compare the size of the estimates we do see a larger effect size for our undergraduate sample relative to the other two (with our MTurk and YouGov estimates being statistically non-distinguishable; see *p*-values marked in the table, based on *F*-tests). Given that the difference we obtain is based on the idea that people often do not have a precise answer to questions like this one stored in long-term memory, these results suggest that undergraduates in college are on average even less aware of how much time they spend following the news relative to those outside of college.

STUDIES 4, 5, 6: AIRLINE OWNERSHIP POLICY

We now move to a second experiment in which we do expect a politically-relevant heterogeneous treatment effect. This one is modeled after canonical framing experiments in which subjects are randomly assigned to read a newspaper article framed positively or negatively around a political issue (Nelson et al. 1997). In our case the issue has to do with ownership of U.S. airlines, and in particular whether foreigners should be allowed to own greater shares of airlines than present law allows. Right now, the Civil Aeronautics Act of 1938 prohibits foreign investors from holding more than 25% of voting stock in any U.S. airline.

Amidst the financial turmoil the industry has endured over the past few years, there have been calls to relax those requirements. The pro and con arguments, which we feature as the main manipulation in our study, have focused mostly on the distribution of economic benefits and costs, particularly as they affect more affluent and densely-populated areas versus less affluent mid-size and smaller cities. The major pro-argument is the possibility of better on-board service enjoyed by the members of the public who fly, whereas the major con-argument is the potential

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

loss of service outside of the country's largest metro areas. The economic aspect of the issue thus speaks directly to distributional questions. To the extent that people are highly concerned about distributional issues, we expect them to view the considerations raised in the frame to be applicable and thus exhibit framing effects (Chong and Druckman 2010). This point is important because, although this issue is obscure for many respondents,[14] there are strong partisan differences in concern about distributional issues. In particular, Democrats demonstrate far greater concern about them than Republicans, with Independents in between.[15] These partisan differences, then, should lead to conditional framing effects (Haider-Markel and Joslyn 2001), in which Democrats are most likely to be moved by our framing while Republicans and Independents are far less so, if at all.

We measure opinions using the following question:

> Would you support or oppose allowing foreign investors to own greater shares of U.S. airlines?

The response options ranged from 1 ("Support strongly") to 7 ("Oppose strongly"). Table 4 presents the results of a shift from the positive to negative frame, in which a positive value represents more opposition to increased foreign ownership of U.S. airlines. All studies included checks for reading ease and comprehension, to ensure that any sample differences were not due to these reading issues.[16]

We first analyze the samples as a whole. We see large framing effects for the whole student and MTurk samples, which is expected given that they both have far more Democrats than the YouGov sample. These overall figures mask important partisan heterogeneities, however. Following our earlier arguments, we disaggregate our sample by partisanship and see a clear pattern. In all three samples, Democrats are framed in precisely the way that we would expect given the differences across treatments. Indeed, when we compare cross-sample group differences (Column 4 of Table 4), we see that the Democrats in all three samples are statistically equivalent.

We see distinctly different patterns among others. In particular, we see marked differences between MTurk participants and the two other samples. While in the

[14]Only 6.5% of students reported that pay a "great deal" of attention to news about the airline industry, which is comparable to the 8.3% of MTurk adults and 5.5% of YouGov adults. Note that this obscurity is consistent with previous framing research (see Druckman et al. 2012).

[15]For example, see the following: http://www.gallup.com/poll/153029/economy-paramount-issue-voters.aspx

[16]We checked that differences in framing effects did not result from failing to read the article by including a comprehension check question immediately following the article. Among undergraduates, 0% of Democrats, 5.3% of Independents and 3.7% of Republicans answered this question incorrectly; among YouGov participants 4.4% of Democrats, 3.1% of Republicans and 2.9% of Independents did so; among MTurk participants 0% of Democrats, 0% of Republicans and 1.45% Independents did so. We also conducted logit analyses to ensure that the likelihood of getting the check question incorrect was not systematically related to demographic or other non-partisanship factors including gender, age, education, experimental group, and news consumption.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Table 4
Framing Effects in Airline Ownership Study

	$\Delta_{opinion}$	s.e.	Cross-sample (a)MTurk, (b)YouGov
Student (All)	**0.63***	(0.33)	*(a)p=0.03*
			(b)p=0.00
Student (Democrats)	**1.33****	(0.51)	(a)p=0.63
			(b)p=0.33
Student (Independents)	−0.92	(0.63)	*(a)p=0.00*
			(b)p=0.12
Student (Republicans)	0.64	(0.61)	*(a)p=0.06*
			(b)p=0.34
MTurk (All)	**1.08*****	(0.23)	*(b)p=0.00*
MTurk (Democrats)	**1.22*****	(0.30)	(b)p=0.42
MTurk (Independents)	**0.73***	(0.39)	*(b)p=0.00*
MTurk (Republicans)	**1.57***	(0.85)	*(b)p=0.01*
YouGov (All)	0.23	(0.21)	
YouGov (Democrats)	**0.98****	(0.49)	
YouGov (Independents)	−0.27	(0.57)	
YouGov (Republicans)	0.27	(0.59)	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, two-tailed t-tests

YouGov and undergraduate samples we see no evidence of framing effects for both Republicans and Independents, we do see evidence for these two groups among MTurk participants (which is especially notable for Republicans, given that they are such a small share of the MTurk sample). In fact, the MTurk Republicans are actually *more* affected by our frame than the MTurk Democrats. Furthermore, while the group differences among Independents and Republicans are statistically similar when we compare YouGov and student subjects (Column 4 of Table 4), our MTurk results are significantly different from both of the other samples. As a next step, and to see if such cross-sample differences persist, we turn to two other studies that consider different approaches to framing. These will focus specifically on subjects' experience taking studies.

ASSESSING SAVVINESS: SOME INITIAL CONSIDERATIONS

Thus far, with the exception of evidence of larger group differences in our first set of studies, students and YouGov samples appear to respond similarly once theoretically-relevant moderators are taken into account. This was not consistently the case for our MTurk sample. Our next step is to investigate a second possible source of heterogeneity in subjects' responses: savviness.

What does it mean for a subject to be experimentally savvy? Following previous research, savviness typically stems from two factors. The first is the sheer number of studies a person has taken (Dalen et al. 2001; Steffens 2004). We have already seen from Table 2 that there is significant variation along this dimension, as students

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

have on average taken far fewer studies than the adults from either our convenience or diverse, national adult sample. All else constant we expect to see that people who have taken more studies are more experimentally savvy. The second factor is learning and training about experimental procedures (Steffens 2004). Participants who receive information about experimental procedures learn more about the experimental process (Morton and Williams 2010). Indeed, informing individuals about the purposes of the study after completion (i.e. "debriefing") is deliberately designed to increase subject knowledge about the experimental process (Brody et al. 2000). Given the additional information included in the debriefing, it is likely that savviness increases with the number of studies taken at a much faster rate among people who are regularly debriefed versus those that are not. This distinction matters for our sample comparison because post-experimental debriefing is a requirement for *all* studies in many undergraduate subject pools (including the one that we used), but typically only a requirement for experiments with adult subjects if the study involves deception (Morton and Williams 2010). Given these two factors affecting savviness, we expect that although the undergraduates have on average taken fewer studies than adults, they may grow equally savvy given debriefing exposure. While the adults in both the convenience and nationally-representative pools are largely forming impressions on their own through repeated participation, the undergraduate subjects are continually informed about the purpose of the studies they have taken.

To examine the consequences of experience taking studies, we create two groups using the number of studies taken (which we call our high and low experience subjects). In order to ensure the robustness of our results, we consider this split in two ways: the median of the number of studies taken and the mean of the number of studies taken. Moreover, in order to ensure that our results are not a function of a particular split, we conduct additional sample-specific tests and use different measures to proxy savviness. Finally, we also conduct a series of checks to ensure that there are no systematic differences among subjects who have taken more studies that might be accounting for our results. We find no evidence of this.[17]

In the next two experiments we use this measure—and the various checks on this measure—to examine the impact of experience via two different types of social information manipulations.

STUDIES 7, 8, 9: DEATH PENALTY

Our first savviness experiment is a 2×2 study in which we manipulate the content of a message subjects receive as well as the types of instructions they receive prior

[17] We estimate a model that could explain experience. We include a number of factors affecting public opinion as independent variables including partisanship, age, education, gender, race, and income (where there is variation on such variables). None are statistically significant. This suggests that it is experience itself—rather than one of the other factors that could be related to experience—that is driving the observed results.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

to the message. Our design is similar to Cook et al. (1970), whose study explicitly considered the way experience with experiments affects responses to stimuli. In their study all subjects received a message about a controversial topic, but were randomly assigned along two factors: the direction of the message (pro or con) and the experimenter's instructions. On this latter factor, some subjects were told to specifically focus on the structure of the message—its wordiness, clarity, and delivery (i.e. the "sentence structure condition")—while others were simply told to listen to the message (i.e. the "basic condition"). Cook et al. (1970) show that highly experienced subjects were more likely to ignore the instructions, leading to equivalent post-treatment outcomes in the sentence structure and basic conditions. This finding was consistent with the idea of savvy subjects searching for a "twist" and disbelieving that an experiment that relies on a controversial topic is really about studying sentence structure. In contrast, among low-experience subjects, there was a statistically-significant difference in the post-treatment responses. This pattern is consistent with the idea that they took the experimenter's instructions at face-value.

We replicate this study through a virtually identical experiment, with the major difference being that the subjects in Cook et al.'s study heard the message while in ours they read it on a computer screen. The controversial issue we use is the death penalty, and we ensure the validity of our message by relying on the treatment Mutz (1992) used about the same issue:

> Many citizens and community leaders on both sides of this issue are convinced that the death penalty [will succeed/will never succeed] in winning the support of the American people and its leaders.

Subjects in our experiment were randomly assigned to either receive the positive or negative version of this message, and then also randomly assigned to receive one of the two following sets of instructions: half were asked to focus on the wording and structure of the statement, while others were simply instructed to read it. All subjects were then asked a question about their support for the death penalty (using a 1–5 scale).

We present our findings in Table 5. Our focus here is not on what differences, if any, occur between the two content cues (i.e. the positive and negative cues). Rather, we are interested in whether those differences differ among people based on the instructions that they received (and their degree of experience taking experiments). For this reason, the cells in Table 5 include difference-in-difference estimates. If the positive cue has the same effect in both the basic and sentence structure treatments, then that would suggest that people are paying equal amounts of attention to the content of the cue despite half of them being instructed otherwise (i.e. they are savvy enough to ignore the instructions). On the other hand, if the instructions move people differently, then that would suggest that some people are paying attention to the sentence structure rather than the content.[18]

[18] Although not presented in Table 5, note that the difference-in-difference estimates differ by experience in both the student and YouGov samples, respectively.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Table 5

Death Penalty Opinion (Difference-in-difference estimates of a change from the negative to positive cue across the basic versus sentence structure conditions; s.e. in parentheses)

	Student	MTurk	YouGov	Cross-sample
High Experience				
ΔBasic - ΔSentence Structure	−0.03 (0.24)	−0.11 (0.23)	0.22 (0.25)	Student/MTurk, p=0.63 Student/YouGov, p=0.13 MTurk/YouGov, p=0.24
Low Experience				
ΔBasic - ΔSentence Structure	**−0.55 (0.29)***	0.11 (0.27)	**−1.59 (0.10)****	*Student/MTurk, p=0.00* *Student/ YouGov, p=0.00* *MTurk/ YouGov, p=0.00*

$*p < 0.1, **p < 0.05, ***p < 0.01$, two-tailed t-tests

First, we see that our YouGov sample acts as we would expect given the relationship between high and low experience. Among those with high experience, there is no difference between the responses to the structure and basic conditions. In contrast, we see a significant difference between the structure and basic conditions for the low experience subjects.[19] A similar pattern emerges for our undergraduate students—we see no differences among the high experience subjects and significant differences among those with low experience.

In contrast, here again our MTurk sample looks different than the others. While, as expected, we see no significant differences among our high experience MTurk participants, we also see no differences among those with low experience, a result that does not conform to expectations and previous findings. This pattern is robust to different approaches to measuring experience.[20]

As a final step, we compare cross-sample differences (the rightmost column of Table 5). First, we see no significant cross-sample differences among high experience

[19] To test the robustness of our results we also examine them if we measure experience using only research studies as opposed to all studies. We rely on this approach as research studies may have been more likely to involve a debriefing than others. Notably, however, it turns out that most of the studies our YouGov subjects had taken were research studies. Likely as a result, we find little difference in our substantive results regardless of whether we rely on the all studies or research studies measure.

[20] Unlike for our undergraduate and YouGov participants, we do not have precise measures of study participation for MTurk respondents. As a result, we consider the robustness of our MTurk results in several ways. First, we use self-reported measures of the *number* of studies that our MTurk participants have taken. Our results are robust to both the mean and median of this number. A second question measures the extent of participation by asking the MTurk participants whether they believe they have taken many, some, few, or no previous MTurk surveys. Notably, responses to this question are highly correlated with the self-reports ($r = 0.61$). The average self-reported number of studies among subjects who report that they have taken many studies is 45, while the average number among those participants who have taken few or some is 22. Our results are robust to using this second measure of experience.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

participants in all three samples, nor do we see significant differences between the low experience MTurkers and any of our subjects with high experience (a pattern suggesting high degrees of savviness even among the less experienced MTurkers). Among the low experience subjects, we see differences between the YouGov and undergraduate samples, in which the students are moved much less on average. One possible reason for this distinction is the idea that there are two paths to experience. Since even our lowest experience student participants completed at least one study prior to participating in this experiment, it is possible that learning through experimental debriefing has a more powerful influence on experience than simply taking repeated studies. It seems reasonable that one experience with being debriefed might be sufficient to know that "tricks" can occur during experiments. If that is the case, then it is possible that low experience undergraduates are still slightly savvier than low experience YouGov participants (a suspicion reinforced by the fact that the difference between the basic and structural conditions is smaller for undergraduates). Further reinforcing this point is the fact that when we limit attention to students who participated in only one study prior to this experiment, the difference between the basic and structural conditions is significantly larger than that reported in Table 5 (and statistically equivalent to the YouGov figure).

STUDIES 10, 11, 12: ELECTORAL COLLEGE

Our last study examines savviness in a different way. Conducting this second savviness study is important because, although partisanship is a common moderator in political behavior studies (such as our airline ownership framing study), previous experience taking studies is not (even though it is a common way in which experimental subjects differ). Thus, it is still useful to conduct a robustness check in another situation in which we believe that experience taking studies should moderate responses.

In our final study we consider a somewhat stronger form of social information: a modified "bogus pipeline" (BPL) experiment. The BPL technique (Jones and Sigall 1971; Roese and Jamieson 1993) is a process in which subjects are led to believe that the experimenter has some increased insight into the subjects' thought process. This is generally untrue—the experimenter simply misleads subjects in order to change their behavior. Key to the BPL is the idea that subjects *believe* that the experimenter somehow has an insight into their attitudes and values. Although BPL was initially suggested as a useful means of correcting for social desirability bias (Jones and Sigall 1971), scholars began to criticize this approach because it required a "naive" study participant (i.e. one who did not have much experience taking experiments; see Ostrom 1973, Sigall and Page 1972). The more experienced the participant, this line of thinking suggested, the less likely he would believe that the experimenter really knew something about him (Ostrom 1973). While this critique certainly limits the

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Table 6
Change in Electoral College Opinion Based on Congruence (s.e. in parentheses)

	Student	MTurk	YouGov	Cross-sample
High Experience	0.00 (0.39)	−0.09 (0.26)	0.42 (0.30)	Student/MTurk, p = 0.69 Student/YouGov, p = 0.08 MTurk/YouGov, p = 0.00
Low Experience	0.04 (0.33)	−0.37 (0.34)	**−0.83**** (0.35)	Student/MTurk, p = 0.16 Student/YouGov, p = 0.00 MTurk/YouGov, p = 0.00

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$, two-tailed t-tests

practicality of this approach for obtaining measures of individual beliefs, it makes a BPL-style experiment useful for our purpose.[21]

Here we rely on a modified BPL in which subjects were initially told that our computer algorithm determined that they find statements supported by political scientists to be most persuasive. All subjects were given this information after they had answered a series of various political questions, including knowledge questions, opinion questions, questions about educational background, and other preference questions. This arrangement was deliberate so that it was conceivable to the subject that the researcher might actually have insight about them.[22] Subsequently, subjects were randomly assigned to receive a message that was either "congruent," a cue in line with what we had identified as their most persuasive source (i.e. political scientists), or "incongruent," a cue supported by a different group (i.e. a majority of the public). The particular treatments are again based on Mutz (1992), except this time focused on the Electoral College:

> A majority of [citizens/political scientists], both Republicans and Democrats alike, are in favor of eliminating the Electoral College, the body of electors appointed by each state that formally elect the President and Vice President of the United States. Occasionally in American history the Electoral College vote has differed from the popular vote, which meant that the person who became President did not receive the majority of the popular vote.

After receiving this statement, subjects were asked to report their opinion on whether to abolish the Electoral College (using a 1-5 scale). Table 6 presents the results, again split by subjects' degree of experience taking studies (following our death penalty study as well as past BPL experiments). Each of the numbers in the

[21] Notably, arguments about the impracticality of the BPL have been heavily applied to undergraduate subject pools. However, there is nothing inherent to this argument that should limit the criticism to undergraduates—indeed, Ostrom's (1973) criticism, for example, hinges on the fact that participants take numerous studies as members of subject pools, not that they are undergraduate students.
[22] In addition, we also conducted a check to make sure that—at least for our low experience subjects—the BPL manipulation was believable.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

table refers to the difference in opinion between those who received the congruent cue versus those that received the incongruent cue. We would expect any significant movement to be in the negative direction, as that would indicate that the incongruent cue is less persuasive than the congruent cue.

To the extent there are no differences in opinion, then we would have evidence that subjects saw through the bogus pipeline prompt (which is what we would expect for high experience subjects). As we see in the top half of Table 6, this is precisely what happened among those with high experience. There were no significant differences in opinion based on whether people saw a congruent or incongruent cue.

The pattern looked different among those with low experience. Here we first see that the YouGov respondents act in line with past work on the bogus pipeline. Those with low experience did not see through the stimulus and thus were persuaded by what they learned from the experimenter. This difference is significantly different from what happened with the high experience subjects. We do not observe similar differences among those with low experience in either the MTurk or undergraduate samples.

Our results for the student and MTurk samples do not match theoretical predictions. Yet we believe that it is worth probing the student results a bit further, as an alternative explanation seems reasonable. Our BPL assumes that subjects find political scientists to be a credible source of information regarding the Electoral College. Yet, given that our undergraduate students were taking political science courses when they took our study, it is possible that such credibility was lacking (for a variety of reasons, not least of which is dissatisfaction with their coursework in the middle of a semester). To address this alternative explanation, we conducted an additional test on the student sample—we told them that they found statements supported by the *majority of the public* to be most persuasive. Then, we offered them either congruent or incongruent statements in a manner identical to what we presented above. The results here were altogether different. Here, again, we see no significant differences among those with high experience ($\mu = 0.16, \sigma = 0.33$) yet the low experience subjects exhibit a marginally significant difference ($\mu = -0.54, \sigma = 0.41$). When we conducted this second BPL study on our MTurk subjects, however, we still did not observe any differences among low (or high) experience subjects. Thus, we can conclude that at least our student samples are displaying a degree of savviness that is consistent with what we'd expect when the BPL assumptions are satisfied.

Taken together, the results from our two savviness studies demonstrate that high experience taking studies leads to savviness, regardless of the context in which you take studies (via YouGov, via MTurk, or as part of a student subject pool). This point underscores why researchers need to be careful about using subjects who have participated in many previous studies when attributes of the experiment might heighten demand characteristics. Lastly, we also found evidence consistent with an unusually high degree of savviness among MTurk participants relative to others.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

CONCLUSION

To address concerns about external validity in experiments, in this paper we have focused on two aspects of narrowness—cognitive factors and heightened savviness—that might (a) limit the generalizability of student samples, and (b) initially suggest the superiority of adult samples (even if they are convenience samples).

Having described the results from our twelve different studies, where do we stand? We address this question in two ways. First, did the results follow theoretical expectations and, second, how did the size of the results compare across samples? On the first question, we see a noticeable difference between our MTurk sample and the other two. The student and YouGov samples consistently produced results that were in line with theoretical expectations, especially once we accounted for relevant moderators. Our MTurk results look different, at least for the experiments that required a bit more "buy-in" from our subjects. When subjects were required to read an article, or trust information from the experimenter, our MTurk sample produced results at odds with what we would predict. In this paper we do not have space to thoroughly consider all of the reasons why such divergence might occur. Yet, from the perspective of replicability, our results do serve to sound a note of caution when using MTurk to produce generalizable results for all but the simplest experimental designs.

Second, there were various points where the effect sizes differed, even though directionally the results followed theoretical expectations. When such differences arose, we offered some potential explanations. Yet we realize that future studies are necessary to thoroughly test them. For now, we would simply conclude that these patterns underscore why researchers who rely upon one type of sample should carefully document reasons why the size of any effect they observe might be larger or smaller with an alternative population or at an alternative moment in political history.

While it may not surprise most readers that undergraduate samples can produce results that differ from those with adult samples, it is important to note that convenience sample adults may *also* not produce replicable results. Even more importantly, while we believe we have made a series of reasonable arguments to explain why our undergraduate samples may have differed from our adult samples, we find it more difficult to explain why our MTurk sample produced different results (even once we accounted for factors such as partisanship and savviness).[23] We thus

[23]It is also instructive to compare our results with other work that has examined MTurk, notably Berinsky et al. (2012). They also note that "several aspects of MTurk should engender caution" yet also find that "MTurk subjects appear to respond to experimental stimuli in a manner consistent with prior research" (16). We believe that our findings are entirely consistent with theirs, given that the experiments they replicated with MTurk are similar in their degree of required "buy-in" as our question-wording experiment. One of their replications—that of Kam and Simas (2010)—arguably requires more "buy-in" from subjects than the other two. However, the buy-in in that case did not derive from the tasks that subjects engaged in or the need to trust information provided to them, but instead the fact that they

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

leave this question as an important task for future work, particularly given the increasing use of MTurk samples among political scientists.

Given the broad nature of the question that motivates our work—Under what conditions do sample characteristics affect an experiment's external validity?—we are only able to scratch the surface in one paper. As we noted at the outset, we have limited ourselves to two types of concerns about external validity and, in addressing them, focused on framing studies. In the future it would be fruitful to expand the analysis to non-framing-based experiments, other framing studies with different types of frames, other types of subject attributes, and other types of generalizability concerns (especially other attributes of the "artificiality" of the lab that have raised suspicion). Nevertheless, we believe that our empirical examination with three commonly-used samples across four experiments provides valuable information for researchers to consider when designing experiments and presenting results.

SUPPLEMENTARY MATERIAL

To view supplementary material for this paper, please visit http://dx.doi.org/10.1017/S2052263014000074.

REFERENCES

Arceneaux, K. 2012. Cognitive Biases and the Strength of Political Arguments. *American Journal of Political Science* 56: 271–85.

Banks, A., and Valentino, N. A. 2012. Emotional Substrates of White Racial Attitudes. *American Journal of Political Science* 56: 289–97.

Barabas, J., and Jerit, J. 2010. Are Survey Experiments Externally Valid? *American Political Science Review* 104: 226–42.

Berinsky, A. J., Huber, G. A., and Lenz, G. S. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20: 351–68.

Brady, H. 2000. Contributions of Survey Research to Political Science. *PS: Political Science and Politics* 33: 47–57.

Brody, J. L., Gluck, J. P., and Aragon, A. S. 2000. Participants' Understanding of the Process of Psychological Research: Debriefing. *Ethics & Behavior* 10: 13–25.

Brooks, D. J., and Geer, J. G. 2007. Beyond Negativity: The Effects of Incivility on the Electorate. *American Journal of Political Science* 51: 1–16.

Chong, D., and Druckman, J. N. 2007. Framing Public Opinion in Competitive Democracies. *American Political Science Review* 101: 637–55.

Chong, D. and Druckman, J. N. 2010. Framing Theory. *Annual Review of Political Science* 10: 103–26.

Cook, T. D., Bean, J. R., Calder, B. J., Frey, R., Krovetz, M. L., and Reisman, S. R. 1970. Demand Characteristics and Three Conceptions of the Frequently Deceived Subject. *Journal of Personality and Social Psychology* 14: 185–94.

repeated the same framing experiment twice and received a series of other questions regarding their risk orientation.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Dalen, L. H., Stanton, N. A., and Roberts, A. D. 2001. Faking Personality Questionnaires in Personnel Selection. *Journal of Management Development* 20: 729–42.

Druckman, J. N., Fein, J., and Leeper, T. J. 2012. A Source of Bias in Public Opinion Stability. *American Political Science Review* 106: 430–54.

Druckman, J. N., and Kam, C. D. 2011. Students as Experimental Participants: A Defense of the 'Narrow Data Base'. In *Handbook of Experimental Political Science.* eds. Druckman, Green, Kuklinski, and Lupia, New York: Cambridge University Press.

Druckman, J. N., and Leeper, T. 2012. Learning More from Political Communication Experiments: Pretreatment and Its Effects. *American Journal of Political Science* 56: 875–96.

Gaines, B. J., Kuklinski, J. H., and Quirk, P. J. 2007. The Logic of the Survey Experiment Reexamined. *Political Analysis* 15: 1–20.

Gartner, S. S. 2008. The Multiple Effects of Casualties on Public Support for War: An Experimental Approach. *American Political Science Review* 102: 95–106.

Gerber, A. S., and Green, D. P. 2008. Field Experiments and Natural Experiments. In *The Oxford Handbook of Political Methodology.* eds. Box-Steffensmeier, Brady, and Collier, Oxford: Oxford University Press.

Gerber, A. S., Huber, G. A., Doherty, D., and Dowling, C. M. 2011. Citizens' Policy Confidence and Electoral Punishment: A Neglected Dimension of Electoral Accountability. *Journal of Politics* 73: 1206–24.

Haider-Markel, D. P., and Joslyn, M. R. 2001. Gun Policy, Opinion, Tragedy, and Blame Attribution: The Conditional Influence of Issue Frames. *Journal of Politics* 63: 520–43.

Henry, P. J. 2008. College Sophomores in the Laboratory Redux: Influences of a Narrow Data Base on Social Psychology's View of the Nature of Prejudice. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory* 19: 49–71.

Hopkins, D., and King, G. 2010. Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. *Public Opinion Quarterly* 74: 201–22.

Huber, G. A., and Paris, C. 2013. Assessing the Programmatic Equivalence Assumption in Question Wording Experiments: Understanding why Americans Like Assistance to the Poor More Than Welfare. *Public Opinion Quarterly* 77: 385–97.

Huddy, L. and Khatib, N. 2007. American Patriotism, National Identity, and Political Involvement. *American Journal of Political Science* 51: 63–77.

Iyengar, S. 2011. Laboratory Experiments in Political Science. In *Handbook of Experimental Political Science.* eds. Druckman, Green, Kuklinski, and Lupia, New York: Cambridge University Press.

Jerit, J. 2009. How Predictive Appeals Affect Policy Opinions. *American Journal of Political Science* 53: 411–26.

Jones, E. E. and Sigall, H. 1971. The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude. *Psychological Bulletin* 76: 349–64.

Kam, C. D. 2007. Implicit Attitudes, Explicit Choices: When Subliminal Priming Predicts Candidate Preference. *Political Behavior* 29: 343–67.

Kam, C. D. and Simas, E. 2010. Risk Orientation and Policy Frames. *Journal of Politics* 72 (2): 381–96.

Kam, C. D., Wilking, J. R. and Zechmeister, E. J. 2007. Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research. *Political Behavior* 29: 415–40.

Kinder, D. 2003. Communication and Politics in the Age of Information. In *Oxford Handbook of Political Psychology.* eds. Sears, Huddy, and Jerivs, Oxford University Press.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press

Kinder, D. and Palfrey, T. 1993. On Behalf of Experimental Political Science. In *Experimental Foundations of Political Science*. eds. Kinder and Palfrey, Ann Arbor, Michigan: University of Michigan.

Levine, M. and Ensom, M. H. H. 2001. Post hoc Power Analysis: An Idea Whose Time Has Passed? *Psychometry* 21: 405–09.

McDermott, R. 2002. Experimental Methodology in Political Science. *Political Analysis* 10: 325–42.

McDermott, R. 2011. Internal and External Validity. In *Handbook of Experimental Political Science.* eds. Druckman, Green, Kuklinski, and Lupia, New York: Cambridge University Press.

McGraw, K. 2011. Candidate Impressions and Evaluation. In *Handbook of Experimental Political Science*. eds. Druckman, Green, Kuklinski, and Lupia, New York: Cambridge University Press.

Miller, J. M., and Krosnick, J. A. 2000. News Media Impact on the Ingredients of Presidential Evaluations: Politically Knowledgeable Citizens are Guided by a Trusted Source, *American Journal of Political Science* 44: 301–15.

Morton, R. B., and Williams, K. C. 2010. *Experimental Political Science and the Study of Causality.* New York: Cambridge University Press.

Mutz, D. 1992. Impersonal Influence: Effects of Representations of Public Opinion on Political Attitudes, *Political Behavior* 14: 89–122.

Mutz, D. 2011. *Population-Based Survey Experiments.* Princeton, NJ: Princeton University Press.

Nelson, T. E., Bryner, S. M. and Carnahan, D. 2011. Media and Politics. In *Handbook of Experimental Political Science.* eds. Druckman, Green, Kuklinski, and Lupia, New York: Cambridge University Press.

Nelson, T. E., Clawson, R. A., and Oxley, Z. M. 1997. Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance. *American Political Science Review* 91: 567–83.

Ostrom, T. M. 1973. The Bogus Pipeline: A New Ignis Fatuus? *Psychological Bulletin* 79: 252–59.

Roese, N. J. and Jamieson, D. W. 1993. Twenty Years of Bogus Pipeline Research: A Critical Review and Meta-Analysis. *Psychological Bulletin* 114: 363–75.

Schwarz, N., Hippler, H.-J., Deutsch, B., and Strack, F. 1985. Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly* 49: 388–95.

Sears, D. 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: 515–30.

Shadish, W. R., Cook, T. D., and Campbell, D. T. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Sigall, H., and Page, R. 1972. Reducing Attenuation in the Expression of Interpersonal Affect via the Bogus Pipeline. *Sociometry* 35: 629–42.

Steffens, M. C. 2004. Is the Implicit Association Test Immune to Faking? *Experimental Psychology* 51: 165–79.

Taber, C. and Lodge, M. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50: 755–69.

Transue, J. E., Lee, D. J., and Aldrich, J. H. 2009. Treatment Spillover Effects Across Survey Experiments. *Political Analysis* 17: 143–61.

Vavreck, L. and Rivers, D. 2008. The 2006 Cooperative Congressional Election Study. *Journal of Elections, Public Opinion, and Parties* 18: 355–66.

https://doi.org/10.1017/xps.2014.7 Published online by Cambridge University Press