Comparing Contemporaneous Laboratory and Field Experiments on Media Effects

Jennifer Jerit
Jason Barabas
Scott Clifford

Department of Political Science
Florida State University
Tallahassee, FL 32306

June 22, 2012

Abstract

Experimenters 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). We develop two hypotheses concerning the relationship between treatment effects in lab and field settings that are tested in contemporaneous experiments. In particular, registered voters in a medium-sized city were assigned to a laboratory or a field experiment involving newspaper treatments. 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. There also is suggestive evidence that as lab and field experiments become similar on one particular dimension (the temporal distance between stimulus and outcome), differences in the size of treatment effects moderate.

Word Count: 6482
Experiments are powerful because they demonstrate cause and effect in an especially compelling way. Perhaps as a result, the use of randomized experiments, particularly lab experiments, is on the rise.\(^1\) Notwithstanding this trend, some argue that the findings from lab experiments have limited external validity because of sample characteristics (e.g., the frequent use of undergraduates as research subjects) or the artificial nature of the study setting (Iyengar 2011; Morton and Williams 2010; McDermott 2002). In response to the concerns about potentially unrepresentative convenience samples often used in lab settings, a growing number of scholars employ survey experiments with diverse adult subject populations (e.g., Brooks and Geer 2007, 3; Gilens 2001, 382; Sniderman and Grob 1996). 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 topics traditionally examined in the lab—most notably, several recent studies using mass media and campaign treatments (e.g., Albertson and Lawrence 2009, Arceneaux and Nickerson 2010; Arceneaux and Kolodny 2009, Gerber, Karlan, and Bergan 2009; Gerber, Gimpel, Green and Shaw 2011; Washington, Gerber, and Huber 2010). In these studies and others like them, the rationale for a field study is the importance of examining political phenomena in a naturalistic setting. Lab experiments have impressive levels

\(^1\) For example, in the *American Political Science Review*, two-thirds of the articles involving experiments took place in a laboratory or classroom (Druckman et al. 2006, 633).
of internal validity, the argument goes, but the empirical findings that emerge 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 important differences (either in the timing of lab and field studies or the stimuli) that limit the conclusions one can draw. In fact, no existing study in any social science discipline 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 vs. field experience. Drawing from the same target population (adult registered voters in Florida), we administer archetypical but 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. The stakes of this comparison are substantial, both from a methodological and a practical standpoint. Experiments vary in their cost and ethical considerations. Insofar as scholars reach similar conclusions in lab and field settings, future researchers might devote more resources to lab experiments, which often can be administered more quickly and at less financial cost. In this way, our study has implications for the interpretation of experimental evidence as well as extraordinary practical relevance for the way people conduct research.

**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, 276-7). Likewise, Gerber, Karlan, and Kaplan (2009) examine the effect of newspapers on political attitudes and behavior in a field experiment in the Washington D.C. 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” (2009, 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, 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., Gerber 2011).

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 than the effects that would be observed if the same study were conducted in the field (e.g., Levitt and List 2007). For example, Gerber 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…” (2011, 120, emphasis original).

---

2 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, 264-65; Mutz 2011, 134-5).
This is an empirical question, however, and one that can be answered only with the accumulation of evidence from parallel experiments. 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 below by outlining 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 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, treated 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., Gaines and Kuklinski 2011; Iyengar 2011; Kinder 2007). This aspect of control is one of the primary differences between lab and field experiments, as the latter often rely on outside organizations to administer
the treatment. \(^3\) 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, there are few distractions in laboratory settings 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 on 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 laboratory, but not in field, experiments (Singer and Levine 2003). Thus, participants in a lab experiment know they are being studied, and this awareness may cause demand characteristics or peer influence behaviors to confound the effect of the treatment (e.g., Hawthorne effects, compensatory rivalry, resentful demoralization; see Shadish, Cook, and Campbell 2002 for more). \(^4\) 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. We

---

\(^3\) In other words, there can be failure to treat problems regarding *delivery*; i.e., the researcher’s ability to distribute the treatment accurately. Failure to treat also can occur with the *receipt* of the treatment and *adherence* to treatment assignment (Nickerson 2005).

\(^4\) Gerber, Green, and Larimer (2008)’s study of social pressure is a notable exception. As with most field experiments, informed consent was not obtained. However, as part of some of the treatments, participants were told that their voter participation was being studied.
surmise that these situational factors will cause subjects to pay greater than usual attention to the stimuli in laboratory settings.

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 terms of minutes or, possibly, hours. By contrast, in field settings, days or even weeks may pass in between the application of the stimulus and the measurement of the outcome. This difference matters because as the time that passes between treatment and outcome increases, treatment effects are likely 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.

Hypotheses

As a result of differences in impact, obtrusiveness, and time distance, the treatment effects from a lab experiment may not correspond to the findings that would be observed in a more naturalistic setting. Indeed, the “behavioral latitude” of the field may dilute, magnify, or even reverse the findings from a laboratory setting (Gerber 2011, 120). Thus, one basic hypothesis emerging from the literature is that treatment effects from lab experiments will be different than those that would be observed in a comparable field study (H1). Naturally, one may operationalize “different” in various ways, such as the magnitude and/or direction of treatment effects. For now, we simply articulate the general critique against lab experiments and explore several operational definitions of this construct later, in the empirical section.

---

5 There is variation in lab and field experiments on this dimension. Recent lab studies have examined the duration of experimental effects, which implies long-term measurement of outcomes (e.g., Mutz and Reeves 2005; Chong and Druckman 2010). Generally speaking, the time that passes between stimulus and outcome is shorter in the lab than in the field.
One of the characteristic features of a randomized experiment, irrespective of where it takes place, is that treatment is exogenous. 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 stimuli appear by itself, say in a single paragraph, or is it embedded in a longer news story or series of stories?) as well as semantic features of the treatment (e.g., does the stimulus imply a particular response or must subjects make an inference?). We expect that the closer the connection between a stimulus and an outcome, the more likely lab and field treatment effects will be different from one another (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, the greater impact of lab stimuli and the (generally) shorter time distance between stimulus and outcome promote the activation of relevant underlying attitudes. Put another way, factors such as impact, obtrusiveness, and time will make little difference in the estimation of treatment effects if the correspondence between the stimulus and the outcome measures is distal, or absent altogether.

Study Design

The purpose of our design was to deliver information exogenously in 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 in more detail below, we went to great lengths to make sure 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 hold 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 treatments reflect the defining features of each style of experimentation (e.g., in terms of impact, time distance). Thus, the lab and field studies were designed to be “typical” of experiments in each area. This aspect of our design was essential. As one of the first studies to assess the congruence between treatment effects in the lab and the field, it was particularly important that we administer characteristic experiments of each type. Whatever differences in treatment effects are observed—if any—can then become the object of future study, with researchers manipulating aspects of the experimental context in one or both settings.

In terms of the substance of the design, we implemented a canonical political communications experiment with a single treatment and control group. The lab study was modeled after past experiments in which treated subjects are exposed to actual media reports or faux stories that are realistically inspired (e.g., Berinsky and Kinder 2006; Iyengar and Kinder 1987; Nelson, Clawson, and Oxley 1997). 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 point in time. In the empirical analyses that follow, 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 using samples drawn from the same population. This design enables us to determine if the findings 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. Participants were drawn from a list of registered voters, obtained from the Leon County Supervisor of Elections Office in February of 2011. Working with staff members at the Tallahassee Democrat—the only local major newspaper serving the area—we identified 98,862 people in Leon County who were not current subscribers to the paper from a list of more than 171,000 voters. From this list, we selected individuals who were located in voting precincts within a five mile radius of the city center, where the experimental lab is located.

---

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

7 We did this in order to boost our response rate in the lab experiment. Removing remote areas of the county also helped ensure delivery of the newspapers in the field. The main sample eligibility screens were residence and newspaper subscribing status, but we also excluded faculty members in our college, members of the university’s Internal Review Board, anyone who was not an active voter (i.e., mail sent to their residence by the election supervisor was returned), anyone
Participants were randomly assigned to be in the field experiment (n=6,000) or the laboratory experiment (n=6,000). In the field experiment, 3,000 households were randomly assigned to receive a one-month Sunday only subscription to the *Tallahassee Democrat* beginning on Sunday, February 27, 2011. Along with the first free Sunday paper, households received a flyer telling them that they had won a “complementary one-month subscription to the *Tallahassee Democrat*” in a “drawing.” The purpose of the flyer was to inform treated subjects about their free subscription and to discourage them from drawing a connection between the subscription and the post-treatment questionnaire. Three thousand people were randomly assigned to the control group; i.e., these individuals did not receive a subscription of any type. We sent a mail survey to all 6,000 people in the field experiment the week of March 7, two weeks after their free subscription began. As an inducement to participate the cover letter indicated there would be a raffle for two $300 awards for those who returned their survey.

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 requiring special assistance at the polls (which might pose complications for participating in the lab experiment), or anyone living in an apartment setting. Finally, when more than one person lived at an address, we randomly selected one individual. After applying these screens, we randomly selected 12,000 from 18,866 who met the above criteria.
computer, upon arrival).\(^8\) Not only did participants in the two arms of our study answer the same questions, but they also were answering the questions over roughly the same time period.\(^9\)

_Treatment and Outcome Measures_

In this study treatment is the presentation of newspaper articles. Thus, in the lab treated subjects were presented with several stories about local politics from the _Tallahassee Democrat_. In the field, treated subjects received complimentary newspapers containing those same stories.\(^10\)

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.\(^11\) Treated subjects in the field experiment received two Sunday papers (on February 27 and March 6) before they got our mail survey. Treated subjects in the laboratory experiment read the four most prominent stories about state politics from the February 27 and March 6 papers. This included the front page story from each paper, as well as two visible interior stories from the March 6 issue. To the extent possible, we maximized the similarity of the treatments across the two contexts. In terms

---

\(^8\) See the Appendix for additional details regarding the experimental protocol.

\(^9\) Approximately 40% of field participants returned their survey during the week of the lab study.

\(^10\) Naturally, treated subjects in either context could choose not to read the newspaper articles.

\(^11\) Tallahassee is the state capitol of Florida. Each year, the spring legislative session begins on the first Tuesday following the first Monday in March. We examined past coverage in the _Tallahassee Democrat_ and found that the paper usually previewed issues facing the legislature in the Sunday issue preceding the start of session (which in our case was Sunday, March 6). The 2011 session was expected to be particularly newsworthy because the state’s newly elected governor faced a $3.6 billion budget deficit.
of content, we included the most salient stories about local politics across the two issues.\textsuperscript{12} When it came to appearance, the treatments in the lab were screen shots of the \textit{Tallahassee Democrat} so they were exact replicas of the stories appearing in the newspaper. Table 1 describes the central features of each article (the actual treatments are shown in a Supplementary Appendix).

Table 1 about here.

As for the outcome measures, we 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 Supplementary Appendix, but in brief, we examined items on job approval, knowledge, policy preferences, and attention to local politics.\textsuperscript{13} Given the nature of the project (and the necessity of receiving Human Subjects Committee approval), the questionnaire had to be designed well 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 \textit{Tallahassee Democrat}—we consulted with members of the local media and other organizations.

\textsuperscript{12} Treated subjects in the field received two additional Sunday papers (on March 13 and March 20) as part of their complimentary subscription. However, auxiliary content analyses indicate that there was virtually no coverage of the topics on our questionnaire in those later newspapers.

\textsuperscript{13} We chose these particular outcome measures for several reasons. From a theoretical standpoint, it is important to examine what happens to knowledge and attitudes when people are exposed to information about politics from the mass media. Indeed, this general style of treatment has been employed in scores of lab experiments and several recent field experiments. From a practical standpoint, these particular outcomes could be measured the \textit{same} way in both contexts.
that regularly conduct surveys in Florida.\textsuperscript{14} We were relatively successful, though there is variation in the degree to which our real-world treatments addressed the topics on the questionnaire. This variation will allow us to test our second hypothesis regarding the outcomes on which the lab and the field are most likely to diverge. In the analyses below, we create a variable called \textit{Correspondence Score}, which represents the number of treatment stories (0 to 4) that were related to a question.\textsuperscript{15}

\textbf{Response Rates and Randomization}

Figure 1 gives an overview of the study design and information regarding response rates.

We randomized 12,000 individuals into one of the two arms of our study (lab or field). Of the 6,000 people who were sent a mail survey in our field experiment, 19\% 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\% response rate.\textsuperscript{16} Despite differences in the response

\textsuperscript{14} We also studied the Governor’s proposed budget, which was released several weeks prior to the start of the spring legislative session.

\textsuperscript{15} 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 \textit{Correspondence Score}. In contrast, none of the articles mentioned immigration, resulting in a value of 0 on the measure. Two coders, working separately, read the full text of the treatment stories and assigned correspondence ratings (Krippendorff’s alpha = .72).

\textsuperscript{16} In the field, 653 surveys were not delivered; in the lab study, 251 invitations were returned.

We used the American Association for Public Opinion Research (AAPOR)'s Response Rate 3 for
rate across the two settings, the factors predicting participation in the field or lab study are remarkably similar (see Appendix for details).\textsuperscript{17}

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 a slightly higher concentration of African-Americans and lower number of 2010 voters in the treatment condition of the field experiment who returned a survey ($p < .05$; see the Appendix for more details on the randomization and participation analyses). As a precaution, we confirmed that the patterns presented below hold with controls for demographic factors and household size.

**Empirical Results**

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.\textsuperscript{18} In the both calculations. 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 (and, at any rate, only one completed a field survey).

\textsuperscript{17} The higher field response rate is due to the fact that while non-blacks were more likely to take part in either arm of the study, this effect was especially strong in the field. In addition, Democrats and older people were significantly less likely to participate in the lab relative to Independents and younger people. Once we account for these patterns, differences in the response rates across the two settings become indistinguishable (see Table A1).
first six columns of Table 2, we present the condition means in each experimental setting as well as the treatment minus control (T – C) differences, which appear under the “Effect” columns. We describe the entries under the “DID” and “Correspondence” columns as we get to those results. In order to facilitate comparisons across the variables in our analysis, the outcomes are dichotomized, but in all instances the same patterns obtain in analyses with the original measures.

Table 2 about here.

For many of the outcomes in Table 2, public opinion looks similar across the lab and the field. For example, about two-thirds of our participants approve of President Obama, regardless of whether they participated in the lab experiment or the field experiment. This pattern reflects the political views of the surrounding area where 56% of Leon County residents are registered Democrats and 28% are Republicans. Irrespective of the experimental setting, about a third of the sample expressed approval of Republican Governor Rick Scott, and about 9 in 10 knew that he was a businessman before taking office. Likewise, subjects in both settings reported low levels of trust in state government and satisfaction with the way things are going in Florida. This similarity across the lab and the field is important because it suggests that the same type of

18 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. Doing this allows us to estimate the same quantity across the lab and the field (see Cook, Shadish, and Wong 2008, 728 for discussion). In other analyses, we calculate the complier average causal effect (CACE) (see the Supplementary Appendix for details). Adjusting our analysis to account for noncompliance has little effect on treatment effects in either context, leaving our substantive conclusions unchanged.
person chose to participate in each arm of our study (a point confirmed via statistical analysis; see Table A1).

At the same time, our central question is whether treatment effects differ across the two experimental settings. On this score, there are six instances (appearing in the first six rows of Table 2; see gray shading) in which the lab effect is statistically significant (at $p < .10$ or better), while the field effect is null. By contrast, there is only one case in which the field effect is statistically significant and the lab effect is null. Based simply on a count of significant findings, the lab context differs from the field, a pattern that is broadly consistent with Hypothesis 1.

However, a more compelling approach entails the comparison between the treatment minus control differences in each context—i.e., the difference-in-differences (DID). That quantity appears in the seventh column of Table 2 and it is represented by the following calculation: $\left[\text{Treatment}_{\text{Lab}} - \text{Control}_{\text{Lab}}\right] - \left[\text{Treatment}_{\text{Field}} - \text{Control}_{\text{Field}}\right]$. To illustrate, beginning with the first entry (Job Approval-Governor Rick Scott), approval in the baseline condition is .32, which means that about a third of untreated lab subjects approved of the job the Governor was doing. Moving across the table, approval drops to .23 for treated subjects, resulting in a .09 difference that is statistically significant ($p < .05$, two-tailed $t$-test). The next set 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 it is statistically significant ($p < .10$, two-tailed).

19 In this particular case, not only is there a significant finding in the lab without a

---

19 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
corresponding finding in the field, but the difference between the two treatment effects also is statistically significant.

Returning to the six outcomes in which there was a statistically significant lab treatment effect, the DID is significant in four instances (the two Approval items, Most Important Issue—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), and in three of the four instances the direction of the lab effect is different than the field effect. Thus, in the context of our study, when there was a statistically significant lab effect, the treatment effect tended to be significantly different than the field effect and it was larger in magnitude. However, differences in the direction of treatment effects were rare and were driven largely by field effects that hovered around zero. From this perspective, then, 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 explore in further detail below.

Variations in Treatment-Outcome Correspondence

In the next series of analyses, we look for evidence for our second hypothesis, which states that as the connection between the stimulus and outcome measure becomes closer, difference 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 (which appears in descending order in the final column of Table 2).

with individual-level data analyses that estimate separate lab and field treatment coefficients on a stacked dataset and then test the two treatment coefficients against each other using a Wald test.
In four of the five instances of a significant DID, the question scored above the median on *Correspondence Score* (taking on the highest value in three cases). For example, the approval items explicitly mention 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 to help eliminate the budget deficit. Likewise, in the fourth story, the Democratic State Representative for Leon County is 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 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.” Not too surprisingly, treated subjects in the lab were 13 percentage points less likely to want tax cuts (effect= -.13, *p* < .05) while the field effect was a smaller and insignificant, resulting in a DID of -.10 (*p* < .10). 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 that had 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
simply 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.

Before moving on to the aggregate-level tests of our hypotheses, we discuss two other noteworthy patterns in Table 2. The first occurs on an item asking what job Rick Scott held before being elected Governor (answer: he was a businessman). There is a significant DID even though the value on Correspondence Score was low, a pattern that runs counter to H2. As Table 2 shows, treated subjects in the lab and in the field moved in opposite directions on this outcome. Even though neither of these effects is particularly large or reaches statistical significance, the difference between the lab treatment and field treatment coefficients is significant ($p = .10$).

The second notable finding occurs on Prefer Spending Cuts vs. Raise Taxes, which was a forced choice item about how best to balance the state budget (1=cut spending). Treated subjects move in the same direction (against spending cuts) in both contexts, though here the field effect is statistically significant while the lab effect is not. This last finding may seem surprising but in this case, there is reason to believe that people in the field experiment received a stronger treatment than subjects in the lab. More specifically, spending cuts figured prominently in a story from the Sunday, March 6 issue that was *not* included as one of our four lab treatment stories. The title of this story was “Plans to cut funds for Department of Corrections frustrate Dee Christensen,” and it appeared on page A7, directly across from the “Pension” and “Nina Hayden” stories. Much like the “Nina Hayden” story, which featured episodic coverage of an African American public defender, the “Dee Christensen” article featured a Department of Corrections employee who commented on the Governor’s plans to cut the prison system budget in half. In this particular case, the field treatment may have been stronger, if only because this
particular story was not shown as part of our lab treatment (i.e., there was no exposure in the lab). At any rate, the difference between the lab and field treatment effects is insignificant.

*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 dataset. 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.

Table 3 about here.

According to H1, lab treatment effects will be different 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; \( p < .05 \)), implying a bigger treatment effect in the laboratory setting by roughly 3 percentage points.\(^{20}\)

According to Hypothesis 2, 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 *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

\(^{20}\) We also performed an analysis in which we bootstrap 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 \( (p < .001) \).
this model, the lab dummy remains positive and statistically significant (coeff. = .03; p < .05) and the coefficient on Correspondence Score is positive and statistically significant (coeff. = .01; p < .05). 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).

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 not expect treatment effects in either context, and hence no difference between the two experimental settings.21

* 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 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 making this observation, we do not mean to suggest that treatment effects from laboratory settings are artifacts of the experimental context. They *are* real. However, there are several ways in which the lab and field context differ (e.g., impact, obtrusiveness, the time that elapses between stimulus and outcome) that are consequential for the estimation of treatment effects.

---
21 We obtain similar results when we operationalize the dependent variable as a binary indicator of either the presence of a statistically significant effect or a larger effect in the lab.
Our last series of analyses take advantage of natural variation on one of these dimensions, the time between stimulus and outcome measure in the field experiment. We expect 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 five 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).

In three out of five 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 include an interaction term for the early return of the survey among those who were treated in the field (results not shown). More importantly, on the two approval items, the coefficient on the interaction term is negatively signed and statistically significant (p < .05)—

---

22 We obtain this result either by re-calculating 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 representing the product of the field treatment indicator and a dummy term for respondents returning their survey early. These analyses yield an insignificant DID (i.e., Wald test) between the coefficient on the lab treatment indicator and the coefficient on the early-returner interaction term.
exactly the pattern we observed for the coefficient on the lab treatment indicator. For the tax cuts item, the coefficient on the interaction term shares the same sign as the lab treatment indicator (i.e., it is negative), but it is not statistically significant.\textsuperscript{23} Thus, some of the starkest differences between lab and field moderate when we concentrate on the subsample that experienced the least time between the treatment and measurement of the outcome.\textsuperscript{24}

What do we make of the remaining differences between the lab and the field? Greater attention to the stimulus in the lab probably contributed to larger treatment effects. Indeed, reaction timers in the lab reveal that the vast majority of subjects spent considerable time viewing the stories. However, like many previous field experiments, we are unable to measure attention to the stimulus in the field.\textsuperscript{25} Differential attention across contexts could be due to the obtrusiveness of the laboratory setting, the greater impact of the treatment resulting from forced

\textsuperscript{23} For the other two outcomes (knowledge and naming the budget as the most important issue), the difference between the treatment effect in the lab and the treatment effect among early returners in the field remains statistically significant.

\textsuperscript{24} Overall, the correlation between the lab and field effect sizes is .20; when we restrict our attention to earlier returners in the field the corresponding value is .31. 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, \( p < .05 \)). Controlling for demographic factors does not change the results of the interaction analysis.

\textsuperscript{25} Treated subjects in the field were more likely to state that they were “following stories about the upcoming legislative session” (\( p < .05 \)) relative to control group subjects, giving us some confidence that people were paying attention to the free newspapers.
exposure, or the combination of both factors. The value of conducting additional parallel experiments lies in the promise of (one day) being able to isolate the effect of particular contextual factors (also see Krupnikov and Levine 2011).

**Discussion**

It is well known that the findings from experimental and observational research may arrive at contrasting results (e.g., Ansolabehere, Iyengar, and Simon 1994; LaLonde 1986; 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). 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; Arceneaux and Nickerson 2010; Gerber et al. 2011; Valentino, Traugott, and Hutchings 2002), but no study has compared treatment effects from simultaneous lab and field experiments on the same topic.

Our empirical analyses show that treatment minus control differences 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. This suggests that 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. However, we find little evidence that the field context reverses the effects found in a lab setting. And auxiliary analyses showed that 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 will
converge upon the same set of empirical findings.26

Perhaps because our study is the first to manipulate the lab/field experience, it has raised
more questions than it has answered. For example, if experimental control is the hallmark of a
laboratory experiment, what happens when we lessen that control—or, alternatively, increase the
degree of control in a field experiment? Similarly, what happens to the size of lab treatment
effects when we treat obtrusiveness as a factor to be manipulated? Finally, at what temporal
distance do lab treatment effects start to resemble the effects from a field experiment? Answering
these types of questions is necessary for the development of experimentation as a method in
political science. But doing so will require a sustained program of experimental interventions
that replicate and extend the work presented here, if only to establish the external validity of the
lab-field differences we report in this study. As Aronson et al. (1990, 82) observe, “Bringing the
research out of the laboratory does not necessarily make it more generalizable or ‘true’; it simply
makes it different.” It is this phenomenon—described once as the “interaction between research
design and research setting” (Aronson, Wilson, and Brewer 1998, 135)—that should concern all
researchers, experimenters and non-experimenters alike.

26 This conclusion is limited to the subset of topics for which it makes sense to conduct either a
lab or a field experiment. Some questions are best explored with one style of experimentation.
Appendix

This appendix provides information about lab experimental protocol, randomization, sample selection, and noncompliance. A separate Supplementary Appendix describes the bootstrapping in more detail, shows the news stories, and provides the questionnaire wording.

Lab Experimental Protocol

As described earlier, 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 was also provided). The lab study ran from the afternoon of Sunday, March 6 until the afternoon of Saturday, March 12. 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 am until 7:00 pm. Lab hours on the other days were: 4:00-6:00 pm (Sunday), 8:00 am-3:00 pm (Friday), and 9:00 am-12:00 pm (Saturday). 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.

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 $p > .10$). Randomization into treatment and control 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 (see note 17), 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 and this pattern was reinforced in the field. Democrats and older people were less likely to participate in the lab, with the effect of both factors working in the opposite direction in the field. Because of these compositional differences, we re-estimated the original lab vs. field comparisons from Table 2 as individual-level models with demographic controls. We also estimated a Heckman selection model to adjust the estimates for non-randomness in the decision to participate. In both cases, our substantive conclusions remain unchanged and we observe the same pattern of treatment effects.

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

27 Despite this finding, our sample remained fairly diverse (i.e., 23 to 28 percent black).
do so \((n=22\) out of 3,000). Few of these individuals returned a survey, which means the problem only affected 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%) reported not receiving home delivery of the TD. There were no differences in responses to this question in the lab vs. field conditions \((|t| = .09, \text{df}=1,684, p < .93)\), or among treatment and control groups in the lab \((|t| = 1.18, \text{df}=415, p < .24)\). 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 controls in the field \((|t| = 3.41, \text{df}=1,134, p < .01)\). Indeed, in the open-ended portion of the manipulation check, several respondents alluded to the complementary subscription (e.g., “won a free month of Sunday only delivery”), which gives us confidence that the free subscriptions were being received.28 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, Karlan, and Bergan (2009), we estimate intent-to-treat (ITT) effects. In the Supplementary Appendix, 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.

---

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


Druckman, James N., and Cindy D. Kam. 2011. “Students as Experimental Participants: A Defense of the ‘Narrow Database.’” In James Druckman, Donald P. Green, James H.


Figure 1. Research Design Overview

$n = 12,000$ Tallahassee Voters
Randomly Assigned to:

- Randomized Field Experiment ($n = 6,000$ recruited)
- Randomized Lab Experiment ($n = 6,000$ recruited)

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td>$n = 555$</td>
<td>$n = 580$</td>
</tr>
<tr>
<td></td>
<td>(completed survey)</td>
<td>(completed survey)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab</td>
<td>$n = 198$</td>
<td>$n = 219$</td>
</tr>
<tr>
<td></td>
<td>(came to lab)</td>
<td>(came to lab)</td>
</tr>
</tbody>
</table>
### Table 1. Summary of Laboratory Treatments

<table>
<thead>
<tr>
<th>Date</th>
<th>First Story</th>
<th>Second Story</th>
<th>Third Story</th>
<th>Fourth Story</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td>Sunday, March 6</td>
<td>Sunday, March 6</td>
<td>Sunday, March 6</td>
<td>Sunday, February 27</td>
</tr>
<tr>
<td><strong>Headline</strong></td>
<td>&quot;State Employees haven't had this much at state since 1950s&quot;</td>
<td>&quot;Pension hot topic for 2011 session&quot;</td>
<td>&quot;Nina Hayden: Passion for public service helps balance reduced pay&quot;</td>
<td>&quot;Pro-union rally held at Capitol&quot;</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>Front page (A1); above fold</td>
<td>Page A6</td>
<td>Page A6</td>
<td>Front page (A1); above fold</td>
</tr>
<tr>
<td><strong>Length</strong></td>
<td>18 paragraphs</td>
<td>9 paragraphs</td>
<td>12 paragraphs</td>
<td>17 paragraphs</td>
</tr>
<tr>
<td><strong>Images</strong></td>
<td>Column graph showing size of state work force over past four years. Graph shows work force as small in 2010 as it was in 2006.</td>
<td>Pie chart showing breakdown of participants in Florida's retirement system.</td>
<td>Photograph of African-American woman, Hayden, who is described as public defender from Clearwater, FL.</td>
<td>Two photographs of protestors, one with a &quot;stop corporate greed&quot; sign, the other in support of Governor Rick Scott.</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>Gov. Scott plans to cut spending by increasing state employees' medical premiums, mandating contributions to retirement fund, cutting jobs and weakening collective bargaining.</td>
<td>Florida Retirement System is underfunded. Gov. Scott and legislators propose mandating employee contributions. Article discusses potential changes to the pension plan.</td>
<td>Perspective of a state employee on proposed budget cuts. State employees are underappreciated and proposed cuts undermine the incentive to work for the state.</td>
<td>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.</td>
</tr>
<tr>
<td><strong>Key quotes</strong></td>
<td>&quot;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.&quot;</td>
<td>&quot;The state proposes to make public employees...chip in for their pensions that are now fully paid by their government employers.&quot;</td>
<td>&quot;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.&quot;</td>
<td>&quot;Scott has called for $5 billion in spending cuts, eliminating about 8,600 state job positions...&quot;</td>
</tr>
</tbody>
</table>

*Note:* Stories are arrayed in the order in which they were viewed by laboratory subjects.
<table>
<thead>
<tr>
<th></th>
<th>Laboratory (n=417)</th>
<th>Field (n=1,135)</th>
<th>DID</th>
<th>Correspondence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Treatment</td>
<td>Effect</td>
<td>Control</td>
</tr>
<tr>
<td>Approve of Governor Rick Scott</td>
<td>.32 (0.03)</td>
<td>.23 (0.03)</td>
<td>-0.09 **</td>
<td>.26 (0.02)</td>
</tr>
<tr>
<td>Approve of Governor Scott on Economy</td>
<td>.33 (0.03)</td>
<td>.23 (0.03)</td>
<td>-10 **</td>
<td>.28 (0.02)</td>
</tr>
<tr>
<td>Most Important Issue: Florida Budget</td>
<td>.23 (0.03)</td>
<td>.33 (0.03)</td>
<td>10 **</td>
<td>.20 (0.02)</td>
</tr>
<tr>
<td>Prefer State Workers Pay Retirement</td>
<td>.48 (0.01)</td>
<td>.40 (0.01)</td>
<td>-0.08 *</td>
<td>.47 (0.01)</td>
</tr>
<tr>
<td>Most Important Issue: FL Gov't Workers</td>
<td>.04 (0.01)</td>
<td>.08 (0.01)</td>
<td>0.04 *</td>
<td>.03 (0.01)</td>
</tr>
<tr>
<td>Prefer Property &amp; Business Tax Cuts</td>
<td>.38 (0.03)</td>
<td>.25 (0.03)</td>
<td>-13 **</td>
<td>.30 (0.02)</td>
</tr>
<tr>
<td>Believe Budget Problems are Serious</td>
<td>.56 (0.03)</td>
<td>.60 (0.03)</td>
<td>0.05</td>
<td>.57 (0.02)</td>
</tr>
<tr>
<td>Trust Florida State Government</td>
<td>.34 (0.03)</td>
<td>.31 (0.03)</td>
<td>-0.03</td>
<td>.29 (0.02)</td>
</tr>
<tr>
<td>Know Gov. Scott was a Businessman</td>
<td>.88 (0.02)</td>
<td>.91 (0.02)</td>
<td>0.03</td>
<td>.93 (0.01)</td>
</tr>
<tr>
<td>Prefer Spending Cuts vs. Raise Taxes</td>
<td>.41 (0.03)</td>
<td>.39 (0.03)</td>
<td>-0.01</td>
<td>.43 (0.02)</td>
</tr>
<tr>
<td>Know Size of Projected Budget Deficit</td>
<td>.41 (0.03)</td>
<td>.40 (0.03)</td>
<td>-0.01</td>
<td>.28 (0.02)</td>
</tr>
<tr>
<td>Know Lt. Governor Carroll</td>
<td>.65 (0.04)</td>
<td>.71 (0.04)</td>
<td>0.06</td>
<td>.82 (0.02)</td>
</tr>
<tr>
<td>Approve of President Barack Obama</td>
<td>.67 (0.03)</td>
<td>.66 (0.03)</td>
<td>-0.01</td>
<td>.67 (0.02)</td>
</tr>
<tr>
<td>Know Department Elimination</td>
<td>.67 (0.03)</td>
<td>.66 (0.03)</td>
<td>-0.01</td>
<td>.68 (0.02)</td>
</tr>
<tr>
<td>Know Pledge to Cut State Workforce</td>
<td>.61 (0.03)</td>
<td>.63 (0.03)</td>
<td>0.02</td>
<td>.60 (0.02)</td>
</tr>
<tr>
<td>Believe Florida is Satisfactory</td>
<td>.30 (0.03)</td>
<td>.29 (0.03)</td>
<td>-0.01</td>
<td>.30 (0.02)</td>
</tr>
<tr>
<td>Prefer Traffic Stops for Immigrants</td>
<td>.51 (0.03)</td>
<td>.49 (0.03)</td>
<td>-0.02</td>
<td>.54 (0.02)</td>
</tr>
</tbody>
</table>

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., [TreatmentLab-ControlLab]-[TreatmentField-ControlField]). 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)
Table 3. Predicting Treatment Effects by Context and Correspondence

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R-Squared</strong></td>
<td>.25</td>
<td>.40</td>
<td>.62</td>
</tr>
<tr>
<td><strong>Number of Obs.</strong></td>
<td>34</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td><strong>Lab Condition</strong></td>
<td>.03 **</td>
<td>.03 **</td>
<td>-.00</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.01)</td>
</tr>
<tr>
<td><strong>Correspondence Score</strong></td>
<td>--</td>
<td>.01 **</td>
<td>-.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td><strong>Lab x Correspondence</strong></td>
<td>--</td>
<td>--</td>
<td>.02 **</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.00</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>.01 **</td>
<td>-.00</td>
<td>.02 **</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.01)</td>
<td>(.00)</td>
</tr>
</tbody>
</table>

Note: The dependent variable is the absolute value of the treatment effect, treating the lab and field effects as separate observations. 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 Correspondence Score variable measures the number of treatment articles (0-4) that cover the issue raised in each survey question. ** p < .05, * p < .10 (two-tailed)
<table>
<thead>
<tr>
<th></th>
<th>Randomization Check</th>
<th>Selection Bias Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Field Control</td>
<td>Field Treatment</td>
</tr>
<tr>
<td>Female</td>
<td>.03 (.04)</td>
<td>.03 (.04)</td>
</tr>
<tr>
<td>Female Not Available</td>
<td>-.14 (.14)</td>
<td>-.05 (.14)</td>
</tr>
<tr>
<td>Black</td>
<td>-.01 (.04)</td>
<td>.04 (.04)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-.04 (.11)</td>
<td>-.05 (.11)</td>
</tr>
<tr>
<td>Race Not Available</td>
<td>.06 (.18)</td>
<td>.17 (.18)</td>
</tr>
<tr>
<td>Age</td>
<td>.00 (.00)</td>
<td>.00 (.00)</td>
</tr>
<tr>
<td>Democrat</td>
<td>-.04 (.05)</td>
<td>-.07 (.05)</td>
</tr>
<tr>
<td>Republican</td>
<td>-.04 (.06)</td>
<td>-.07 (.06)</td>
</tr>
<tr>
<td>Voted 2010</td>
<td>.01 (.05)</td>
<td>.03 (.05)</td>
</tr>
<tr>
<td>Voted 2010 N/A</td>
<td>-.10 (.22)</td>
<td>-.02 (.21)</td>
</tr>
<tr>
<td>Voted 2010 Primary</td>
<td>.04 (.05)</td>
<td>.03 (.05)</td>
</tr>
<tr>
<td>Voted 2010 Primary N/A</td>
<td>-.09 (.14)</td>
<td>.13 (.13)</td>
</tr>
<tr>
<td>Voted 2008</td>
<td>-.01 (.06)</td>
<td>-.01 (.06)</td>
</tr>
<tr>
<td>Voted 2008 N/A</td>
<td>.03 (.11)</td>
<td>.06 (.11)</td>
</tr>
<tr>
<td>Voted 2008 Primary</td>
<td>-.03 (.05)</td>
<td>-.05 (.05)</td>
</tr>
<tr>
<td>Voted 2008 Primary N/A</td>
<td>.03 (.06)</td>
<td>-.01 (.06)</td>
</tr>
<tr>
<td>Assigned to Field</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Log-Likelihood: -12468.60  -282.89  -4059.88
Model $X^2$ p-value: .99  .66  .00
Number of Obs.: 12,000  417  12,000

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, * p < .10 (two-tailed)
Supplementary Appendix

This appendix explores noncompliance in our experiments. We also show the treatment stories from the lab experiment and the mail survey that was sent as part of the field experiment. The survey questions were the same in the lab experiment (both the actual text and the formatting of the response options). The only exception is that in the lab experiment, the opening banner (“Public Opinion in Leon County”) and introductory paragraph were excluded.

Accounting for Noncompliance

In the context of our study, noncompliance occurred when someone refused to take treatment (i.e., by declining the complimentary newspaper subscription), or when a participant treated themselves on their own accord (e.g., by buying the Sunday newspaper at the store). Green and Gerber (2012) refer to this situation as “two-sided noncompliance” and note that it is common in experiments that employ encouragement designs (see Albertson and Lawrence [2009] for an empirical example). Following the procedures from Green and Gerber (2012) we estimate the complier average causal effect (CACE) by dividing the intent-to-treat (ITT) effect by the proportion of subjects who are compliers (denoted as ITTD). The ITTD is represented as:

\[
\text{ITTD} = \frac{\text{Ntr}}{\text{Nt}} - \frac{\text{Ncr}}{\text{Nc}}
\]

2 Under the assumptions of non-interference, excludability, and monotonicity, the CACE estimator provides consistent estimates of the average treatment effect among compliers (see Green and Gerber 2012, Ch. 6 for discussion).
where the first proportion is the number of newspaper readers in the treatment group ($N_{tr}$) divided by the total number of people in the treatment group ($N_t$), and the second proportion is the number of newspaper readers in the control group ($N_{cr}$) divided by the total number of people in the control group ($N_c$). Recall that in the lab experiment there are 198 people in the treatment group and 219 people in the control group. In the field experiment, there are 555 people in the treatment group and 580 people in the control group.

Using administrative data from the Tallahassee Democrat, we calculated the ITT$_D$ for the lab and field. This version takes into account people who either refused the free subscription ($n = 5$) or who started a subscription on their own during our study (4 lab controls and 15 field controls fell into this category). This yields:

$$\text{ITT}_D \text{ lab } = \frac{198}{198} - \frac{4}{219} = .98 \quad \text{ITT}_D \text{ field } = \frac{550}{555} - \frac{15}{580} = .96$$

Using the previously described question regarding newspaper usage, we calculated a second version of the ITT$_D$ for the lab and field. This question is a self-reported measure of newspaper usage and so it is subject to the usual caveats regarding self-reported measures of behavior. However, the question allows us to incorporate a wider array of non-compliance. In particular, this version of the ITT$_D$ takes into account people who stated they currently received the paper either because they were already subscribing or because they were borrowing the paper from a friend, buying it at the store, viewing it online, or obtaining it at work. Forty-seven lab controls and 116 field controls fell into this category, yielding:

$$\text{ITT}_D \text{ lab } = \frac{198}{198} - \frac{47}{219} = .79 \quad \text{ITT}_D \text{ field } = \frac{550}{555} - \frac{116}{580} = .79$$

Regardless of how we calculate it, the proportion of compliers is relatively high (ranging from .79 to .98). More importantly, the difference in this proportion across the lab and field is
minuscule, giving us confidence that the substantive conclusions reported in the paper are robust to noncompliance problems. We confirmed this conclusion by re-estimating the individual–level models with instrumental variables (and using random assignment to lab or field treatment group as our instruments; see Gerber and Green APSR 2000 or Sovey and Green AJPS 2011).