Public Opinion Quarterly Advance Access published April 11, 2013

Public Opinion Quarterly

COMPARING CONTEMPORANEOUS LABORATORY AND FIELD EXPERIMENTS ON MEDIA EFFECTS

JENNIFER JERIT*
JASON BARABAS
SCOTT CLIFFORD

Abstract Researchers employing experiments often conduct their studies in the laboratory or in the field. Each mode has specific advantages (e.g., the control of the lab versus the realistic atmosphere of the field). Two hypotheses concerning the relationship between treatment effects in lab and field settings were tested in contemporaneous experiments. Registered voters in a medium-size city were randomly assigned to a laboratory or a field experiment involving newspaper stimuli. The analyses show significantly larger treatment effects in the laboratory experiment, especially for public opinion outcomes in which the content of the treatment could be readily linked to the dependent variable. Evidence also suggests that differences in the size of treatment effects moderate as lab and field experiments become similar on one dimension—namely, the temporal distance between stimulus and outcome.

JENNIFER JERIT and JASON BARABAS are Associate Professors of Political Science at Stony Brook University, Stony Brook, NY, USA. SCOTT CLIFFORD is a Ph.D. candidate in the Department of Political Science at Florida State University and a pre-doctoral fellow at Duke University's Social Science Research Institute, Durham, NC, USA. The authors thank Kevin Arceneaux, Charles Barrilleaux, Adam Berinsky, Bob Crew, Don Green, Bob Jackson, Cindy Kam, Jim Kuklinski, Rick Lau, Cherie Maestas, Becky Morton, Spencer Piston, Evan Parker-Stephen, Mary Stutzman, Carlisle Rainey, David Redlawsk, John Ryan, Gaurav Sood, and the editors for helpful comments and suggestions. They also thank the staff of the *Tallahassee Democrat*, including the general circulation manager, Richard Kay, and the business systems manager, Mike Hoke. Previous versions of this paper were presented at the Vanderbilt University Conference on Laboratory Experiments, the New York Area Political Psychology Meeting, the Department of Political Science at Florida State University, and the annual meetings of the American Political Science Association and the International Society of Political Psychology. This work was supported by Florida State University's Council on Research and Creativity [Project 023839 to J. J. (lead PI)]. *Address correspondence to Jennifer Jerit, Department of Political Science, Social and Behaviorial Sciences Building, 7th Floor, Stony Brook University, Stony Brook, NY 11794-4392, USA; e-mail: jennifer.jerit@stonybrook.edu.

doi:10.1093/poq/nft005

Experiments are powerful because they demonstrate cause and effect in an especially compelling way. As a result, the use of randomized experiments—particularly lab experiments—is on the rise. Notwithstanding this trend, some argue that the findings from lab experiments have limited external validity because of (1) sample characteristics (e.g., the frequent use of undergraduates as research subjects), or (2) the artificial nature of the study's setting (McDermott 2002; Morton and Williams 2010; Iyengar 2011). In response to criticism about unrepresentative convenience samples often used in lab settings, a growing number of scholars employ survey experiments with diverse adult subject populations (e.g., Gilens 2001; Brooks and Geer 2007). Others, like Druckman and Kam (2011), argue that concerns about the alleged "college sophomore" problem are overstated. They offer a vigorous defense of convenience samples (including those composed of students) and demonstrate via simulation that the conditions under which such samples affect causal inference are rare.

Our study is motivated by the second critique, namely the purported artificiality of the laboratory setting. This is an important topic because concerns about the lack of realism have resulted in a wave of field experiments on subjects traditionally examined in the lab—most notably, several recent studies using mass media and campaign treatments (e.g., Albertson and Lawrence 2009; Arceneaux and Kolodny 2009; Gerber, Karlan, and Bergan 2009; Arceneaux and Nickerson 2010; Gerber, Huber, and Washington 2010; Gerber et al. 2011). In these studies, the rationale for a field experiment is the importance of examining political phenomena in a naturalistic setting. Lab experiments have impressive levels of internal validity, the argument goes, but the empirical findings emerging from them may not be reliable indicators of the effects that would be observed in the real world.

Yet the empirical evidence on this point is virtually non-existent. There have been some attempts in economics (e.g., Benz and Meier 2008) and psychology (e.g., Mitchell 2012) to compare effect sizes across lab and field studies, but there remain differences (in either the timing of lab and field studies, the stimuli in each setting, or the participants) that limit the conclusions one can draw. In fact, no existing study has compared *contemporaneous* lab and field experiments with a *similar* treatment. Thus, the issue of whether the insights from the lab extrapolate to the "world beyond" (Levitt and List 2007) remains an empirical question. We explore this topic in a study that manipulates the lab versus field experience. Drawing from the same target population, we administer archetypical and contemporaneous experiments in the laboratory and in the field. In doing so, this study represents one of the first attempts to compare treatment effects from two types of experimental settings.

State of the Literature

In recent years, researchers have begun to implement field experiments in substantive areas that once had been the purview of laboratory experimenters.

In these studies, scholars note the importance of conducting research in a naturalistic setting, such as an actual election campaign or, simply, the "real world." For example, in their examination of the durability of broadcast media effects, Albertson and Lawrence describe their decision to conduct a randomized field experiment this way: "[our] design allows respondents to view programs in their own homes, thus more closely approximating regular viewing conditions" (2009, pp. 276–7). Likewise, Gerber, Karlan, and Bergan (2009) examine the effect of newspapers on political attitudes and behavior in a field experiment in the Washington, DC, area. The authors state that "[field] experimentation has some advantages over … previous research strategies [e.g., lab experiments], namely the use of a naturalistic setting" (p. 37). Finally, in their investigation of negative and positive campaign ads, Arceneaux and Nickerson (2010) conduct a field experiment so they can "estimate the effects of message tone in the context of an actual campaign" (2010, p. 56). The common thread across these studies is the notion that field experiments combine the internal validity of randomized experiments *and* increased external validity because the study is conducted in a real-world setting (see, e.g., Arceneaux 2010; Gerber 2011).[1]

In addition to the benefits of administering a study in the environment in which the phenomenon of interest naturally occurs, a corresponding claim about the disadvantage of the lab is often made. In particular, there is concern that laboratory effects are different from the effects that would be observed if the same study were conducted in the field (e.g., Levitt and List 2007). For example, Gerber (2011) observes, "Although it is often remarked that a laboratory experiment will reliably indicate the *direction* but not the magnitude of the effect that would be observed in a natural setting, to my knowledge that has not been demonstrated …" (p. 120, emphasis original). Our study takes a step in that direction by assessing the degree of correspondence in the findings from contemporaneous lab and field experiments involving similar experimental treatments. We begin by outlining below some of the defining characteristics of lab and field settings that may cause outcomes to diverge in the two settings.

SOME CONTEXTUAL DIFFERENCES ACROSS LAB AND FIELD

One of the essential features of a laboratory experiment is that it takes place in a controlled setting (Aronson et al. 1990; Field and Hole 2003). This heightened degree of control has several consequences. First, it allows for the standardization of procedures across treatment and control groups (McDermott 2002). With the exception of the manipulation, everything about the experiment—both the procedures and how they are implemented—is the same for all participants. Moreover, there is little behavioral leeway when it comes to

1. For some, the fact that field experiments take place in naturalistic settings makes them more ecologically valid, not necessarily more externally valid (Morton and Williams 2010; Mutz 2011).

reacting to the stimulus (e.g., participants often are confined to a computer terminal and not allowed to communicate with others unless that is an explicit feature of the study, as in Druckman and Nelson [2003]). When combined with random assignment, standardization ensures that any difference in outcomes between the treatment and control groups can be attributed to the stimulus, not extraneous factors. A second and related dimension of experimental control pertains to the delivery of the treatment. Aside from subject inattentiveness or computer malfunction, treatment subjects are exposed to the stimulus. In fact, exposure to the treatment is effectively forced—a characteristic that some scholars have come to view as a liability (e.g., Kinder 2007; Gaines and Kuklinski 2011; Arceneaux, Johnson, and Murphy 2012). This aspect of control is one of the primary differences between lab and field experiments, since several types of failure to treat problems may arise in the field (Nickerson 2005). A final element of control pertains to the pristine environment of most laboratory settings (Kinder 2007). Unlike field experiments where the manipulation must compete with noise from the real world, laboratory settings have few distractions unless they are intended (i.e., controlled) by the experimenter. As a result of these differences between the lab and the field, the impact of lab treatments will likely be greater than comparable stimuli administered in a naturalistic setting.

A second dimension in which lab and field experiments differ is the obtrusiveness of the experiment and, thus, the potential for subject reactivity (Webb et al. 2000). It is common practice to obtain informed consent from participants in the laboratory but not in the field (Singer and Levine 2003). Thus, participants in a lab experiment *know* they are being studied, and this awareness may cause demand characteristics to confound the effect of the treatment (see Shadish, Cook, and Campbell [2002] for discussion).[2] Even if the typical subject in a lab study (i.e., a college student) is unmotivated, the situational forces of the experimental context are strong. In most cases, subjects come to a specified location to participate and they have made an appointment in advance. They also are explicitly asked to give their consent. These situational factors, we surmise, will cause subjects to pay greater than usual attention to the stimuli in laboratory settings. A related difference concerns the assessment of outcomes. Field experimenters often measure outcomes unobtrusively (e.g., with administrative data), whereas lab experimenters use more obtrusive methods, such as a questionnaire. Thus, treatment effects may be magnified as a result of the precision and demand that characterize lab-based outcome measures.

The third way lab and field experiments differ is the distance, in terms of time, between the stimulus and the outcome measure(s). In the typical laboratory experiment, this distance is measured in minutes or, possibly, hours.

2. Gerber, Green, and Larimer's (2008) study of social pressure is a notable exception. As part of some of the treatments, participants were told that their voter participation was being studied.

By contrast, in field settings, days or even weeks may pass in between the application of the stimulus and the measurement of the outcome.[3] This difference matters because the greater the time between treatment and outcome, the more likely the treatment effects are to attenuate. Even if one were able to perfectly replicate a lab finding in a field setting, the mere passage of time might complicate efforts to measure that effect.

Study Design

The purpose of our design was to deliver exogenously the same information in two different experimental contexts—one highly controlled, the other naturalistic—and to examine how evaluations of public officials, policy preferences, and political knowledge were affected. As we describe below, we went to great lengths to ensure that the lab and field treatments were comparable. In addition, the similarity of the lab and field studies was maximized in other ways. The participants in each experiment were random draws from the same target population, and the studies were conducted in the same city and at roughly the same moment in time. As a result, we held constant many of the factors that have prevented previous scholars from comparing the results of lab and field experiments (e.g., differences related to the subject population or the political environment). It was equally important, however, that the experimental interventions reflect the defining features of each style of experimentation in terms of control, obtrusiveness, and time distance. Thus, the lab and field studies were designed to be "typical" of experiments in each area.

In terms of the substance of the design, we implemented a political communications experiment with a single treatment group and a single control group. The lab study was modeled after past experiments in which treatment subjects are exposed to actual media reports or faux stories that are realistically inspired (e.g., Iyengar and Kinder 1987; Nelson, Clawson, and Oxley 1997; Berinsky and Kinder 2006). In our case, participants came to a computer lab where they completed a self-administered questionnaire. The experiment featured an information treatment (in this case, stories from the local paper) followed by questions measuring knowledge and attitudes regarding the topics covered in the news stories. Likewise, our field study was modeled after previous field experiments using an information treatment (e.g., Albertson and Lawrence 2009; Gerber, Karlan, and Bergan 2009). As with those studies, our participants were unobtrusively treated with information from a local media outlet, and outcomes were assessed in a survey administered at a later

3. There is variation in lab and field experiments on this dimension. Recent lab studies have examined the duration of experimental effects, which necessitates the measurement of outcomes over a longer period of time (e.g., Mutz and Reeves 2005; Chong and Druckman 2010).

point in time.[4] In the empirical analyses discussed below, we focus on the effect of treatment assignment because this is the causal quantity estimated in many earlier lab and field experiments. What distinguishes our study from past efforts to explore the generalizability of lab experiments is that we implemented lab and field studies on the same topic at the same moment in time and with samples drawn from the same population. This design enabled us to determine if the results of an experiment differ depending on whether the study was conducted in the field or in the lab.

SAMPLE RECRUITMENT AND TIMELINE

The sample for the study was 12,000 individuals in Leon County, Florida. We started with a list of 171,187 registered voters from the Leon County Supervisor of Election's office based upon administrative records as of late January 2011. From that list we eliminated anyone who resided in a precinct outside a five-mile radius of the county election office in the city center.[5] Next, we discarded anyone deemed ineligible according to the Institutional Review Board application (see the appendix for more details). Finally, working with staff members at the *Tallahassee Democrat*—the only local major newspaper serving the area—we identified households (i.e., addresses) that were not current subscribers to the paper. After applying all these procedures, we were left with 18,668 individuals who could potentially be selected for inclusion in the study.

Participants were randomly assigned to be in the field experiment ($n = 6{,}000$) or the laboratory experiment ($n = 6{,}000$). In the field experiment, half of the households (3,000) were randomly assigned to the control group, which did not receive a free newspaper subscription, and the other half of the households were randomly assigned to the treatment group, which received a one-month Sunday-only subscription to the *Tallahassee Democrat* beginning on Sunday, February 27, 2011. Along with the first free Sunday paper, these households received a flyer telling them that they had won a "complimentary one-month subscription to the *Tallahassee Democrat*" in a drawing. The purpose of the flyer was to inform treatment subjects about their free subscription and to discourage them from drawing a connection between the subscription and the post-treatment questionnaire. We sent a mail survey to all 6,000 people in the field experiment the week of March 7, two weeks after the treatment group's free subscription began. As an inducement to participate, the cover letter indicated that there would be a raffle for two $300 awards for those who returned their survey.

4. Many field experiments measure outcomes with administrative data (e.g., Gerber, Green, and Larimer 2008), but several recent studies assess outcomes with a survey (e.g., Albertson and Lawrence 2009; Gerber, Karlan, and Bergan 2009; Arceneaux and Nickerson 2010; Gerber et al. 2011).

5. The only exception to this procedure was the elimination of a small number of people without a permanent address who were listed as registered to vote at that governmental office. The use of a five-mile radius was intended to increase the response rate in the lab experiment and to ensure delivery of the newspapers in the field. The appendix reports other details on sample eligibility.

In the lab experiment, we recruited 6,000 people through the U.S. mail to come to a university campus to participate "in a study that examines people's political attitudes and behaviors." In exchange for their participation, invitees were told they would receive $30 and that free parking would be arranged by the study organizers. The lab experiment was timed to coincide with the field experiment. Thus, the lab sessions ran from Sunday, March 6, until Saturday, March 12, with subjects being randomly assigned to treatment and control via computer (see appendix for additional details on experimental protocol). The participants in the two arms of our study not only answered the same questions, but they also were answering the questions at roughly the same time.[6]

TREATMENT AND OUTCOME MEASURES

In this study, the treatment is the presentation of newspaper articles. In the lab, treatment subjects were asked to "take a moment to read" several stories about local politics from the *Tallahassee Democrat*. In the field, treatment subjects received complimentary newspapers containing those same stories.[7] As a result of our decision to treat people with actual newspaper stories, the "real world" dictated the content of our treatments. In our case, we anticipated that there would be coverage of local politics due to the start of the spring legislative session on Tuesday, March 8.[8] Treatment subjects in the field experiment received two Sunday papers (on February 27 and March 6) before they got our mail survey. Treatment subjects in the laboratory experiment read the four most prominent stories about state politics from the February 27 and March 6 papers. These included the front-page story from each paper, as well as two prominent interior stories from the March 6 issue. To the extent possible, we maximized the similarity of the treatments across the two contexts. In terms of content, we included the most salient stories about local politics across the two issues.[9] When it came to appearance, the treatments in the lab were screen shots of the *Tallahassee Democrat*, so they were exact replicas of the stories appearing in the newspaper. Table 1 describes the central features of each article.

6. Approximately 40 percent of field participants returned their survey during the week of the lab study.

7. Naturally, treatment subjects in either context could choose not to read the newspaper articles. Reaction timers in the lab revealed that the majority of subjects viewed the stories. As with many previous field experiments, we are unable to measure attention to the stimulus. However, treatment subjects in the field were more likely than were control subjects to state in the survey that they were "following stories about the upcoming legislative session" ($p < .05$), suggesting that these people were paying attention to the free newspapers.

8. Tallahassee is the state capital of Florida. Historically, the *Tallahassee Democrat* has previewed the issues facing the legislature the Sunday before the start of the session (which in our case was March 6). The 2011 session was expected to be newsworthy because the state's newly elected governor faced a $3.6 billion budget deficit.

9. Treatment subjects in the field received two additional Sunday papers (on March 13 and March 20) as part of their complimentary subscription. Auxiliary content analyses indicate that there was virtually no coverage of the topics on our questionnaire in those later newspapers.

Table 1. Summary of Laboratory Treatments

	First story	Second story	Third story	Fourth story
Date	Sunday, March 6	Sunday, March 6	Sunday, March 6	Sunday, February 27
Headline	"State employees haven't had this much at stake since 1950s"	"Pension hot topic for 2011 session"	"Nina Hayden: Passion for public service helps balance reduced pay"	"Pro-union rally held at Capitol"
Location	Front page (A1); above fold	Page A6	Page A6	Front page (A1); above fold
Length	18 paragraphs	9 paragraphs	12 paragraphs	17 paragraphs
Images	Column graph showing size of state workforce over past four years. Graph shows workforce as small in 2010 as it was in 2006.	Pie chart showing breakdown of participants in Florida's retirement system.	Photograph of African American woman, Hayden, who is described as public defender from Clearwater, FL.	Two photographs of protestors, one with a "stop corporate greed" sign, the other in support of Governor Rick Scott.
Summary	Gov. Scott plans to cut spending by increasing state employees' medical premiums, mandating contributions to retirement fund, cutting jobs and weakening collective bargaining.	Florida retirement system is underfunded. Gov. Scott and legislators propose mandating employee contributions. Article discusses potential changes to the pension plan.	Perspective of a state employee on proposed budget cuts. State employees are underappreciated, and proposed cuts undermine the incentive to work for the state.	State employees protesting Gov. Scott's proposals, which are perceived as balancing the budget on the backs of state workers. Comparisons made between Gov. Rick Scott and Wisconsin Gov. Scott Walker.
Key quotes	"For state employees, the 2011 legislative session will be a pivotal 60 days…as legislators grapple with a $3.6 billion shortage in state revenues."	"The state proposes to make public employees … chip in for their pensions that are now fully paid by their government employers."	"Hayden said Scott, a wealthy hospital executive who never held office before he ran on an anti-bureaucracy platform last year, doesn't seem to relate to working people—particularly those in government."	"Scott has called for $5 billion in spending cuts, eliminating about 8,600 state job positions…"

NOTE.—Stories are arrayed in the order in which they were viewed by laboratory subjects.

The outcome measures asked about a wide variety of topics, most having to do with local political issues and the spring 2011 legislative session. The full text of the questionnaire is provided in the appendix, but in brief, it contained items on job approval, knowledge, policy preferences, and attention to local politics. Given the nature of the project, the questionnaire had to be designed in advance of the actual study period. To ensure that the questions covered newsworthy topics—and thus had a chance of being covered in the *Tallahassee Democrat*—we studied the governor's proposed budget and consulted with members of the local media and organizations that regularly conduct surveys in Florida. We were relatively successful, though there is variation in the degree to which our real-world treatments addressed the topics on the questionnaire. In the analyses below, we create a variable called *Correspondence Score*, which represents the number of treatment stories (0 to 4) that were related to an outcome measure. For example, all four treatment stories had negative coverage of Governor Rick Scott, and so the gubernatorial approval question received a value of 4 on *Correspondence Score*. In contrast, none of the articles mentioned immigration, resulting in a value of 0 for a question that assessed support for a tough anti-immigration bill.[10]

Hypotheses

As a result of differences in impact, obtrusiveness, and time distance, treatment effects should be more likely to manifest in a laboratory setting. Thus, with a similar stimulus administered in both experimental settings, we expected the magnitude of treatment effects to be larger in the lab (H1). That said, there is tremendous variation in how closely an experimental treatment is related to the outcome measure(s). The connection may be defined by the physical appearance of the treatment (e.g., Does the stimulus appear by itself, say in a single paragraph, or is it embedded in a longer news story or series of stories?) as well as by semantic features of the treatment (e.g., Does the stimulus imply a particular response or must subjects make an inference?). We expected that the closer the connection between a stimulus and an outcome, the more likely the laboratory setting would exaggerate the effect of the treatment, relative to the field (H2). Here, our earlier discussion of the contextual differences across experimental settings is instructive. When the treatment-outcome connection is close, demand effects are more likely to occur because participants may intuit what the researcher is "looking for." Similarly, when the impact of the stimulus is greater and the time distance between stimulus and outcome is shorter, the greater the chance that relevant underlying attitudes become activated. Conversely, factors such as impact, obtrusiveness, and time should

10. Two coders, working separately, read the treatment stories and assigned correspondence ratings (Krippendorff's alpha = .72).

make little difference in the estimation of treatment effects if the correspondence between the stimulus and the outcome measure(s) is distal or absent altogether.

Empirical Results

Figure 1 recaps the study design and provides information regarding participation patterns.

In our study, 12,000 individuals were randomized into a lab experiment or a field experiment. Of the 6,000 people who were sent a mail survey in our field experiment, 19 percent completed the questionnaire (n = 1,135), a response rate that did not differ significantly across the treatment and control conditions (p = .39). In the laboratory experiment, 417 invitees came to our lab and participated in the study, resulting in a 7-percent response rate.[11] Despite

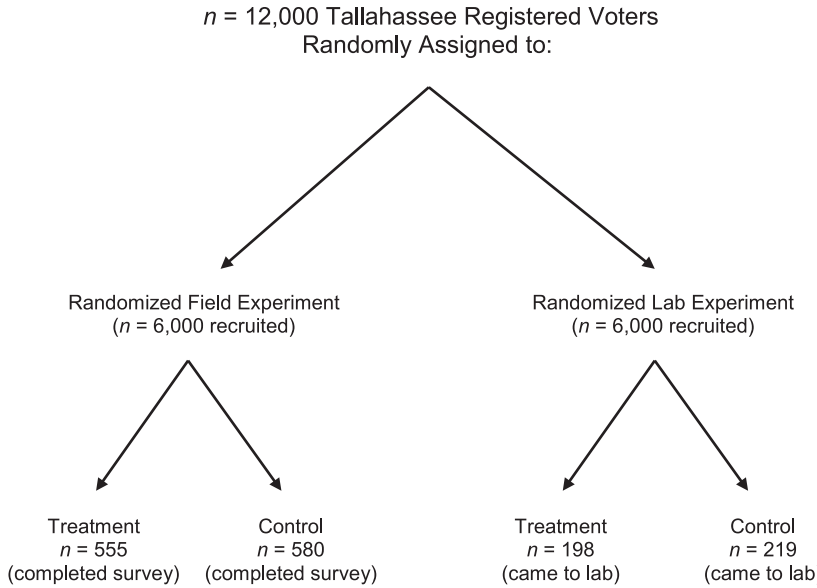

Figure 1. Research Design Overview.

11. In the field, 653 surveys were not delivered; in the lab study, 253 invitations were returned (American Association for Public Opinion Research [AAPOR] Response Rate 3). The higher field response rate is due to the fact that although non-blacks were more likely to take part in either arm of the study, this effect was stronger in the field. Additionally, Democrats and older people were significantly less likely to participate in the lab relative to Independents and younger people. Once we account for these differences, the lab and field response patterns become statistically indistinguishable.

differences in the response rates, the factors predicting participation in the field or lab study were remarkably similar (see appendix for details).

Using covariate information from the voter file, we conducted extensive randomization checks. For the most part, randomization was successful, both at the initial stage (into the field versus the lab) and later, when participants were assigned into treatment and control groups in each context. The only exception to this pattern is that in the treatment condition of the field experiment a slightly higher concentration of African Americans and a lower number of 2010 voters returned a survey ($p < .05$; see the appendix for more details on the randomization and participation analyses). As a precaution, we confirmed that the results hold with controls for demographic factors and household size.

TREATMENT EFFECTS ACROSS EXPERIMENTAL MODES

Our analysis focuses on the effect of treatment assignment, which corresponds to the effect of being presented with news stories either at a computer lab or at one's home.[12] In order to facilitate comparisons across the variables in our analysis, the outcomes are dichotomized, but the same patterns obtain in analyses with the original measures. The first six columns of table 2 show the condition means in each experimental setting as well as the treatment minus control (T – C) differences, which appear under the "Effect" columns. The "DID" column represents the difference-in-differences, or the difference between the treatment effects in each context. The final column presents the value of *Correspondence Score* for each outcome, with higher values representing questions with the closest connection to the treatment.

When viewed as a whole, table 2 reveals six instances (appearing in the first six rows in gray shading) in which the lab effect is statistically significant, while the field effect is indistinguishable from zero. There is only one case in which the field effect is statistically significant and the lab effect is null.

The first hypothesis predicts that treatment effects in the laboratory experiment will be larger than those in the field experiment. To evaluate this hypothesis, we compared the treatment minus control differences in each context (i.e., the difference-in-differences). That quantity appears in the seventh column of table 2 and is represented by the following calculation: $[\text{Treatment}_{\text{Lab}} - \text{Control}_{\text{Lab}}] - [\text{Treatment}_{\text{Field}} - \text{Control}_{\text{Field}}]$. Beginning with the first entry (Approve of Governor Rick Scott), approval in the baseline condition is .32, which means that about a third of lab subjects not in the treatment group approved of the job the governor was doing. Moving across the table, approval drops to .23 for treatment subjects, resulting in a –.09 difference that is statistically significant ($p < .05$, two-tailed t-test). The next set

12. Noncompliance issues (e.g., not taking the paper when assigned it or getting the paper when assigned to the control) were minimal, and following Gerber, Karlan, and Bergan (2009), we estimate intent-to-treat (ITT) effects.

Table 2. Summary of Condition Means, Treatment Effects, Difference-in-Differences, and Treatment Correspondence

	Laboratory (n = 417)			Field (n = 1,135)				
	Control	Treatment	Effect	Control	Treatment	Effect	DID	Correspondence
Approve of Governor Rick Scott	.32 (.03)	.23 (.03)	–.09 (.04)**	.26 (.02)	.27 (.02)	.01 (.03)	**–.10 (.05)***	4
Approve of Governor Scott on economy	.33 (.03)	.23 (.03)	–.10 (.04)**	.28 (.02)	.28 (.02)	–.00 (.03)	**–.10 (.05)***	4
Most imp. issue: Florida budget	.23 (.03)	.33 (.03)	.10 (.04)**	.20 (.02)	.20 (.02)	–.00 (.02)	**.10 (.05)****	4
Prefer state workers pay retirement	.48 (.03)	.40 (.03)	–.08 (.05)**	.47 (.02)	.46 (.02)	–.01 (.03)	**–.08 (.06)**	4
Most imp. issue: FL gov't workers	.04 (.01)	.08 (.02)	.04 (.02)*	.03 (.01)	.05 (.01)	.02 (.01)	**.02 (.02)**	4
Prefer property & business tax cuts	.38 (.03)	.25 (.03)	–.13 (.05)**	.30 (.02)	.27 (.02)	–.03 (.03)	**–.10 (.05)***	3
Believe budget problems are serious	.56 (.03)	.60 (.03)	.05 (.05)	.57 (.02)	.56 (.02)	–.00 (.03)	**.05 (.06)**	3
Trust Florida state government	.34 (.03)	. 31 (.03)	–.03 (.05)	.29 (.02)	.30 (.02)	.01 (.03)	**–.04 (.05)**	3
Know Gov. Scott was a businessman	.88 (.02)	.91 (.02)	.03 (.03)	.94 (.01)	.92 (.01)	–.02 (.02)	**.05 (.03)**	1
Prefer spending cuts vs. raise taxes	.41 (.03)	.39 (.03)	–.01 (.05)	.43 (.02)	.37 (.02)	–.06 (.03)**	**.05 (.06)**	1
Know size of projected budget deficit	.36 (.03)	.33 (.03)	–.03 (.05)	.53 (.02)	.49 (.02)	–.04 (.03)	**.01 (.06)**	1
Know Lt. Governor Carroll	.67 (.03)	.73 (.03)	.06 (.05)	.82 (.02)	.82 (.02)	.00 (.02)	**.06 (.05)**	0
Approve of President Barack Obama	.67 (.03)	.66 (.03)	–.01 (.05)	.67 (.02)	.68 (.02)	.02 (.03)	**–.02 (.05)**	0
Know department elimination	.67 (.03)	.66 (.03)	–.01 (.05)	.68 (.02)	.68 (.02)	–.00 (.03)	**–.01 (.05)**	0
Know pledge to cut state workforce	.61 (.03)	.63 (.03)	.02 (.05)	.60 (.02)	.61 (.02)	.01 (.03)	**.01 (.06)**	0
Believe Florida is satisfactory	.30 (.03)	.29 (.03)	–.01 (.04)	.30 (.02)	.28 (.02)	–.02 (.03)	**.01 (.05)**	0
Prefer traffic stops for immigrants	.51 (.03)	.49 (.04)	–.02 (.05)	.54 (.02)	.52 (.02)	–.02 (.03)	**–.00 (.06)**	0

NOTE.—The entries are means with standard errors in parentheses. In some instances, the number of observations may be reduced due to item nonresponse. The cell entries may not sum properly due to rounding. The areas in gray shading denote statistically significant experimental effects in either the lab or field context. The entries in boxes signal a statistically significant difference between the lab and field contexts. DID = The difference-in-differences between the lab and field treatment effects (i.e., $[\text{Treatment}_{\text{Lab}} - \text{Control}_{\text{Lab}}] - [\text{Treatment}_{\text{Field}} - \text{Control}_{\text{Field}}]$). Correspondence denotes the number of treatment condition news articles, out of four in the lab, that covered the topic in question.

**$p < .05$, *$p < .10$ (two-tailed)

of entries shows the field experimental groups, with a control group mean of.26 and a treatment group mean of.27, resulting in a small and statistically insignificant treatment effect (.01). The next column ("DID") reports the difference-in-differences between the lab and the field (i.e., the –.09 lab effect minus the.01 field effect). The DID is –.10 and is statistically significant ($p < .10$, two-tailed).[13] In this particular case, not only is there a significant finding in the lab without a corresponding finding in the field, but the difference between the two treatment effects also is statistically significant.

Considering the six outcomes in which there was a statistically significant lab treatment effect, the DID is significant in four instances (the two gubernatorial approval items, Most Important Issue: Florida Budget, and Prefer Property & Business Tax Cuts). For each of these four questions, the lab effect is larger in magnitude (DID calculations are on the order of 10 points). Thus, in the context of our study, when there was a statistically significant lab effect, this effect tended to be significantly larger in magnitude than the field treatment effect. Viewing table 2 in its entirety, however, the support for H1 is mixed, with only a handful of instances in which there was a significant difference in the treatment effects across the two contexts. Indeed, it was more common to observe null effects in both contexts, a topic we return to below.

VARIATIONS IN TREATMENT-OUTCOME CORRESPONDENCE

Next we look for evidence for the second hypothesis, which states that as the connection between the stimulus and outcome measure becomes closer, differences in treatment effects across the two settings will be more likely to emerge. Operationally, this implies that we should observe significant DIDs for questions having the highest values of *Correspondence Score*.

In all four instances of a significant DID, the question scored above the median on *Correspondence Score* (taking on the highest value in three cases). For example, two items ask about approval of Governor Rick Scott and he received negative coverage in the treatment stories. In the first treatment story, subjects learned that the governor planned to cut more than 8,600 state positions. Likewise, in the fourth story, a local public official was quoted as saying, "I want Rick Scott to … bring jobs to Florida but I don't want him to do it on the backs of state employees." This same story drew an analogy between Florida and states like Wisconsin, where Republican governors were described as "patching budget deficits with major budget cuts that crimp state employee benefits." Finally, Nina Hayden (third story) said that the governor "doesn't seem to relate to working people." In the case of Prefer Property & Business

13. We calculate the DID with Stata 12's ttesti command using the raw aggregate-level data from the lab and field effect columns (boxes denote a significant DID). These patterns are confirmed with individual-level data analyses that estimate separate lab and field treatment coefficients on a stacked data set and test the two treatment coefficients against each other using a Wald test.

Tax Cuts (*Correspondence Score* = 3), several articles criticized the governor's plan to cut property and business taxes. In the third story, Nina Hayden is quoted as saying, "It seems like [the governor] is targeting public servants, and then giving the tax breaks to large corporations." The fourth article made a similar claim, with a reference to an "assault on the middle class in order to give tax breaks to the rich." Treatment subjects in the lab were 13 percentage points less likely to want tax cuts (effect = -.13, $p < .05$, two-tailed), whereas the field effect was smaller and insignificant, resulting in a DID of –.10 ($p < .10$, two-tailed). On the whole, when we observed statistically significant differences in treatment effects across the two contexts, it was also the case that there was a close substantive connection between the stimulus and the particular outcome measure.

The remaining rows in table 2 show the comparisons for questions having the weakest link to the news stories. For several of these items (Obama approval, Lt. Governor Jennifer Carroll, the elimination of the Department of Community Affairs, and immigration), there was no mention of the topic in our treatment stories (meaning that one can view these outcomes as placebos). For these questions there was no treatment effect in either context and the differences between lab and field treatment effects were insignificant, as one would expect.

AGGREGATE-LEVEL HYPOTHESIS TESTS

We summarize the question-by-question results with an analysis that examines whether the lab treatment effects are, on average, larger in magnitude than the field effects. To explore this issue, we predict the absolute value of the 34 treatment effects (17 outcome measures from the lab and another 17 from the field). In these analyses, the lab and field data are combined into a single data set. The key predictors are dummy terms indicating lab condition or not, level of correspondence (which ranges from 0 to 4), and the interaction of these two variables.

According to H1, lab treatment effects will be larger than field treatment effects. We test this hypothesis with the first model in table 3, which includes an indicator for lab condition. The coefficient on this term is positive and statistically significant (coeff. = .03; $p < .05$, two-tailed), implying a bigger treatment effect in the laboratory setting by roughly 3 percentage points.[14]

According to H2, this effect should be conditional on the relationship between the treatment and the outcome measure. The closer the substantive connection, the larger the discrepancy in effect size across the two contexts. The key test of H2 is the interaction between the lab condition dummy and

14. We also bootstrapped the individual-level data to account for the uncertainty around the estimated treatment effects. Our findings converge, showing that the average lab effect was .04 larger than the field effect ($p < .001$, two-tailed; mean = .037 with a 95-percent interval from .036 to .038).

Table 3. Predicting Treatment Effects by Context and Correspondence

	DV = abs (treatment effect)		
	Model 1 (n = 34)	Model 2 (n = 34)	Model 3 (n = 34)
Lab condition	.03 (.01)**	.03 (.01)**	–.00 (.01)
Correspondence score	—	.01 (.00)**	–.00 (.00)
Lab X correspondence	—	—	.02 (.00)**
Constant	.02 (.00)**	.00 (.01)	.02 (.01)**
R-squared	.24	.38	.61

NOTE.—The dependent variable is the absolute value of the treatment effect, treating the lab and field effects as separate observations. Entries are coefficients from ordinary least squares regressions. Standard errors are clustered by question (i.e., each question appears twice, once in the lab and once in the field) and reported in parentheses. The full equation (i.e., for model 3) is: abs(Treatment Effect) = constant + b1*Lab + b2*Correspondence + b3*(Lab*Correspondence). Lab is a dichotomous indicator of a laboratory effect (1 = lab, 0 = field), and Correspondence is a variable that measures the number of treatment articles (0 to 4) that cover the issue raised in each survey question.

**$p < .05$ (two-tailed)

Correspondence Score, which we expect to be positive and statistically significant. We begin with the second column of results, which shows the coefficients from a model with separate terms for lab condition and *Correspondence Score*. In this model, the lab dummy remains positive and statistically significant (coeff. = .03; $p < .05$, two-tailed) and the coefficient on *Correspondence Score* is positive and statistically significant (coeff. = .01; $p < .05$, two-tailed). Model 3 includes the lab dummy, *Correspondence Score*, and the crucial interaction term. There is a small but statistically significant effect of nearly two points (coeff. = .02; $p < .05$, two-tailed). Thus, the various factors that make lab effects larger have the most dramatic effect at high values of *Correspondence Score*—that is, when there is a close relationship between treatment and outcome. The lab dummy, which represents the effect of the laboratory setting when *Correspondence Score* is 0, is near zero and no longer significant. This last finding is consistent with earlier results. When the treatment is unrelated to the outcome measure, we would expect no treatment effects in either context, and hence no difference between experimental modes.[15]

BRIDGING (ONE OF) THE DIFFERENCES BETWEEN LAB AND FIELD

One of the lessons we draw from the preceding analyses is that treatment effects from laboratory studies will be difficult to reproduce in other experimental

15. Similar results obtain when we operationalize the dependent variable as a binary indicator of the presence of a statistically significant effect or a larger effect in either context (these findings are more speculative due to the small number of cases).

settings (especially field experiments) when lab manipulations are the strongest—that is, when it is almost impossible for subjects to miss the relationship between the stimulus and subsequent outcome measures. In this way, mode differences (e.g., impact, obtrusiveness, time distance) may be consequential for the estimation of treatment effects.

Our last series of analyses takes advantage of natural variation on one of these dimensions: the time between stimulus and outcome measure in the field experiment. We expected that differences in treatment effects could be partially explained by the distance, in terms of time, between stimulus and outcome. This proposition was tested by examining field respondents who completed their surveys during the week of the lab experiment. Operationally, this corresponds to people who were *below* the median in terms of the length of time to return their survey ($n = 448$). Recall from table 2 that there were significant differences between lab and field treatment effects for four of our outcome measures. We expected that there would be greater correspondence in the treatment effects for the subset of field respondents who returned their survey early (i.e., those for whom less time had elapsed between stimulus and outcome).[16]

In three out of four cases, the differences between the lab and field moderate when we refine our analysis to look for treatment effects among field respondents who returned their survey within the first week. In particular, the statistically significant differences between the lab and field shown in table 2 for the two gubernatorial approval items and the tax cut proposal *disappear* when we focus on early returners.[17] Thus, some of the starkest differences between lab and field moderate when we concentrate on the subsample experiencing the shortest time between the treatment and measurement of the outcome.[18]

The time analyses also help us explore one of the central limitations of our study—namely, the potential confound between treatment and mode effects. In table 2, we assume that the significant DIDs are the result of *mode* differences relating to the characteristic features of lab and field experiments. An alternative explanation is that the results stem from *treatment* differences across the two settings (e.g., lab subjects received newspaper articles whereas field subjects received a newspaper subscription). If our results can be explained by differences in experimental treatment rather than mode of experimentation, the

16. Here, our results are purely observational since the date of survey completion was not randomized. However, treatment assignment is uncorrelated with date of return. The only factor predicting early return of the field survey was gender (women are less likely to return in the first week, $p < .05$, two-tailed). Controlling for demographic factors does not change the results of this analysis.

17. We obtain this result either by recalculating the means reported in table 2 for the subset of field respondents who returned their survey earlier or via individual-level analyses with an interaction term.

18. Overall, the correlation between the lab and field effect sizes is .21 ($p < .42$, two-tailed); when we restrict our attention to earlier returners in the field, the corresponding value is .43 ($p < .10$, two-tailed).

findings should not become more similar as mode effects are diminished (i.e., when we focus on early returners in the field). The fact that we do observe this pattern undercuts the claim that treatment, rather than mode, differences are driving the observed results.

Discussion

It is well known that the findings from experimental and observational research may arrive at contrasting results (e.g., LaLonde 1986; Ansolabehere, Iyengar, and Simon 1994; Gerber, Green, and Kaplan 2003). There has been renewed attention to this topic in recent years in response to the growth of experimental research (Druckman et al. 2006; see also Barabas and Jerit 2010). In particular, some critics of lab experiments question whether laboratory findings provide a clear indication of the magnitude, *or even the direction*, of real-world causal effects. Previous researchers have attempted to recover lab-induced findings in naturalistic settings (e.g., Ansolabehere, Iyengar, and Simon 1994; Valentino, Traugott, and Hutchings 2002; Arceneaux and Nickerson 2010; Gerber et al. 2011), but no study has compared treatment effects from simultaneous lab and field experiments on the same topic.

Our empirical analyses show that treatment effects were larger in the lab context, and that this discrepancy was most likely to occur when the content of our treatments could be readily linked to an outcome measure. Thus, factors such as control, obtrusiveness, and temporal distance matter most when there is a clear substantive connection between stimulus and outcome. In these situations, we observed statistically significant treatment effects in the laboratory but not in the field. At the same time, some of the lab–field differences dissipate when we focus on the subset of field respondents who most closely resembled lab participants in terms of the distance between stimulus and outcome. This last result, although non-experimental, is important because it suggests that there are conditions under which lab experiments and field experiments can converge upon the same set of empirical findings.

Perhaps because our study is among the first to manipulate the lab–field experience, there was little precedent for many of the decisions we had to make. One of the challenges, already noted, was the need to select treatments that were comparable yet representative of the stimuli commonly used in each experimental setting. Another decision involved the use of a survey to assess outcomes in the lab and in the field. Though several recent field experiments employ surveys (Gerber, Huber, and Washington 2010; Esterling, Neblo, and Lazer 2011; Gerber et al. 2011; Green et al. 2011; Karpowitz et al. 2011; Shaw and Gimpel 2012), behavioral measures are arguably more common. We opted for attitudinal measures because outcomes could be assessed in exactly the same way across the lab and the field. That said, one avenue for future work is to conduct parallel experiments with behavioral measures in both

contexts. Here, however, the opposite problem exists in that only a small number of political behaviors can reasonably be measured in a laboratory setting. Moreover, the "behaviors" most easily measured in the lab context—intent to vote and vote intention—exhibit a substantial amount of conceptual distance from the corresponding behaviors in the field (e.g., turnout, vote choice).

Ordinarily, design choices are a crucial element of the research process. The importance of such choices is amplified when conducting a study with parallel experiments because of the need for comparability across research settings. Notwithstanding these challenges, a sustained program of experimental interventions replicating and extending the work presented here seems essential, if only to establish the external validity of the lab–field differences we report in this study. As Aronson et al. (1990) observe, "Bringing the research out of the laboratory does not necessarily make it more generalizable or 'true'; it simply makes it different" (p. 82). It is this phenomenon—described once as the "interaction between research design and research setting" (Aronson, Wilson, and Brewer 1998, p. 135)—that should concern all researchers, experimenters and non-experimenters alike.

Appendix

LAB EXPERIMENTAL PROTOCOL

Subjects in the lab study were told they were participating "in a study that examines people's political attitudes and behaviors." In the recruitment letter, invitees were instructed to make an appointment through a webpage (a phone number also was provided). The lab study ran from the afternoon of Sunday, March 6, 2011, until the afternoon of Saturday, March 12, 2011. Lab sessions took place at a variety of times. With the exception of the two weekend days and Friday, the lab was open from 9:00 A.M. until 7:00 P.M. Lab hours on the other days were 4:00–6:00 P.M. (Sunday), 8:00 A.M.–3:00 P.M. (Friday), and 9:00 A.M.–12:00 P.M. (Saturday). At the lab, participants were escorted to the computer room where the study was taking place, but a proctor was not present while subjects completed the questionnaire. After giving consent, treatment subjects saw the following message before viewing the newspaper stories and answering the questionnaire: "We're going to begin by showing you a few news articles. Please take a moment to read them." Control subjects proceeded directly to the outcome measures after viewing the consent form.

After answering our questions, subjects responded to some additional (unrelated) items. They also viewed an informational screen about a campus organization for adults over the age of 50. Control-group subjects did not read any of the stories from the *Tallahassee Democrat* (*TD*), but they did answer these other questions and view the informational screen. The entire questionnaire was self-administered, which means that subjects could advance through the screens as quickly or as slowly as they liked.

SAMPLE RECRUITMENT

As noted in the text, the main sample eligibility screens were residence and newspaper-subscribing status. In addition, we excluded faculty members in our college, members of the university's Institutional Review Board, anyone who was not an active voter (because mail sent to their residence by the election supervisor was returned), anyone requiring special assistance at the polls (which might pose complications for participating in the lab experiment), and anyone living in an apartment setting or in a household with more than three voters. Finally, when two voters lived at the same address, we randomly selected one individual. After applying these screens, we used the sampling procedure in Stata to randomly select 12,000 people from the 18,866 who met the above criteria.

RANDOMIZATION TESTS AND SAMPLE SELECTION

Unlike many laboratory studies, we drew our sample from a list of registered voters, which means that we have an array of (pre-treatment) covariates that can be used to characterize randomization and participation in the study. The first two columns of table A1 show multinomial probit estimates that confirm the success of the randomization into various conditions. There were no randomization problems across an array of demographic and voting history covariates (all variables are $p > .10$; the model chi-square p-value is highly *in*significant). Randomization into treatment and control groups in the lab also was successful (all terms are $p > .10$ in the third column of table A1).

Much as one might expect, the decision to come to the lab or to return the field survey was *non*-random. As we noted earlier, however, the factors predicting participation in the field or lab were remarkably similar. The fourth and fifth columns of table A1 show the results of a probit model predicting participation into either arm of the study (i.e., 1 = participation in lab or field). The coefficients in the fourth column represent the effect of a covariate in predicting participation in the lab context; the coefficients in the next column indicate whether these effects were strengthened or weakened in the field. Thus, for example, Republican identifiers were less likely to participate in either arm of our study, whereas people who had participated in previous elections showed the opposite pattern. As we noted earlier, there were some imbalances across the two arms of our study. When it came to the lab experiment, African Americans were less likely to participate than other respondents, and this pattern was reinforced in the field.[19] Democrats and older people were less likely to participate in the lab than other subjects, but the effect of both factors worked in the opposite direction in the field. Because of these compositional differences, we reestimated the original lab versus field comparisons as individual-level models with demographic controls. In those analyses our substantive conclusions remain unchanged and we observe the same pattern of treatment effects.

19. Despite this finding, our sample remained fairly diverse (roughly 25 percent black).

Table A1. Randomization Check and Selection Bias Analysis

	Randomization check			Selection bias analysis	
	Field control	Field treatment	Laboratory treatment	Lab or field participation	Field interaction terms
	n = 12,000		n = 417	n = 12,000	
Female	.03 (.04)	.03 (.04)	.06 (.13)	.06 (.05)	.02 (.07)
Female not available	–.14 (.14)	–.05 (.14)	— #	–.02 (.23)	–.06 (.30)
Black	–.01 (.04)	.04 (.04)	.11 (.16)	–.16 (.06)**	–.21 (.08)**
Hispanic	–.04 (.11)	–.05 (.11)	.45 (.36)	.03 (.15)	.03 (.19)
Race not available	.06 (.18)	.17 (.18)	— #	–.37 (.36)	.08 (.43)
Age	.00 (.00)	.00 (.00)	–.00 (.00)	–.00 (.00)**	.01 (.00)**
Democrat	–.04 (.05)	–.07 (.05)	–.03 (.18)	–.17 (.08)**	.19 (.10)**
Republican	–.04 (.06)	–.07 (.06)	.01 (.20)	–.25 (.08)**	.06 (.11)
Voted 2010	.01 (.05)	.03 (.05)	–.02 (.20)	.43 (.07)**	.10 (.09)
Voted 2010 N/A	–.10 (.22)	–.02 (.21)	–.75 (.73)	.48 (.07)**	–.12 (.37)
Voted 2010 primary	.04 (.05)	.03 (.05)	–.08 (.18)	.29 (.05)**	–.02 (.08)
Voted 2010 Primary N/A	–.09 (.14)	.13 (.13)	.33 (.49)	.14 (.20)	.17 (.25)
Voted 2008	–.01 (.06)	–.01 (.06)	–.41 (.32)	.14 (.11)	.04 (.14)
Voted 2008 N/A	.03 (.11)	.06 (.11)	–.58 (.49)	.22 (.18)	–.08 (.23)
Voted 2008 primary	–.03 (.05)	–.05 (.05)	.25 (.16)	.20 (.06)**	–.01 (.08)
Voted 2008 primary N/A	.03 (.06)	–.01 (.06)	–.15 (.27)	–.04 (.10)	–.16 (.13)
Assigned to field	—	—	—	—	.23 (.17)
Log-likelihood	–12,468.60		–282.89	–4,059.88	
Model X^2 p-value	.99		.66	.00	

NOTE.—The first two columns contain multinomial probit estimates predicting field treatment or control relative to the omitted laboratory condition. The third column shows probit estimates predicting lab treatment assignment vs. lab control assignment (which was done once the lab participants showed up at the lab). The last two columns show probit estimates for the dichotomous dependent variables predicting participation in the lab or field (= 1) relative to those who were invited but did not come to the lab or complete the field survey (= 0). Standard errors are in parentheses. The constant terms for each model have been suppressed for presentation purposes.

= Term dropped by software due to insufficient variation (i.e., always predicts success or failure)

**$p < .05$ (two-tailed)

Finally, there was one unexpected issue related to the sample. Two respondents assigned to the field condition (one treatment, the other control) showed up at the lab and were assigned to the control condition. Even though they were not treated in the lab, we omit these respondents from our analyses in the paper (and, at any rate, only one completed a field survey).

NONCOMPLIANCE

Three potential noncompliance problems surfaced during the study. First, field participants were given the option to cancel the free subscription, and a small number chose to do so (n = 22 out of 3,000). Few of these individuals returned a survey, which means the problem affected only five field participants. Second, there were some people (n = 175) who subscribed to the *TD* on their own during the study period. Once again, very few of these people ultimately returned a survey or came to the lab (n = 36). Moreover, only 16 started the paper before March 6 (the first day of the lab study), and they were spread fairly evenly across the conditions.

Third, we asked a question at the end of our surveys to assess whether individuals were receiving home delivery of the *TD*. This question served as a manipulation check. It also helped us determine whether subjects, all of whom where non-subscribers according to staff members at the *TD*, were getting the paper some other way (e.g., buying it at the store, sharing with friends). Most respondents (73 percent) reported *not* receiving home delivery of the *TD*. There were no differences in responses to this question in the lab versus field conditions ($|t| = .15$, df = 1,550, $p < .88$, two-tailed), or among treatment and control groups in the lab ($|t| = 1.18$, df = 415, $p < .24$, two-tailed). However, and as we expected, those who were in our treatment condition in the field were more likely to say they were receiving home delivery of the *TD* than were controls in the field ($|t| = 3.40$, df = 1,133, $p < .01$, two-tailed). Indeed, in the open-ended portion of the manipulation check, several respondents alluded to the complimentary subscription (e.g., "won a free month of Sunday-only delivery"), which gives us confidence that the free subscriptions were being received.[20] On the whole, noncompliance issues (e.g., not taking the paper when assigned it or getting the paper when assigned to the control) were minimal, and following Gerber et al. (2009), we estimate intent-to-treat (ITT) effects. In other analyses (available upon request), we calculate the complier average causal effect (CACE). Adjusting our analysis to account for noncompliance has little effect on treatment effects in either context, leaving our substantive conclusions unchanged.

QUESTIONNAIRE WORDING

The survey questions were administered only in English, and they were the same in each experiment (both the actual text and the formatting of the response

20. We also confirmed paper delivery at ten randomly selected addresses for field subjects.

options). The only exception is that in the lab experiment, the opening banner ("Public Opinion in Leon County") and introductory paragraph were excluded.

PUBLIC OPINION IN LEON COUNTY

Your opinions are very important to us and this is your chance to voice your concerns. Your individual responses will be kept confidential and the results will only be reported in the aggregate.

1. Do you approve or disapprove of the way Barack Obama is handling his job as President?

Approve → Do you strongly approve or somewhat approve?
- ☐ Strongly approve
- ☐ Somewhat approve

Disapprove → Do you strongly disapprove or somewhat disapprove?
- ☐ Strongly disapprove
- ☐ Somewhat disapprove

2. Do you approve or disapprove of the way Rick Scott is handling his job as Governor?

Approve → Do you strongly approve or somewhat approve?
- ☐ Strongly approve
- ☐ Somewhat approve

Disapprove → Do you strongly disapprove or somewhat disapprove?
- ☐ Strongly disapprove
- ☐ Somewhat disapprove

3. Do you approve or disapprove of the way Rick Scott is handling his job when it comes to improving Florida's economy?

Approve → Do you strongly approve or somewhat approve?
- ☐ Strongly approve
- ☐ Somewhat approve

Disapprove → Do you strongly disapprove or somewhat disapprove?
- ☐ Strongly disapprove
- ☐ Somewhat disapprove

4. In general, how satisfied are you with the way things are going in Florida today?

Satisfied → Are you very satisfied or somewhat satisfied?
- ☐ Very satisfied
- ☐ Somewhat satisfied

Dissatisfied → Are you very dissatisfied or somewhat dissatisfied?
- ☐ Very dissatisfied
- ☐ Somewhat dissatisfied

5. How serious do you think Florida's budget problems are?

Serious → Would you say very serious or somewhat serious?
- ☐ Very serious
- ☐ Somewhat serious

Not serious → Would you say not at all serious or not too serious?
- ☐ Not at all serious
- ☐ Not too serious

6. What do you think is the most important problem for the state government of Florida to address? Please write your response in the space below.

7. **How much of the time do you think you can trust Florida state government to do what is right?**

- ☐ Just about always
- ☐ Most of the time
- ☐ Only some of the time

8. **Governor Scott wants to cut about 2 billion dollars in property and business taxes to make Florida more attractive to business and attract jobs. Others say the state can't afford to lose the revenue. Do you think these tax cuts are a good idea or bad idea?**

- ☐ Good idea
- ☐ Bad idea

9. **Governor Scott says that state workers, who currently are not required to contribute to their retirement funds, should pay part of the cost. Do you think this is a good idea or bad idea?**

- ☐ Good idea
- ☐ Bad idea

10. **To balance the state budget, if you had to choose, would you prefer raising taxes or cutting government programs and services?**

- ☐ Raising taxes
- ☐ Cutting services

11. **Would you support or oppose a law that allows police officers to ask for immigration documentation during routine traffic stops?**

Support → Would you say strongly support or just somewhat support?
- ☐ Strongly support
- ☐ Somewhat support

Oppose → Would you say strongly oppose or just somewhat oppose?
- ☐ Strongly oppose
- ☐ Somewhat oppose

Next, we are going to ask you some factual questions about politics. Not everyone will know the correct answer to these questions, but please answer every question as best you can.

12. **What job or position did Rick Scott hold <u>before</u> being elected as Governor? Was he…**

- ☐ Attorney General for the state of Florida
- ☐ A former member of the Florida state legislature
- ☐ A business man (CORRECT ANSWER)

13. **Governor Rick Scott's budget will almost entirely eliminate one state agency. Do you know which agency this is?**

- ☐ Department of Environmental Protection
- ☐ Department of Community Affairs (CORRECT ANSWER)
- ☐ Department of Business and Professional Regulation

14. **What job or political office is currently held by Jennifer Carroll? Is she the…**

- ☐ Lieutenant Governor for the state of Florida (CORRECT ANSWER)
- ☐ Attorney General for the state of Florida
- ☐ Chief Financial Officer for the state of Florida

15. **There has been talk about the size of the projected state budget deficit. Do you know which is closer to the approximate size of the projected Florida state budget deficit?**

- ☐ $4 billion (CORRECT ANSWER)
- ☐ $6 billion
- ☐ $13 billion

16. **Governor Rick Scott pledged to cut the state workforce. Do you know by approximately what percentage he pledged to cut the state workforce?**

- ☐ 1-2%
- ☐ 6-7% (CORRECT ANSWER)
- ☐ 15-16%

17. **How closely have you been following stories about the upcoming legislative session? Would you say you've been FOLLOWING STORIES ABOUT THE SESSION . . .**

- ☐ Very closely
- ☐ Fairly closely
- ☐ Not too closely
- ☐ Not at all closely

18. **Do you <u>currently</u> receive home delivery of the *Tallahassee Democrat* newspaper?**

- ☐ Receive home delivery every day
- ☐ Receive home delivery only on Sunday
- ☐ Do not receive home delivery
- ☐ Other *[For example, read online, buy at store, etc.] Please Specify:* _____

19. **During a typical week, how many <u>days</u> do you watch news about local politics on television?**
[Please circle the number of days]

0 1 2 3 4 5 6 7

Background Information

What is the last grade or class that you completed in school?

- ☐ Eight years or less
- ☐ 9-11 years
- ☐ Completed high school
- ☐ Business or technical school
- ☐ Some college
- ☐ Completed college
- ☐ Graduate or professional school

Gender:

- ☐ Male
- ☐ Female

In what year were you born? _____

What is your occupation? _____

What is your race/ethnicity?

☐ White, not Hispanic

☐ Black, not Hispanic

☐ Hispanic

☐ Asian/ Pacific Islander

☐ American Indian/ Alaskan Native

THANK YOU FOR YOUR PARTICIPATION

References

Albertson, Bethany, and Adria Lawrence. 2009. "After the Credits Roll: The Long-Term Effects of Educational Television on Public Knowledge and Attitudes." *American Politics Research* 37:275–300.

Ansolabehere, Stephen D., Shanto Iyengar, and Adam Simon. 1994. "Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout." *American Political Science Review* 93:901–9.

Arceneaux, Kevin. 2010. "The Benefits of Experimental Methods for the Study of Campaign Effects." *Political Communication* 27:199–215.

Arceneaux, Kevin, Martin Johnson, and Chad Murphy. 2012. "Polarized Political Communication, Oppositional Media Hostility, and Selective Exposure." *Journal of Politics* 74:174–86.

Arceneaux, Kevin, and Robin Kolodny. 2009. "Educating the Least Informed: Group Endorsements in a Grass Roots Campaign." *American Journal of Political Science* 53:755–70.

Arceneaux, Kevin, and David Nickerson. 2010. "Comparing Negative and Positive Campaign Messages: Evidence from Two Field Experiments." *American Politics Research* 38:54–83.

Aronson, Elliot, Phoebe C. Ellsworth, J. Merrill Carlsmith, and Marti Hope Gonzales. 1990. *Methods of Research in Social Psychology*. 2nd ed. New York: McGraw-Hill.

Aronson, Elliot, Timothy D. Wilson, and Marilynn B. Brewer. 1998. "Experimentation in Social Psychology." In *The Handbook of Social Psychology*, edited by Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, 99–142. 4th ed. Boston: McGraw-Hill.

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

Benz, Matthew, and Stephen Meier. 2008. "Do People Behave in Experiments as in the Field? Evidence from Donations." *Experimental Economics* 11:268–81.

Berinsky, Adam J., and Donald R. Kinder. 2006. "Making Sense of Issues Through Media Frames: Understanding the Kosovo Crisis." *Journal of Politics* 68:640–56.

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

Chong, Dennis, and James Druckman. 2010. "Dynamic Public Opinion: Communication Effects over Time." *American Political Science Review* 104:663–80.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research Political Science." *American Political Science Review* 100:627–35.

Druckman, James N., and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Database.'" In *Handbook of Experimental Political Science*, edited by James Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, 41–57. New York: Cambridge University Press.

Druckman, James N., and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47:729–45.

Esterling, Kevin M., Michael A. Neblo, and David M. J. Lazer. 2011. "Means, Motive, and Opportunity in Becoming Informed about Politics: A Deliberative Field Experiment with Members of Congress and Their Constituents." *Public Opinion Quarterly* 75:483–503.

Field, Andy, and Graham Hole. 2003. *How to Design and Report Experiments*. New York: Sage.

Gaines, Brian J., and James Kuklinski. 2011. "Experimental Estimation of Heterogeneous Treatment Effects for Treatments Related to Self-Selection." *American Journal of Political Science* 55:724–36.

Gerber, Alan. 2011. "Field Experiments in Political Science." In *Handbook of Experimental Political Science*, edited by James Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, 115–40. New York: Cambridge University Press.

Gerber, Alan S., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. "How Large and Long-Lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105:135–50.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2003. "The Illusion of Learning from Observational Research." In *Problems and Methods in the Study of Politics*, edited by Ian Shapiro, Rogers M. Smith, and Tarek E. Masoud, 251–73. New York: Cambridge University Press.

Gerber, Alan S., Donald P. Green, and Chris W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102:33–48.

Gerber, Alan, Gregory Huber, and Ebonya Washington. 2010. "Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment." *American Political Science Review* 104:720–42.

Gerber, Alan, Dean Karlan, and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics* 1:35–52.

Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95:379–96.

Green, Donald P., Peter M. Aronow, Daniel E. Bergan, Pamela Greene, Celia Paris, and Beth I. Weinberger. 2011. "Does Knowledge of Constitutional Principles Increase Support for Civil Liberties? Results from a Randomized Field Experiment." *Journal of Politics* 73:463–76.

Iyengar, Shanto. 2011. "Laboratory Experiments in Political Science." In *Handbook of Experimental Political Science*, edited by James Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, 73–88. New York: Cambridge University Press.

Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.

Karpowitz, Christopher F., J. Quin Monson, Lindsay Nielson, Kelly D. Patterson, and Steven A. Snell. 2011. "Political Norms and the Private Act of Voting." *Public Opinion Quarterly* 75:659–85.

Kinder, Donald. 2007. "Curmudgeonly Advice." *Journal of Communication* 57:155–62.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76:604–20.

Levitt, Steven D., and John A. List. 2007. "Viewpoint: On the Generalizability of Lab Behaviour to the Field." *Canadian Journal of Economics* 40:347–70.

McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science* 5:31–61.

Mitchell, Gregory. 2012. "Revisiting Truth or Triviality: The External Validity of Research in the Psychological Laboratory." *Psychological Science* 7:109–17.

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

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

Mutz, Diana, and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *American Political Science Review* 99:1–16.

Nelson, Thomas E., Rosalee Clawson, and Zoe Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91:567–83.

Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13:233–52.

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

Shaw, Daron R., and James G. Gimpel. 2012. "What If We Randomize the Governor's Schedule? Evidence on Campaign Appearance Effects from a Texas Field Experiment." *Political Communication* 29:137–59.

Singer, Eleanor, and Felice J. Levine. 2003. "Protection of Human Subjects of Research: Recent Developments and Future Prospects for the Social Sciences." *Public Opinion Quarterly* 67:148–64.

Valentino, Nicholas A., Michael Traugott, and Vincent L. Hutchings. 2002. "Group Cues and Ideological Constraint: A Replication of Political Advertising Effects Studies in the Lab and in the Field." *Political Communication* 19:29–48.

Webb, Eugene J., Donald T. Campbell, Richard D. Schwarz, and Lee Sechrest. 2000. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.