Institute for Policy Research
Northwestern University
Working Paper Series

WP-09-05

Students as Experimental Participants: A Defense of the "Narrow Data Base"

James N. Druckman
Faculty Fellow, Institute for Policy Research
Associate Professor of Political Science
Northwestern University

Cindy D. Kam
Associate Professor of Political Science
Vanderbilt University

Version: May 11, 2009

DRAFT
Please do not quote or distribute without permission.

Abstract

In this chapter, we investigate the extent to which using students as experimental participants creates problems for causal inference. First, we discuss the impact of student subjects on a study's internal and external validity. In contrast to common claims—including Sear's (1986) widely cited proclamation of students being a "narrow data base"—we argue that student subjects do *not* intrinsically pose a problem for a study's external validity. Second, we use simulations to identify situations when student subjects are likely to constrain experimental inferences. We show, perhaps surprisingly, that such situations are relatively limited. Third, we briefly survey empirical evidence that provides guidance on when researchers should be particularly attuned to taking steps to ensure appropriate generalizability from student subjects. We conclude with a discussion of the practical implications of our findings. In short, we argue that student subjects are not an inherent problem to experimental research; moreover, a case can be made that the burden of proof—of student subjects being a problem—should lie with critics rather than experimenters.

An experiment entails randomly assigning participants to various conditions or manipulations. Given common consent requirements, this means experimenters need to recruit participants who, in essence, agree to be manipulated, often in controlled environments. The ensuing practical and ethical challenges of subject recruitment have led many researchers to rely on convenience samples of college students. For political scientists who put particular emphasis on generalizabilty to relevant political situations, the use of student participants often constitutes a critical, and according to some reviewers, fatal problem for experimental studies.

In this chapter, we investigate the extent to which using students as experimental participants creates problems for causal inference. First, we discuss the impact of student subjects on a study's internal and external validity. In contrast to common claims—including Sear's (1986) widely cited proclamation of students being a "narrow data base"—we argue that student subjects do *not* intrinsically pose a problem for a study's external validity. Second, we use simulations to identify situations when student subjects are likely to constrain experimental inferences. We show, perhaps surprisingly, that such situations are relatively limited. Third, we briefly survey empirical evidence that provides guidance on when researchers should be particularly attuned to taking steps to ensure appropriate generalizability from student subjects. We conclude with a discussion of the practical implications of our findings. In short, we argue that student subjects are not an inherent problem to experimental research; moreover, a case can be made that the burden of proof—of student subjects being a problem—should lie with critics rather than experimenters.

The Validity of Using Student Subjects

Although internal validity may be the "sine qua non" of experiments, most researchers use experiments to make *generalizable* causal inferences (Shadish et al. 2002: 18-20). For

example, a researcher might wish to assess whether a media story about a welfare program causes viewers to become more supportive of the program. An experiment aims to isolate the nature of the relationship between the stimulus (story) and the response (welfare support) (e.g., is there a causal relationship?; is it strong?). Focusing on causal inference differs from descriptive inference, where the point might be to portray the percentage of voters who support welfare or the extent of a given individual's support (e.g., a low or high score on an evaluation scale) (e.g., Gerring 2001).

A critical element in making causal inference is the assurance of internal validity: "inferences about whether observed covariation between A and B reflects a causal relationship from A to B…" (Shadish et al. 2002: 53). For example, there may exist covariation between viewing the aforementioned media story and welfare support. Internal validity refers to the confidence one can have that the story causes support. This is a tricky question since it may be that support for welfare causes news attention or some third factor such as partisanship stirs viewing and support. Experiments employ random assignment that, when successful, ensures near definitive causal documentation. If individuals randomly assigned to watch the news story exhibit significantly greater support for welfare than those randomly assigned to not watch (on average), confidence can be taken that the story caused support, at least in the context of the study with the particular participants. When random assignment is successfully carried out, experiments constitute the "gold standard" of causal inference (Shadish et al. 2002: 13).[1] Internal validity is critical—"if a study has low internal validity—if it doesn't clearly demonstrate a causal relation between the independent and dependent variables—then there is nothing to generalize" (Anderson and Bushman 1997: 21; also see McDermott 2002: 334-335).

[1] McDermott's chapter discusses threats to internal validity in experiments. Bowers' chapter explores failed random assignment.

As mentioned, many experimentalists also seek to "generalize" a documented causal relationship, and this introduces a host of other issues. For example, upon finding a causal connection between the welfare story and support in a laboratory study with students, one might ask whether the relationship also exists within a heterogeneous population, in a large media marketplace, over time. This is largely a question of external validity, which refers to the extent to which the "causal relationship holds over variations in persons, settings, treatments [and timing], and outcomes" (Shadish et al. 2002: 83). McDermott (2002: 334) explains that "External validity… tend so preoccupy critics of experiments. This near obsession… tend[s] to be used to dismiss experiments…" (also see, e.g., Anderson and Bushman 1997, Levitt and List 2007).

A point of particular concern involves generalization from the sample of experimental participants—especially when, as is often the case, the sample consists of students—to a larger population of interest. Indeed, this was the focus of Sears' (1986) widely cited article, "College Sophomores in the Laboratory: Influences of a Narrow Data base on Social Psychology's View of Human Nature." And, many political scientists employ "the simplistic heuristic of 'a student sample lacks external generalizability'" (Kam et al. 2007: 421) (e.g., Lijphart 1971, Bartels 1993: 267, McGraw and Hoekstra 1994, Jacoby 2000: 753).[2] Gerber and Green (2008: 358) note the same reaction in political science, explaining that "If one seeks to understand how the general public responds to social cues or political communication, the external validity of lab studies of undergraduates has inspired skepticism (Sears 1986, Benz and Meier 2006)." In short, social scientists in general and political scientists in particular view student subjects as a major hindrance to drawing inferences from experimental studies.

[2] Through 2008, Sears (1986) article has been cited an impressive 446 times according to the Social Science Citation Index. It is worth noting that Sears' argument is conceptual—he does not offer empirical evidence that student subjects create problems (although see, e.g., Peterson 2001, which we will discuss later).

Assessing the extent to which using student subjects is problematic has particular current relevance. First, many political science experiments use student subjects; for example, Kam et al. (2007: 419-420) report that from 1990 through 2006, a quarter of experimental articles in general political science journals relied on student subjects while over 70% did so in more specialized journals (also see Druckman et al. 2006).[3] Are the results from these studies of questionable validity? Second, there are practical issues. A common rationale for moving away from laboratory studies, in which student subjects are relatively common, to survey and/or field experiments is that these latter venues facilitate using non-student participants (e.g., Sniderman and Grob 1996, Lee et al. 2005, Brooks and Geer 2007: 2, Gerber and Green 2008: 358). When evaluating the pros and cons of laboratory versus survey or field experiments, should substantial weight be given to whether participants are students? Similarly, those implementing lab experiments have increasingly put forth efforts (and paid costs) to avoid student subjects (e.g., Lau and Redlawsk 2006: 65-66; Kam 2007). Are these costs worthwhile? To address these questions, we next turn to a broader discussion of what external validity demands.

The Dimensions of External Validity

To assess the external validity or generalizability of a causal inference, one must consider *from what* we are generalizing and *to what* we hope to generalize. When it comes to "from what," a critical, albeit often neglected, point is that external validity is best understood as being assessed over a range of studies on a single topic (McDermott 2002: 335). Liyanarachchi (2007: 55) explains:

> According to experts on methodology, true external validity of findings can only be obtained by converging the results of many studies in an area (e.g., validity by convergence proposed by Campbell and Fiske 1959, meta-analysis developed by Hunter et al. 1982). Reiterating this point in social sciences, McGrath et al. (1982: 105) suggested: "No one 'finding' is evidence, and no one study yield[s] "knowledge;'' empirical information can gain credence *only* by accumulation of convergent results."

[3] This is even more of an issue in psychology (see Sherman et al. 1999 for a content analysis).

Assessment of any single study, regardless of the nature of its participants, must be done in light of the larger research agenda to which it hopes to contribute.[4]

Moreover, when it comes to generalization from a series of studies, the goal is to generalize across *multiple* dimensions. External validity refers to generalization not only of individuals but also across settings/contexts, times, and operationalizations. There is little doubt that institutional and social contexts play a critical role in determining political behavior, and consequently that they can moderate causal relationships. One recent powerful example comes from the political communication literature; a number of experiments, using both student *and* non-student subjects, show that when exposed to political communications (e.g., in a laboratory), individuals' opinions often reflect the content of those communications (see, e.g., Kinder 1998, Chong and Druckman 2007b). The bulk of this work, however, ignores the contextual reality that people outside of the controlled study setting have choices (i.e., they are not captive). Arceneaux and Johnson (2008) show that as soon as participants in communication experiments can choose whether to receive a communication (i.e., the captive audience constraint is removed), results about the effects of communications drastically change (and become less dramatic). In this case, ignoring the contextual reality of choice appears to have constituted a much greater threat to external validity than the nature of the subjects.[5]

[4] This is consistent with a Popperian approach to causation that suggests causal hypotheses are never confirmed and evidence accumulates via multiple tests, even if all of these tests have limitations. Campbell (1969: 361) offers a fairly extreme stance on this when he states, "…had we achieved one, there would be no need to apologize for a successful psychology of college sophomores, or even of Northwestern University coeds, or of Wistar staring white rats."

[5] A related example comes from Barabas and Jerit's (2009) study that compares the impact of communications in a survey experiment against analogous dynamics that occurred in actual news coverage. They find the survey experiment vastly over-stated the effect, particularly among certain sub-groups. Sniderman and Theriault (2004) and Chong and Druckman (2007a) also reveal the importance of context; both studies show that prior work that limits competition between communications (i.e., by only providing participants with a single message rather than a mix that is typically found in political contexts) likely misestimate the impact of communications on public opinion.

Timing also matters; experiments implemented at one time may not hold at other times given the nature of world events. Gaines et al. (2007) further argue that survey experiments in particular may misestimate effects due to a failure to consider what happened prior to the study (also see Gaines and Kuklinski's chapter). Building on this insight, Druckman (2009) asked survey respondents for their opinions about a publicly owned gambling casino, which was a topic of "real world" ongoing political debate. Prior to expressing their opinions, respondents randomly received no information (i.e., control group) or information that emphasized either economic benefits or social costs (e.g., addiction to gambling). Druckman shows that the opinions of attentive respondents in the economic information condition did not significantly differ from attentive individuals in the control group.[6] The non-effect likely stemmed from the economic information—which was available outside the experiment in ongoing political discussion—having already influenced all respondents. Another exposure to this information in the experiment did not add to the prior, pre-treatment effect. In other words, the ostensible non-effect lacked external validity—not because of the sample—but because it failed to account for the timing of the treatment (also see Slothuus 2009).[7]

A final dimension of external validity involves how concepts are employed. Finding support for a proposition means looking for different ways of administering and operationalizing the treatment (e.g., delivering political information via television ads, newspaper stories, interpersonal communications, survey question text) and operationalizing the dependent variables (e.g., behavioral, attitudinal, physiological, implicit responses). For example, in their study of the relationship between altruism, partisanship, and participation, Fowler and Kam

[6] For reasons explained in his paper, Druckman (2009) also focuses on individuals more likely to have formed prior opinions about the casino.
[7] Another relevant timing issue concerns the duration of any experimental treatment effect (see, e.g., Gaines et al. 2007, Gerber et al. 2007).

(2007) move away from conventional attitudinal measures of altruism (e.g., that use self-reported evaluations of statements such as, "One of the problems of today's society is that people are often not kind enough to others") by using an experimental dictator game. In the game, individuals decide how much, if any, of a pot of money to share with another player (where there is no penalty for sharing nothing). The other player is either an anonymous individual, a registered Republican, or a registered Democrat. They find a strong connection between these measures—where sharing more money with the anonymous individual indicates increased altruism, and where sharing more money with the registered partisan indicates social identification with the party—and participation.[8]

In short, external validity does *not* simply refer to whether a specific study, if re-run on a different sample, would provide the same results. It refers more generally to whether "conceptually equivalent" (Anderson and Bushman 1997) relationships can be detected across people, places, times, and operationalizations. This introduces the other end of the generalizability relationship—that is, "equivalent" to what? For many, the "to what" refers to behavior as observed outside of the study, but this is not always the case. Experiments have different purposes; Roth (1995:22) identifies three non-exclusive roles that experiments can play: "search for facts," "speaking to theorists," or "whispering in the ears of princes," which facilitates "the dialogue between experimenters and policymakers" (also see Guala 2005: 141-160). These types likely differ in the target of generalization. Of particular relevance is that theory oriented experiments typically are not meant to "match" behaviors observed outside the study *per se*, but rather the key is to generalize to the precise parameters put forth in the given theory. Plott (1991: 906) explains that "The experiment should be judged by the lessons it

[8] Also see Loewen (2009) who examines the link between the dictator game giving and support for public spending. Another example comes from Lupia and McCubbins' (1998) who test their theory of persuasion using both economic and psychological types of experiments.

teaches about the theory and not by its similarity with what nature might have happened to have created." This echoes Mook's (1983) argument that much experimental work is aimed at developing and/or testing a theory, not at establishing generalizability. Even experiments that are designed to demonstrate "what can happen" (e.g., Milgram, Zimbardo, Asch) can still be useful, even if they do not mimic everyday life.[9] In many of these instances, the nature of the subjects in the experiments are of minimal relevance, particularly given experimental efforts to ensure their preferences and/or motivations match those in the theory (e.g., see Dickson's chapter on induced value theory).

Assessment of how student subjects influence external validity depends on three considerations: (1) the research agenda on which the study builds (e.g., has prior work already established relationship with student subjects, meaning incorporating other populations may be more pressing?), (2) the relative generalizability of the subjects, compared to the setting, timing, and operationalizations (e.g., a study using students may have more leeway to control these other dimensions), and (3) the goal of the study (e.g., to build a theory or to generalize one).

Evaluating External Validity

The next question is how to evaluate external validity. While this is best done over a series of studies, we acknowledge the need to assess whether a particular study contributes or detracts from the validity of a research agenda. Individual studies can be evaluated in at least two ways (Aronson and Carlsmith 1968, Aronson et al. 1998). First, experimental realism refers to whether "an experiment is realistic, if the situation is involving to the subjects, if they are forced to take it seriously, [and] if it has impact on them" (Aronson et al. 1985: 485). Second, mundane realism concerns "the extent to which events occurring in the research setting are likely to occur

[9]Aronson et al. (1998: 132) explain that it "is often assumed (perhaps mindlessly!) that all studies should be as high as possible in external validity, in the sense that we should be able to generalize the results as much as possible across populations and settings and time. Sometimes, however, the goal of the research is different."

in the normal course of the subjects' lives, that is, in the 'real world.'" (Aronson et al. 1985: 485).[10]

Much debate about samples focuses on mundane realism. When student subjects do not match the population to which a causal inference is intended (Kam et al. 2007: 419), many conclude that the study has low external validity. Emphasis on mundane realism, however, is misplaced (e.g., see McDermott 2002, Morton and Williams 2008: 345): of much greater importance is experimental realism. Failure of participants to take the study and treatments "seriously" compromises internal validity, which in turn, renders external validity of the causal relationship meaningless (e.g., Dikhaut et al. 1972: 477, Liyanarachchi 2007: 56).[11] In contrast, at worst, low levels of mundane realism simply constrain the breadth of any generalization but do not make the study useless.

Moreover, scholars have yet to specify clear criteria for assessing mundane realism, and, as Liyanarachchi (2007: 57) explains, "any superficial appearance of reality (e.g., a high level of mundane realism) is of little comfort, because the issue is whether the experiment 'captures the intended essence of the theoretical variables' (Kruglanski 1975: 106)."[12] That said, beyond superficiality, we recognize student subjects—while having no ostensibly relevant connection with experimental realism—may limit mundane realism that constrain generalizations of a particular study. *This occurs when features of the subjects affect the nature of the causal*

[10] A third evaluative criterion is psychological realism which refers to "the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life" (Aronson et al. 1998: 132). The relevance of psychological realism is debatable, and depends on one's philosophy of science (c.f., Friedman 1953 and Simon 1963, 1979: 475-476; also see MacDonald 2003).

[11] We do not further discuss steps that can be taken to ensure experimental realism, as this moves into the realm of other design issues (e.g., subject payments, incentives; see Dickson's chapter).

[12] Berkowitz and Donnerstein (1982: 249) explain that "The meaning the subjects assign to the situation they are in and the behavior they are carrying out [i.e., experimental realism] plays a greater part in determining generalizabilty of an experiment's outcome than does the sample's demographic representatives or the setting's surface realism."

relationship being generalized. When this occurs and with what consequences are questions to which we now turn.

Statistical Framework

In this section, we examine the "problem" of convenience samples from a statistical point of view. This allows us to specify the conditions *under which* student samples might constrain casual generalization (in the case of a single experiment). Our focus, as in most political science analyses of experimental data, is on the magnitude of some experimental treatment, *T*, on an attitudinal or behavioral dependent measure, *y*.[13] Suppose, strictly for presentational purposes, we are interested in the effect of a persuasive communication (*T*) on a subject's post-stimulus policy opinion (*y*) (we could use virtually any example from any field). *T* takes on a value of 0 for subjects randomly assigned to the control group and takes on a value of 1 for subjects randomly assigned to the treatment group.[14] Suppose the true data generating process is:

$$y_i = \beta_0 + \beta_T T_i + \varepsilon_i \qquad [1]$$

Assuming that ε_i is a well-behaved disturbance term with mean zero, variance of σ^2, and $\text{Cov}(\varepsilon_i, \varepsilon_j)=0$, and all other assumptions of the classical linear regression model are met, the OLS estimate for β_T should be unbiased, consistent, and efficient.[15] The results derived from

[13] Psychologists typically use analysis of variance, but it is identical in practice.
[14] For ease of exposition, our example only has one treatment group. The lessons easily extend to multiple treatment groups.
[15] We could have specified a data generating process that also includes a direct relationship between *y* and some individual-level factors such as partisanship or sex (consider a vector of such variables, **X**). Under random assignment, the expected covariance between the treatment and **X** is zero. Hence, if we were to estimate the model without **X**, omitted variable bias would technically not be an issue. If the data generating process does include **X**, and even though we might not have an omitted variable bias problem, including **X** in the model may still be advisable. Inclusion of relevant covariates (that is, covariates that, in the data generating process, actually have a nonzero effect on y) will reduce e_i (the difference between the observed and predicted y), which in turn will reduce s^2, resulting in more precise estimated standard errors for our coefficients (see Franklin 1991). Moreover, it is only in *expectation* that $\text{Cov}(\mathbf{X},T)=0$. In any given sample, $\text{Cov}(\mathbf{X},T)$ may not equal zero. Inclusion of covariates can mitigate against incidental variation in cell composition. In advising inclusion of control variables, Ansolabehere and Iyengar (1995: 172) note, "…randomization does not always work. Random assignment of treatments provides a general safeguard against biases but it is not foolproof. By chance, too many people of a particular type may end up in one of the treatment groups, which might skew the results…" (also see Bowers' chapter).

estimation on a given sample would be fully generalizable to those that would result from estimation on any other sample.

Specific samples will yield various distributions across a wide span of individual covariates. To continue with our running example about persuasive communication, samples may differ in the distribution of *attitude crystallization*.[16] Student samples may yield a disproportionately high group of subjects that are low in crystallization, with only a small proportion that is high in it (Sears 1986). A random sample from the general population might lead to a group that is normally distributed and centered at the middle of the range. A sample from politically active individuals (such as conventioneers) might result in a group that is disproportionately high in crystallization, with very few (if any) respondents who are low in crystallization.[17]

For illustrative purposes, consider the following samples with varying distributions on attitude crystallization. In all cases, N=200 and treatment is randomly assigned to half of the cases. Let attitude crystallization range from 0 (low) to 1 (high). Consider one sample where 90% of the sample is at a value of "0" and 10% of the sample is at a value of "1". Call this the "Student Sample." Consider a second sample where the sample is normally distributed and centered on 0.5 with standard deviation of 0.165. Call this the "Random Sample." Consider a third sample where 10% of the sample is at a value of "0" and 90% of the sample is at a value of 1. Call this the "Conventioneers Sample."[18]

[16] This example is inspired by Sears' (1986) discussion of "Uncrystallized Attitudes."
[17] And, of course, crystallization might vary across different types of issues. On some issues (e.g., financial aid policies), students might have highly crystallized views, whereas conventioneers might have less crystallized views.
[18] Now, if our goal was to use our three samples to make descriptive inferences about the general population's mean level of attitude crystallization, then both the Student Sample and the Conventioneers Sample would be inappropriate. The goal of an experimental design is expressly *not* to undertake this task. Instead, the goals of an experimental design are to estimate the causal effect of some treatment and then to generalize it.

Suppose the true treatment effect (β_T) takes a value of 4. We set up a Monte Carlo experiment with the parameter $\beta_T = 4$, that estimated Equation [1] 1,000 times, each time drawing a new ε term. We repeated this process for each of the three types of samples (student, random, and conventioneers). The sampling distributions for b_T appear in Figure 1.

[Figure 1 about here]

The results in Figure 1 demonstrate that when the true data generating process produces a single treatment effect, estimates on any sample—whether it is drawn from students, the general population, or conventioneers—will produce an unbiased estimate of the true underlying treatment effect. Perhaps this point seems obvious, but we believe it has escaped notice from many who criticize experiments that rely on student samples. We repeat: *If the underlying data generating process is characterized by a homogeneous treatment effect (i.e., the treatment effect is the same across the population), then any convenience sample should produce an unbiased estimate of that single treatment effect, and, thus, the results from any convenience sample should be easily generalizable to any other group of individuals.* Put another way, if the treatment effect is the same across populations, the nature of a particular sample is largely irrelevant for establishing that effect.

Suppose, however, the "true" underlying data generating process contains a heterogeneous treatment effect: that is, the effect of the treatment is moderated[19] by individual-level characteristics (i.e., the size or direction of the treatment effect varies within subgroups of the population). For example, the size of the treatment effect might depend upon some subject

[19] See Baron and Kenny (1986) for the distinction between moderation and mediation. Political scientists often use the terms interchangeably to refer to instances where some variable, Z, influences the effect of some other variable (such as a treatment). Psychologists refer to the case where Z affects the *effect* of X as moderation (i.e., an interaction effect). Psychologists refer to mediation when some variable X influences the *level* of some variable Z, whereby X affects Y through its effect on the level of Z. A mediating variable might be thought of as a mechanism. For an extended treatment of interaction effects in regression analysis, see Kam and Franzese (2007). For a discussion of mediation, see Bullock and Ha's chapter.

characteristic, such as gender, race, age, education, sophistication, etc. Another way to say this is that there may be an "interaction of causal relationship with units" (Shadish et al. 2002: 87). Under this line of theorizing, we note that a cause-effect relationship derived from a particular sample may not necessarily generalize to another sample.

As one method of overcoming this issue, researchers may use random sampling to ensure external validity; if a researcher can randomly sample experimental subjects, then the researcher can be assured that:

> the average causal relationship observed in the sample will be the same as (1) the average causal relationship that would have been observed in any other random sample of persons of the same size from the same population and (2) the average causal relationship that would be been observed across *all* other persons in that population who were not in the original random sample. That is, random sampling eliminates possible interactions between the causal relationship and the class of persons who are studied versus the class of persons who are not studied within the same population (Shadish et al. 2002: 91).

Although random sampling has advantages for external validity, Shadish et al. (2002: 91) note that "it is so rarely feasible in experiments." For political scientists who conduct experiments, the way to move to random sampling might be to use survey experiments, where respondents are (more or less) a random sample of some population of interest. We will say a bit more about this possibility, below. For now, let us assume that a given researcher has a specific set of reasons for not using a random sample (cost, instrumentation, desire for laboratory control, etc.), and let's examine what challenges a researcher using a convenience sample might face in this framework.

To do so, we revise our data generating process to reflect the possibility that some individual-level characteristic moderates the treatment effect. We take our basic Equation in [1] and theorize that some individual-level characteristic, Z, influences the magnitude of the treatment effect:[20]

[20] The discussion that follows could also be considered within a hierarchical linear modeling approach.

$$\beta_1 = \gamma_{10} + \gamma_{11}Z_i \qquad [2]$$

We also theorize that the individual-level characteristic, *Z*, might influence the intercept:

$$\beta_0 = \gamma_{00} + \gamma_{01}Z_i$$

Substituting into [1]:

$$y_i = (\gamma_{00} + \gamma_{01}Z_i) + (\gamma_{10} + \gamma_{11}Z_i)T_i + \varepsilon_i$$

$$y_i = \gamma_{00} + \gamma_{01}Z_i + \gamma_{10}T_i + \gamma_{11}Z_i*T_i + \varepsilon_i \qquad [3]$$

If our sample includes sufficient variance on this moderator, and we have *ex ante* theorized that the treatment effect depends upon this moderating variable, *Z*, then we can (and should) *estimate* the interaction. If, however, the sample does not contain sufficient variance, not only can we not identify the moderating effect, but we may misestimate the on-average effect—depending on what specific range of *Z* is present in our sample.

In short, the question of generalizing treatment effects reduces to asking if there is a single treatment effect or a set of treatment effects, the size of which depends upon some (set of) covariate(s). Note that this should be a *theoretically oriented* question of generalization. It is not just whether "student samples are generalizable" but rather, what particular characteristics of student samples might lead us to question whether the causal relationship detected in a student sample experiment would be systematically different from the causal relationship in the general population.

Suppose, to revisit our running example, we are interested in the extent to which a subject's level of attitude crystallization *(Z)* influences the effect of a persuasive communication (*T*) on a subject's post-stimulus policy opinion (*y*). The theory is that the more crystallized someone's attitude is, the smaller the treatment effect should be. The less crystallized a person's

attitude is, the greater the treatment effect should be. Using this running example, based on equation [3], assume that the true relationship has the following (arbitrarily selected) values:

$\gamma_{00} = 0$

$\gamma_{01} = 0$

$\gamma_{10} = 5$

$\gamma_{11} = -5$

Assume that *Z*, attitude crystallization, ranges from 0 (least crystallized) to 1 (most crystallized). γ_{10} tells us the effect of the treatment when *Z*=0, that is, the treatment effect among the least crystallized subjects. γ_{11} tells us how crystallization moderates the effect of the treatment.

Substituting these values into Equation [2], we see that $\beta_1 = 5 - 5Z_i$. The true treatment effect (β_1) linearly declines with values of *Z*, attitude crystallization. In other words, the lower the level of crystallization, the higher treatment effect is. The higher the level of crystallization, the lower the treatment effect is. At the highest levels of crystallization, there is no treatment effect. We can graph this hypothetical relationship as shown in Figure 2. Here, we see that the treatment effect is a linear function of crystallization.[21]

[Figure 2 about here]

We can set up a Monte Carlo experiment with the parameters laid out above:

$$y_i = 0 + 0Z_i + 5T_i - 5Z_i*T_i + \varepsilon_i$$

First, consider what happens when we estimate [1], the simple (but theoretically incorrect, given it fails to model the moderating effect) model that looks for the "average" treatment effect: $y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$. We estimated this model 1,000 times, each time drawing a new ε term. We repeated this process for each of the three types of samples. The results appear in Figure 3.

[21] In this simple example, we assume linearity in how Z affects the treatment effect. Nonlinearities are, of course, possible.

[Figure 3 about here]

When we estimate a "simple" model, looking for an average treatment effect, our estimates for β_1 diverge from sample to sample. In cases where we have a student sample, and where low levels of crystallization increase the treatment effect, we will systematically overestimate the treatment effect relative to what we would get in estimating the same model on a random sample with moderate levels of crystallization. In cases where we have a conventioneers sample, and where high levels of crystallization depress the treatment effect, we will systematically underestimate the treatment effect, relative to the estimates obtained from the general population.

Note that in these three cases, we have obtained three different results because we have estimated a model based on Equation [1]. Equation [1] should be estimated when the data generating process produces a *single* treatment effect: the value of β_1. Instead, we have "mistakenly" estimated Equation [1] when the true data generating process produces a series of treatment effects (governed by the function: $\beta_1 = 5 - 5Z_i$). The sampling distributions above produce an "average" treatment effect depends directly upon the mean value of Z within a given sample: $5 - 5*E(Z)$.

Recall that the Student Sample is distributed such that 90% of the sample is at a value of 0 and 10% of the sample is at a value of 1. As Figure 1 demonstrates, the sampling distribution for this sample is centered on 4.5, which is: $5 - 5*(0*.90+1*.10) = 4.5$. Similarly, the sampling distribution for the Conventioneers Sample is centered on 0.5, which maps exactly onto the distribution of Z in the sample: $5 - 5*(0*.10+1*.90) = 0.5$. Finally, the sampling distribution for the Random Sample is centered on 2.5, which represents $5 - 5*\int_0^1 Z_i p(Z_i) dZ = 5 - 5*0.5 = 2.5$.

Are the results from one sample more trustworthy than the results from another sample? As Shadish et al (2002) note, conducting an experiment on a random sample will produce an "average" treatment effect; hence, to some degree the results from the Random Sample might be more desirable than the results from the other two convenience samples. However, we would argue that all three sets of results reflect a fundamental disjuncture between the model that is estimated and the true data generating process. If we have a theoretical reason to believe that the data generating process is more complex, then we should embed this *theoretical model* into our *statistical model*. Hence, if we have genuine beliefs that there is an interaction between T and Z, then we should explicitly model this interaction.[22]

So, let's see what happens when we do. We returned to Equation [3] and estimated the model 1,000 times, each time drawing a new ε term. We repeated this process three times, for each of the three types of samples (Student Sample, Random Sample, and Conventioneers Sample). The results appear in Figure 4 and Table 1.

[Figure 4 about here]

[Table 1 about here]

First, notice that the sampling distributions for b_T are all centered on the same value: 5, and the sampling distributions for b_{TZ} are also all centered on the same value: -5. In other words, Equation [3] produces *unbiased point estimates* for the terms β_T and β_{TZ}, regardless of which sample is used to estimate it. We are able to uncover unbiased point estimates even in smallish samples (N=200) where only 10% of the sample provides that key variation on Z (Student Sample and Conventioneers Sample).

[22] This, of course, assumes there is some variation in Z in our sample. Below, we will elaborate upon how much variation is needed and for what purposes.

Next, notice the spread of the sampling distributions. We have the most certainty about b_T in the Student Sample and substantially less certainty in the Random Sample and the Conventioneers Sample. The greater degree of certainty in the Student Sample results from the greater mass of the sample that is located at 0 in the Student Sample (since the point estimate for β_T, the un-interacted term in Equation [3], represents the effect of T when Z happens to take on the value of 0).[23]

For the sampling distribution of b_{TZ}, we have higher degrees of certainty (smaller standard errors) in the Student Sample and the Conventioneers Sample. This is an interesting result. By using samples that have higher variation on Z, we yield more precise point estimates of the heterogeneous treatment effect.[24] Moreover, we are still able to uncover the interactive treatment effect, since these samples still contain some variation across values of *Z*.

How much variation in Z is sufficient? So long as Z varies to any degree in the sample, the estimates for b_T and b_{TZ} will be *unbiased*. But being "right on average" may be little comfort if the degree of uncertainty around the point estimate is large. If Z does not vary very much in a given sample (that is, its range is constrained), then the estimated standard error for b_{TZ} will be large. But this degree of uncertainty is a run-of-the-mill concern when estimating a model on any dataset: more precise estimates arise from analyzing datasets that maximize variation in our independent variables.

Our discussion thus suggests that experimentalists (and their critics) need to consider the underlying data generating process: that is, *theory* is important. If a single treatment effect is theorized, then testing for a single treatment effect is appropriate. If a heterogeneous treatment

[23] See Kam and Franzese (2007) for guidance on interpretation of coefficients in interactive models.
[24] Uncovering more certainty in the Student and Conventioneers Samples (compared to the Random Sample) derives from the specific ways in which we have constructed the distributions of Z. If the Random Sample were, say, uniformly distributed rather than normally distributed along Z, then the same result would not hold. The greater precision in the estimates depends upon the underlying distribution of Z in a given sample.

effect is theorized, then researchers should be explicit in explaining how the treatment effect should vary along a specific (set of) covariate(s), and researchers can thereby estimate such relationships so long as there is sufficient variation in the specific (set of) covariate(s) in the sample. We hope to push those who launch vague criticisms regarding the "ungeneralizability" of student samples to instead think more deeply: to consider whether and in what ways the underlying data generating process would suggest a heterogeneous treatment effect that depends upon a particular (set of) covariate(s).

In sum, we have identified three distinct situations. First, in the homogenous case—where the data generating process produces a single effect β_T, of *T* on *y*—we showed the estimated treatment effect derived from a student sample is an unbiased estimate of the true treatment effect. Second, when there is a heterogeneous case (where the treatment effect is moderated by some covariate *Z*) *and* the researcher fails to recognize the contingent effect, a student sample may misestimate the effect (if the student sample is non-representative on the particular covariate *Z*). However, in this case, even a representative sample would mis-specify the effect due to a failure to model the interaction. Third, when the researcher appropriately models the heterogeneity with an interaction, then the student sample, even if it is non-representative on the covariate Z, will mis-estimate the effect *only* if there is virtually no variance (i.e., literally almost none) on the moderating dynamic. Moreover, a researcher can empirically assess the degree of variance on the moderator within a given sample, and/or use simulations to evaluate whether limited variance poses a problem for uncovering the interactive effect.

Contrasting Student Samples with Other Samples

We have argued that a given sample constitutes one—and arguably not a critical one—of many considerations when it comes to assessing external validity. Further, a student sample only creates a problem when an ill-informed researcher fails to model a contingent causal effect (when there is an underlying heterogeneous treatment effect), and the students differ from the target population with regard to the distribution of the moderating variable. This situation, which we acknowledge does occur with non-trivial frequency, leads to the question of just how often student subjects empirically differ from representative samples. The greater such differences, the more likely problematic inferences occur.

Kam (2005) offers some telling evidence comparing student and non-student samples on two variables that often affect information professing: political awareness and need for cognition (Bizer et al. 2004). She collected data from a student sample using the exact same items as are used in the National Election Study's (NES) representative sample of adult citizens. She finds the distributions for both variables in the student sample closely resemble those in the 2000 NES.[25] This near identical match in distribution, then, allowed Kam (2005) to more broadly generalize results from an experiment, on party cues, she ran with the student subjects.

Kam focuses on awareness and need for cognition because these variables plausibly moderate the impact of party cues—as explained, in comparing student and non-student samples, one should focus on possible differences that *are relevant to the study in question*. Of course, one

[25] For political awareness, subjects were asked to identify the positions of four political figures: Trent Lott, William Rehnquist, Tony Blair, and John Ashcroft. The four items were averaged to form a scale. The experimental sample mean is 0.34 (with standard deviation of 0.34) compared with 0.27 (s.d. 0.28) in NES 2000; reliability for the scale is 0.71 for the experimental sample and 0.64 for NES 2000. For Need for Cognition, subjects responded to a pair of items. The additive scale composed of the two items ranges from 0 to 1, has a mean of 0.64 (with a standard deviation of 0.18), and $\alpha = 0.48$. There were no significant differences across conditions. In the NES 2000, the additive raw scale ranges from 0 to 1, has a mean of 0.60 (s.d. 0.35) and $\alpha = 0.61$. The difference in the standard deviations can be attributed to differences in response alternative format. Since one of the need for cognition items on the NES was measured in only two (instead of five) categories, it consequently has a higher variance.

may nonetheless wonder whether students differ in others ways that *could* matter (see e.g., Sears 1986: 520). This requires a more general comparison, which we undertake by turning to the 2006 Civic and Political Health of the Nation Dataset (collected by CIRCLE) (for a similar exercise, see Kam et al. 2007).[26]

These data consist of telephone and web interviews with 2,232 individuals 15 and older living in the continental US. The sampling frame included youth ages 15-25 (N=1674) and adults 26 and over (N=547). We limited the analysis to individuals aged 18 and over. We selected all ostensibly politically relevant predispositions available in the data,[27] and then compared individuals currently enrolled in college against the general population. The appendix contains question wording for each item.

[Table 2 about here]

As we can see from Table 2, there are several instances where the means for students and the non-student general population are indistinguishable from zero. Students and the non-student general population are, on average, indistinguishable when it comes to partisanship, ideology, the importance of religion, belief in limited government, views about homosexuality as a way of life, the contributions of immigrants to society, social trust, degree of following and discussing politics, and overall media use. Students are distinguishable from the non-student general population in religious attendance (but not the importance of religion), in level of political information as measured in this particular dataset[28], and in specific types of media use (students use the internet more than the non-student general population to get news; students view national

[26] We use CIRCLE data since it provides a sufficient number of student aged respondents. In contrast, for example, the 2008 NES contained only 65 students and 145 individuals under 22 (out of a total N of 2,323).
[27] We did this before looking at whether there were differences between students and the non-student general population sample; that is, we did not selectively choose variables.
[28] The measure of political information in this dataset is quite different from that typically found in the NES; it is heavier on institutional items and relies on more recall than recognition.

network news less than the non-student general population does). Overall, however, we are impressed by just how similar students are on key covariates often of interest to political scientists to the non-student general population.

In cases where sample differences do occur on variables that are theorized to influence the size and direction of the treatment effect, the next step entails assessing the problem. Most straightforwardly, as explained, the researcher should check for at least some variance in the experimental (student sample) and model the interaction. The researcher also might consider cases where students—despite differing on relevant variables—might be advantageous. In some situations, students facilitate testing a causal proposition. Students are relatively educated, in need of small amounts of money, and accustomed to following instructions (e.g., from Professors) (Guala 2005: 33-34). For these reasons, student samples may enhance the experimental realism of experiments that rely on induced value theory (where monetary payoffs are used to induce preferences) and/or involve relatively complicated, abstract instructions (Frideman and Sunder 1994: 39-40).[29] The goal of many of these experiments is to test theory, and, as mentioned, the match to the theoretical parameters (e.g., the sequence of events if the theory is game theoretic) is of utmost importance (rather than mundane realism).

Alternatively, estimating a single treatment effect upon a student sample subject pool can sometimes make it *harder* to find effects. For example, studies of party cues examine the extent to which subjects will follow the advice given to them by political parties. Strength of party identification might be a weaker cue for student subjects, whose party affiliations are still in the formative stages (Campbell et al. 1960, Niemi and Jennings 1991). If this were the case, then the use of a student sample would make it even more difficult to discover party cue effects. To the

[29] We suspect that this explains why the use of student subjects seems to be much less of an issue in experimental economics (e.g., Guala 2005).

extent that party cues work among student samples, these likely *underestimate* the degree of cue-taking that might occur among the general population, whose party affiliations are more deeply grounded. Similarly, students seem to exhibit relatively lower levels of self-interest and susceptibility to group norms (Sears 1986: 524) meaning that using students in experiments on these topics increases the challenge of identifying treatment effects.[30]

Finally, it is worth mentioning that if the goal of a set of experiments is to generalize a theory, then testing the theory across a set of carefully chosen convenience samples may even be superior to testing the theory within a single random sample. A theory of the moderating effect of attitude crystallization on the effects of persuasive communications might be better tested on a series of different samples (and possibly different student samples) that vary on the key covariate of interest.

Researchers need to consider what particular student sample characteristics might lead a casual relationship discovered in the sample to systematically differ from what would be found in the general population. Researchers then need to elaborate upon the direction of the bias: the variation might facilitate the assessment of causation, and/or it might lead to *either* an overestimation *or* an underestimation of what would be found in the general population.

Conclusion

As mentioned, political scientists are guilty of a "near obsession" with external validity (McDermott 2002: 334). And, this obsession with external validity focuses nearly entirely upon a single dimension of external validity: who is studied. Our goal in this paper has been to situate the role of experimental samples within a broader framework of how one might assess the generalizability of an experiment. Our key points are, as follows.

[30] As explained, students also tend to be more susceptible to persuasion (Sears 1986). This makes them a more challenging population on which to experiment if the goal is to identify conditions where persuasive messages fail (see, e.g., Druckman 2001).

- The external validity of a single experimental study must be assessed in light of an entire research agenda, *and* in light of the goal of the study (e.g., testing a theory or searching for facts).

- Assessment of external validity involves multiple-dimensions including the sample, context, time, and conceptual operationalization. There is no reason *per se* to prioritize the sample as the source of an inferential problem. Indeed, we are more likely to lack variance on context and timing since these are constants in the experiment.

- In assessing the external validity of the sample, experimental realism (as opposed to mundane realism) is critical, and there is nothing inherent to the use of student subjects that reduces experimental realism.

- The nature of the sample—and the use of students—matters in certain cases. However, a necessary condition is: a heterogeneous (or moderated) treatment effect. Then the impact depends on:
    - If the heterogeneous effect is theorized, the sample only matters if there is virtually no variance on the moderator. If there is even scant variance, the treatment effect not only will be correctly estimated but may be estimated with greater confidence. The suitability of a given sample can be assessed (e.g., empirical variance can be analyzed).
    - If the heterogeneous effect is not theorized, it may be misestimated. However, even in this case, evaluating the bias is not straightforward because any sample will be inaccurate (since the "correct" moderated relationship is not being modeled).

- The range of heterogeneous, non-theorized cases may be much smaller than often thought. Indeed, when it comes to a host of politically relevant variables, student samples do not significantly differ from non-student samples.

- There are cases where student samples are desirable since they facilitate causal tests or make for more challenging assessments.

Our argument—that concerns about the sample come down more to a theoretical than an empirical issue—has a number of practical implications. First, we urge researchers to attend more to the potential moderating effects of the other dimensions of generalizabilty: context, time, and conceptualization. The last decade has seen an enormous increase in survey experiments, due in no small way to the availability of more representative samples. Yet scholars must account for the distinct context of the survey interview (e.g., Converse and Schuman 1974, Zaller 1992: 28). Mueller (1974: 1) explains that the survey "interview situation is an odd social

experience [about which] few people are accustomed." Sniderman et al. (1991: 265) elaborates that "the conventional survey interview, though well equipped to assess variations among individuals, is poorly equipped to assess variation across situations." Unlike most controlled lab settings, researchers using survey experiments have limited ability introduce contextual variations.

Second, we encourage the use of dual samples of students and non-students. The discovery of differences should lead to serious consideration of what drives distinctions (i.e., what is the underlying moderating dynamic and can it be modeled?). The few studies that explicitly compare samples (e.g., Gordon et al. 1986, James and Sonner 2001, Peterson 2001, Mintz et al. 2006, Dinah et al. 2009), while sometimes reporting differences, rarely explore the nature of the differences.[31] When dual samples are not feasible, researchers can take a second-best approach by utilizing question wordings that match those in general surveys (thereby facilitating comparisons).

Third, we hope for more discussion about the pros and cons of alternative modes of experimentation. While we recognize the benefits of using survey and/or field experiments, it is critical to assess the advantages in light of the full range of considerations. For example, the control available in laboratory experiments enables researchers to maximize experimental

[31] For example, Mintz et al. (2006) implemented an experiment, with both students and military officers, about counterterrorism decision-making. They find the two samples significantly differed, on average, in the decisions they made, the information they used, the decision strategies they employed, and the reactions they displayed. They (2006: 769) conclude that "student samples are often inappropriate, as empirically they can lead to divergence in subject population results." We would argue that this conclusion is pre-mature. While their results reveal on average differences between the samples, the authors leave unanswered why the differences exist. Mintz et al. (2006: 769) speculate that the differences may stem from variations in expertise, age, accountability, and gender. A thorough understanding of the treatment effect (which, as explained, is the goal of any experiment) would, thus, require exploration of these moderators, which may well be possible with both the military and student samples. Our simulation results suggest that even if the student sample exhibited limited variation on these variables, it could have isolated the same key treatment dynamics as would be found in the military sample.

realism (e.g., by using induced value or simply by more closely monitoring the subjects).[32] Similarly, there is less concern in laboratory settings about compliance × treatment interactions that become problematic in field experiments or spillover effects in survey experiments (Lee et al. 2005). In terms of external validity, increased control often affords greater ability to manipulate context and time, which, we have argued, deserve much more attention. Finally, when it comes to the sample, attention should be paid to the nature of any sample and not just student samples. This includes consideration of non-response biases in surveys (see Groves and Peytcheva 2008) and the impact of using "professional" survey respondents that are common in many web-based panels.[33]

Finally, we hope for greater disciplinary consideration of the practical and ethical issues involved in the construction of student subject pools, mandatory versus voluntary participation (e.g., Korn and Hogan 1992, Padilla-Walker et al. 2005), subject compensation, and the use of non-student convenience samples (Kam et al. 2007). Such conversations should be attuned to the pedagogical potential of experimental participation, which can, with proper follow-up, include demonstrating the working of social science research.[34] Similar questions concern how participation affects subsequent subject behavior (e.g., Stevens and Ash 2001, Bender 2007) including willingness to participate in subsequent studies (e.g., Porter et al. 2003).[35] Additionally, there are a number of sampling and statistical techniques relevant to drawing

[32] It is important to draw a distinction between laboratory experiments and classroom experiments where researchers administer experiments during or after classes. These latter contexts can sometimes work effectively, but they also raise other challenges in terms of controlling the setting.

[33] The use of professional, repeat respondents raises similar issues to those caused by repeated use of participants from a subject pool (see, e.g., Stevens and Ash 2001).

[34] There is a related, long-standing debate about the pros and cons of using experiments in educational settings (e.g., concerning curriculum) (see Cook 2003).

[35] Related to this concern is the impact of deception in experiments on subsequent experimental behavior and participation. Economic experimental laboratories prohibit deception due to concern that it threatens experimental realism; this makes the construction of subject pools that can be shared by economic and psychological approaches (where mild deception is often common) impossible. See Dickson's chapter in this volume.

inferences from common laboratory studies, that have received virtually no attention in political science (e.g., purposive sampling, Pitman test; see, e.g., Shadish et al. 2002, Hedges 2009, Keele et al. 2009).

We have made a strong argument for the increased usage and acceptance of student subjects, suggesting that the burden of proof be shifted from the experimenter to the critic (also see Friedman and Sunder 1994: 16). We recognize that many will not be persuaded; however, at the very least, we hope to have stimulated increased discussion about why and when student subjects may be problematic.

References

Anderson, Craig A., and Brad J. Bushman. 1997. "External Validity of 'Trivial' Experiments: The Case of Laboratory Aggression." *Review of General Psychology* 1:19-41.

Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. New York: Free Press.

Arceneaux, Kevin, and Martin Johnson. 2008. "Choice, Attention, and Reception in Political Communication Research." Presented at the annual meeting of the International Society for Political Psychology, Paris, France, July 9-12.

Aronson, Elliot, Marilynn B. Brewer, and J. Merill Carlsmith. 1985. "Experimentation in Social Psychology." In Gardner Lindzey and Elliot Aronson, eds., *Handbook of Social Psychology* 3rd Edition. New York: Random House.

Aronson, Elliot, and J. Merill Carlsmith. 1968. "Experimentation in Social Psychology." In Gardner Lindzey and Elliot Aronson, eds., *Handbook of Social Psychology* 2nd Edition. Reading, MA: Addison-Wesley.

Aronson, Elliot, Timothy D. Wilson, Marilynn B. Brewer. 1998. "Experimentation in Social Psychology." In Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, eds., *The Handbook of Social Psychology*. 4th Edition. Boston: The McGraw-Hill Companies, Inc.

Barabas, Jason, and Jennifer Jerit. 2009. "The External Validity of Treatments: A Comparison of Natural and Survey Experiments." Unpublished Paper, Florida State University.

Baron, Reuben M., and David A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51: 1173-1182.

Bartels, Larry M. 1993. "Messages Received: The Political Impact of Media Exposure." *American Political Science Review* 87: 267-285.

Bender, Timothy A. 2007. "Time of Participation Effect and Grade-Orientation." *Personality and Individual Differences* 43: 1175-1183.

Benz, Matthias, and Stephan Meier. 2008. "Do People Behave in Experiments as in the Field?: Evidence from Donations." *Experimental Economics* 11: 268-281.

Berkowitz, Leonard, and Edward Donnerstein. 1982. "External Validity is More Than Skin Deep: Some Answers to Criticisms of Laboratory Experiments." *American Psychologist* 37: 245-257.

Bizer, George. Y., Jon A. Krosnick, Allyson L. Holbrook, S. Christian Wheeler, Derek D. Rucker, and Richard E. Petty. 2004. "The Impact of Personality on Cognitive, Behavioral, and Affective Political Processes: The Effects of Need to Evaluate." *Journal of Personality* 72: 995-1027.

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.

Campbell, Angus, Philip Converse, Warren Miller, and Donald Stokes. 1960 [1980 reprint]. *The American Voter.* Chicago: University of Chicago Press.

Campbell, Donald T. 1969. "Prospective: Artifact and Control." In Robert Rosenthal and Robert Rosnow, eds., *Artifact in Behavioral Research*. New York: Academic Press.

Campbell, Donald T., and Donald W. Fiske. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56: 81-105.

Chong, Dennis, and James N. Druckman. 2007a. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101: 637-655.

Chong, Dennis, and James N. Druckman. 2007b. "Framing Theory." *Annual Review of Political Science* 10: 103-126.

Converse, Jean M., and Howard Schuman. 1974. *Conversations at Random: Survey Research as Interviewers See It.* New York: Wiley.

Cook, Thomas D. 2003. "Why Have Educational Evaluators Chosen Not to Do Randomized Experiments?" *The Annals of the American Academy of Political and Social Sciences* 589: 114-149.

Dickhaut, John W., J. Leslie Livingstone, and David J. H. Watson. 1972. "On the Use of Surrogates in Behavioral Experimentation." *The Accounting Review* 47, Suppolment: 455-471.

Dinah, Pura T. Depositario, Rodolfo M. Nayga Jr., Ximing Wu, and Tiffany P. Laude. 2009. "Should Students Be Used as Subjects in Experimental Auctions?" *Economic Letters* 102: 122-124.

Druckman, James N. 2001. "On The Limits of Framing Effects." *Journal of Politics* 63: 1041-1066.

Druckman, James N. 2009. "Competing Frames in a Campaign." Unpublished Manuscript, Northwestern University.

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: 627-636.

Fowler, James H. and Cindy D. Kam. 2007. "Beyond the Self: Social Identity, Altruism, and Political Participation." *Journal of Politics* 69: 813-827.

Franklin, Charles H. 1991. "Efficient Estimation in Experiments." *The Political Methodologist* 4: 13-15.

Friedman, Milton. 1953. *Essays in Positive Economics.* Chicago: University of Chicago Press.

Friedman, Daniel, and Shyam Sunder. 1994. *Experimental Economics: A Primer for Economists*. New York: Cambridge University Press.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15: 1-20.

Gerber, Alan, James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2007. "The Influence of Television and Radio Advertising on Candidate Evaluations: Results from a Large Scale Randomized Experiment." Unpublished paper, Yale University.

Gerber, Alan S., and Donald P. Green. 2008. "Field Experiments and Natural Experiments." In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Gerring, John. 2001. *Social Science Methodology: A Critical Framework*. Cambridge, UK: Cambridge University Press.

Gordon, Michael E., and L.Allen Slade, and Neal Schmitt. 1986. "The 'Science of the Sophomore' Revisited: From Conjecture to Empiricism." *Academy of Management Review* 11: 191-207.

Groves, Robert M., and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72: 167-189.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. New York: Cambridge University Press.

Hedges, Larry. 2009. "Improving Generalizations from Social Experiments." Unpublished paper, Institute for Policy Research, Northwestern University.

Hunter, John E., Frank L. Schmidt., and Gregg B. Jackson. 1982. *Meta-analysis: Cumulating Research Findings Across Studies*. Beverly Hills: Sage.

Jacoby, William G. 2000. "Issue Framing and Public Opinion on Government Spending." *American Journal of Political Science* 44: 750-767.

James, William L., and Brenda S. Sonner. 2001. "Just Say No to Traditional Student Samples." *Journal of Advertising Research* 41: 63-71.

Kam, Cindy D. 2005. "Who Toes the Party Line?: Cues, Values, and Individual Differences." *Political Behavior* 27: 163-182.

Kam, Cindy D. 2007. "When Duty Calls, Do Citizens Answer?" *Journal of Politics* 69: 17-29.

Kam, Cindy, and Robert J. Franzese. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor: University of Michigan Press.

Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research." *Political Behavior* 29: 415-440.

Kinder, Donald R. 1998. "Communication and Opinion." *Annual Review of Political Science* 1: 167-197.

Korn, James H., and Kathleen Hogan. 1992. "Effect of Incentives and Aversiveness of Treatment on Willingness to Participate in Research." *Teaching of Psychology* 19: 21-24.

Kruglanski, Arie W. 1975. "The Human Subject in the Psychology Experiment: Fact and Artifact." In Leonard Berkowitz, ed., *Advances in Experimental Social Psychology*. New York: Academic Press.

Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. Cambridge, UK: Cambridge University Press.

Lee, Daniel J., John E. Transue, and John Aldrich. 2005. "Treatment Spillover Effects Across Survey Experiments." Presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 7-10.

Levitt, Steven D., and John A. List. 2007. "Viewpoint: On the Generalizability of Lab Behavior to the Field." *Canadian Journal of Economics* 40: 347-370.

Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65: 682-93.

Liyanarachchi, Gregory A. 2007. "Feasibility of Using Student Subjects in Accounting Experiments: A Review." *Pacific Accounting Review* 19: 47-67.

Loewen, Peter. 2009. "Dictators and Purses: Altruism and Support for Greater Public Spending." Unpublished Paper, University of British Columbia.

Lupia, Arthur, and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge, UK: Cambridge University Press.

MacDonald, Paul. 2003. "Useful Fiction or Miracle Maker: The Competing Epistemological Foundations of Rational Choice Theory." *American Political Science Review* 97: 551-565.

McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10: 325-342.

McGrath, Joseph Edward, Joanne Martin, and Richard A. Kulka. 1982. *Judgment Calls in Research*. Beverly Hills: Sage.

McGraw, Kathleen M., and Valerie Hoekstra. 1994. "Experimentation in Political Science." *Research in Micropolitics* 3: 3-29.

Mintz, Alex, Steven B. Redd, and Arnold Vedlitz. 2006. "Can we generalize from student experiments to the real world in political science, military affairs, and international relations?" *Journal of Conflict Resolution* 50: 757-776.

Mook, Douglas G. 1983. "In Defense of External Invalidity." *American Psychologist* 38:379-387.

Morton, Rebecca B., and Kenneth C. Williams. 2008. "Experimentation in Political Science." In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Mueller, John E. 1973. *War, Presidents, and Public Opinion*. New York: John Wiley & Sons.

Niemi, Richard G., and M. Kent Jennings. 1991. "Issues and Inheritance in the Formation of Party Identification." *American Journal of Political Science* 35: 970-988.

Padilla-Walker, Laura M., Byron L. Zamboanga, Ross A. Thompson, and Larissa A. Schmersal. 2005. "Extra Credit As Incentive for Voluntary Research Participation." *Teaching of Psychology* 32: 150-153.

Peterson, Robert A. 2001. "On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-analysis." *Journal of Consumer Research* 28: 450-461.

Plott, Charles R. 1991. "Will Economics Become an Experimental Science?" *Southern Economic Journal* 57: 901-919.

Porter, Stephen R., Michael E. Whitcomb, and William H. Weitzer. 2003. "Multiple Surveys of Students and Survey Fatigue." *New Directions for Institutional Research* 121: 63-73.

Roth, Alvin E. 1995. "Introduction to Experimental Economics." In John H. Kagel and Alvin E. Roth, eds., *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

Sears, David O. "College Sophomores in the Laboratory: Influence of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51: 515-530.

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

Sherman, Richard C., Amy M. Buddie, Kristin L. Dragan, Christian M. End, and Lila J. Finney. 1999. "Twenty Years of *PSPB*: Trends in Content, Design, and Analysis." *Personality and Social Psychology Bulletin* 25: 177-187

Simon, Herbert A. 1963. "Problems of Methodology Discussion." *American Economic Review Proceedings* 53: 229-231.

Simon, Herbert A. 1979. "Rational Decision Making in Business Organizations." *The American Economic Review* 69: 493-513.

Slothuus, Rune. 2009. "The Political Logic of Party Cues in Opinion Formation." Presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 2-5.

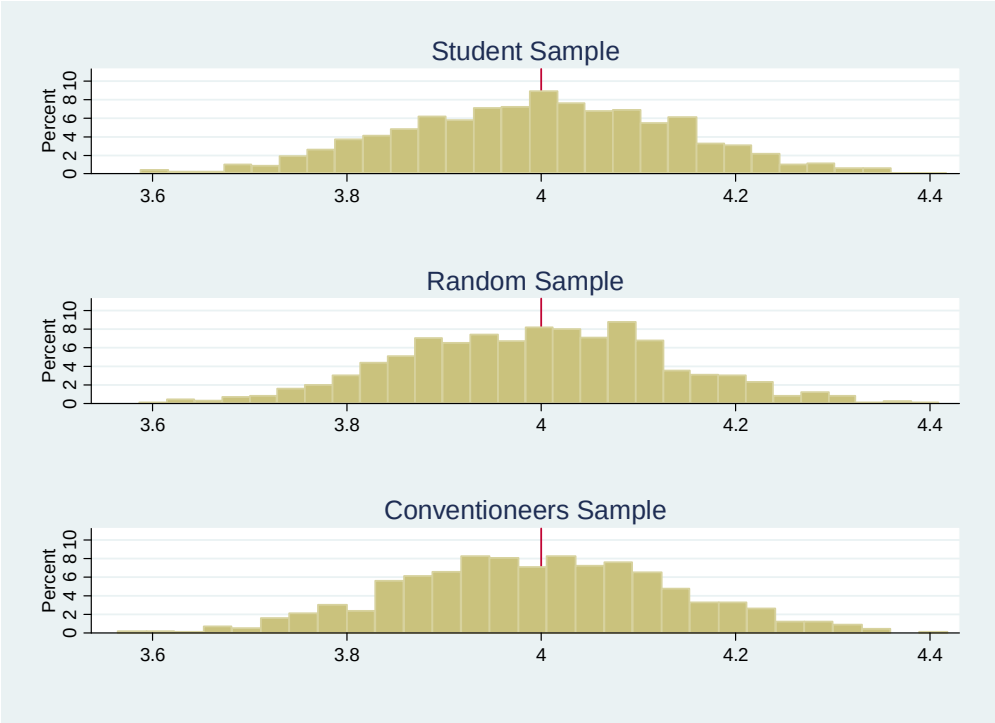
Sniderman, Paul M., and Douglas B. Grob. 1996. "Innovations in Experimental Design in Attitude Surveys." *Annual Review of Sociology* 22: 377-399.

Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In Willem E. Saris, and Paul M. Sniderman (eds.), *Studies in Public Opinion*. Princeton, NJ: Princeton University Press.

Stevens, Charles D., and Ronald A. Ash. 2001. "The Conscientiousness of Students in Subject Pools: Implications of 'Laboratory' Research." *Journal of Research in Personality* 35: 91-97.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Figure 1. Sampling distribution of b_T, single treatment effect



Note: 1,000 iterations, estimated using Eq [1]

Figure 2. Hypothetical heterogeneous (linear) treatment effect

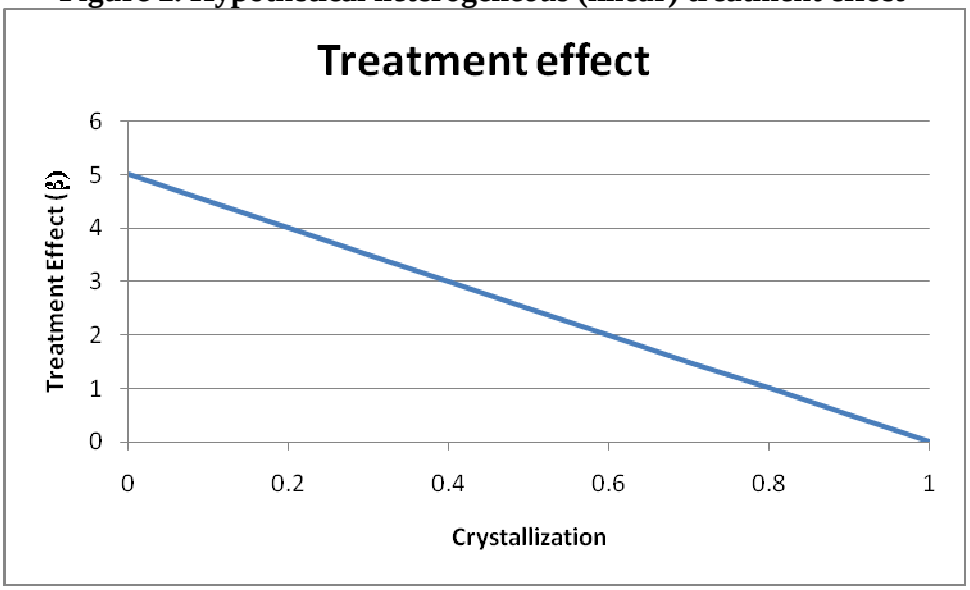
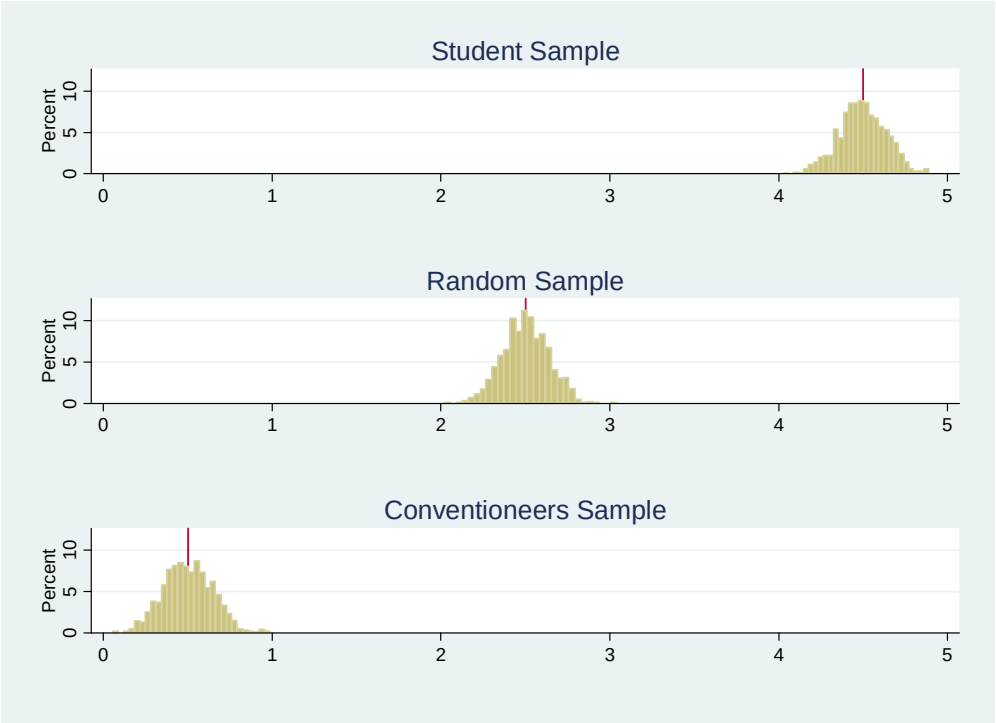
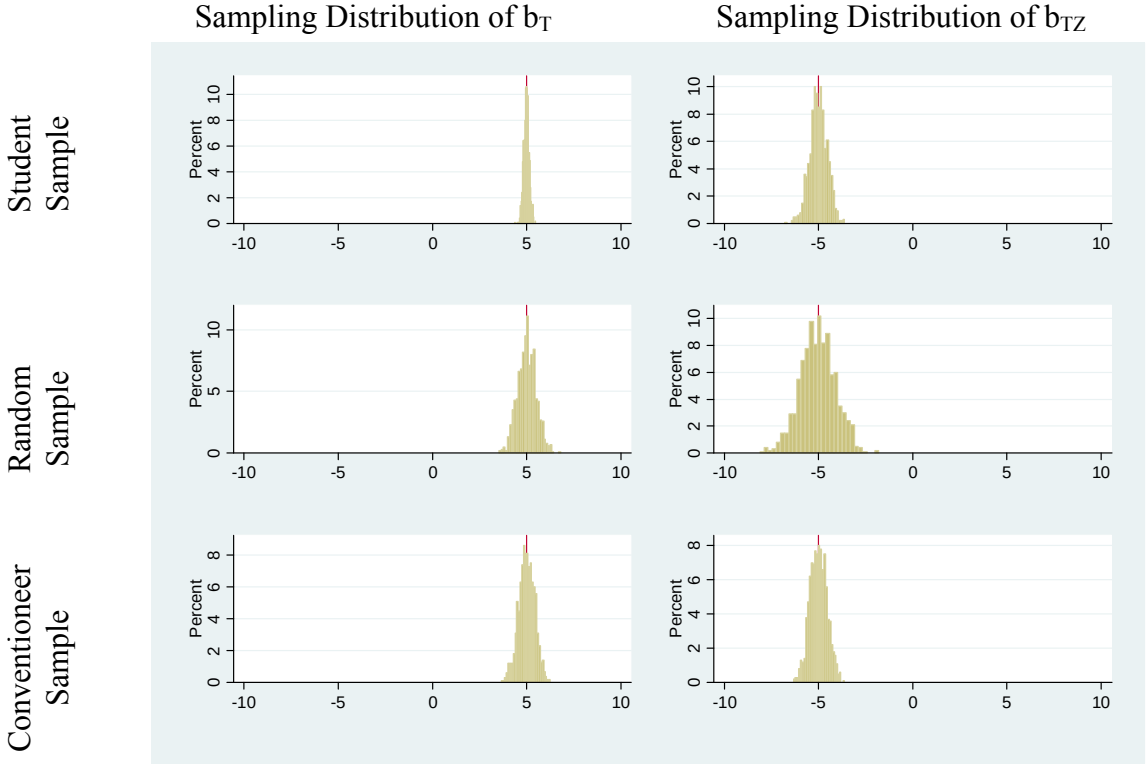


Figure 3. Sampling distribution of b_T, heterogeneous treatment effects



Note: 1,000 iterations, estimated using Eq [1]

Figure 4. Sampling distributions of b_T and b_{TZ}, heterogeneous treatment effects



Note: 1,000 iterations, estimated using Eq [3]

Table 1. Sampling distribution moments for b_T and b_{TZ}, heterogeneous treatment effects

	b_T Mean (s.d.)	b_{TZ} Mean (s.d.)
Student Sample	5.01 (0.15)	-5.01 (0.48)
Random Sample	5.00 (0.49)	-5.00 (0.90)
Conventioneers Sample	4.99 (0.46)	-4.99 (0.49)

Table 2. Comparison of students versus non-student general population
Means with standard errors in parentheses

	Students	Non-student General Population	*p-value*
Partisanship	0.47 (0.02)	0.45 (0.01)	*ns (not significant)*
Ideology	0.50 (0.01)	0.52 (0.01)	*ns*
Religious Attendance	0.56 (0.02)	0.50 (0.01)	<0.01
Importance of Religion	0.63 (0.02)	0.62 (0.02)	*ns*
Limited Government	0.35 (0.03)	0.33 (0.02)	*ns*
Homosexuality as a way of life	0.60 (0.03)	0.62 (0.02)	*ns*
Contribution of immigrants to society	0.62 (0.03)	0.63 (0.02)	*ns*
Social trust	0.34 (0.03)	0.33 (0.02)	*ns*
Follow politics	0.68 (0.02)	0.65 (0.01)	*ns*
Discuss politics	0.75 (0.01)	0.71 (0.01)	*ns*
Political information (0 to 6 correct)	2.53 (0.11)	1.84 (0.07)	<0.01
Newspaper use (0 to 7 days)	2.73 (0.14)	2.79 (0.11)	*ns*
National TV news (0 to 7 days)	3.28 (0.15)	3.63 (0.10)	<0.05
News radio (0 to 7 days)	2.47 (0.16)	2.68 (0.11)	*ns*
Web news (0 to 7 days)	3.13 (0.16)	2.18 (0.10)	<0.01
Overall media use	2.90 (0.09)	2.83 (0.06)	*ns*

Weighted analysis.
See the appendix for variable coding and question text.
Source: 2006 Civic and Political Health Survey.

Appendix: Question Wordings and Codings for Table 2

The 2006 Civic and Political Health Survey consists of telephone and web interviews with 2,232 individuals 15 and older living in the continental US. The sampling frame included youth ages 15-25 (N=1674) and adults 26 and over (N=547). For our purposes, we confine the sample to individuals aged 18 and over. The key comparison we make is between students and the non-student general population. Students were defined as those respondents currently in college (undergraduate) according to the question: "Are you currently enrolled in school?" (Q.8). The non-student general population was defined as those indicating any other response to the enrollment question. All analyses were conducted using the probability weight variable ("weight").

Question wordings and coding:

Partisanship

(0=strong Republican to 1=strong Democrat)

Seven-point partisanship scale calculated by combining responses to three questions:

- In politics today, do you consider yourself a Democrat, Republican, Independent, or something else? (Q.106)
- [If Democrat or Republican in Q.106] Do you consider yourself a strong [Republican/Democrat] or or not so strong...? (Q.107)
- [If not Democrat or Republican in Q.106] Do you lean more toward the Democratic party or more toward the Republican party? (Q.108)

Ideology

(0=Very Conservative to 1 = Very Liberal)

In general, would you describe your political views as very conservative (0), conservative (.25), moderate (.5), liberal (.75), or very liberal (1)? (Q.109)

Religious Attendance

(0=Never to 1=More than once a week)

Aside from weddings and funerals how often do you attend religious services: more than once a week (1), once a week (.8), once or twice a month (.6), a few times a year (.4), seldom (.2), or never (0)? (Q.114)

Importance of Religion

(0=Not very important to 1=Very important)

How important would you say religion is in your own life: very important (1), fairly important (.5), or not very important (0)? (Q.115)

Limited Government

(0=Government should do more; 1 = Government does too many things)

Which statement do you agree with (Q.88):

- Government should do more to solve problems (1), or
- Government does too many things better left to business and individuals (0)

Homosexuality as a way of life

(0=should be discouraged; 1 = should be accepted)

Which statement do you agree with (Q.95):

- Homosexuality is a way of life that should be accepted by society (1), or
- Homosexuality is a way of life that should be discouraged by society (0)

Contribution of immigrants to society

(0=immigrants a burden; 1 = immigrants a strength)

Which statement do you agree with (Q.96):

- Immigrants today strengthen our country because of their hard work and talents (1), or
- Immigrants today are a burden on our country because they take our jobs, housing & health care (0)

Social trust
(0=out for themselves; 1=try to be helpful)
Which statement do you agree with (Q.86):

- Most of the time people try to be helpful (1), or
- Most of the time people are just looking out for themselves (0)

Follow politics
(0=never to 1=most of the time)
Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election or not. Others aren't that interested. Do you follow what's going on in government and public affairs most of the time (1), some of the time (.67), rarely (.33), or never (0)? (Q.49)

Discuss politics
(0=never to 1=very often)
How often do you talk about current events or things you have heard about in the news with your family and friends: very often (1), sometimes (.67), rarely (.33), or never (0)? (Q.50)

Political information scale
(0 correct to 6 correct)
Additive index of six measures (Cronbach's α=0.70)

- Would you say that one of the parties is more conservative than the other on the national level? If yes, which is more conservative? (Q.100)
  - Yes, Republican Party more conservative (1)
  - All other responses (0)
- How much of a majority is required for the U.S. senate and house to override a presidential veto? (Q.101) [Open-ended response, coded 1 for two-thirds or 67%; and 0 otherwise]
- Which of the following best describes who is entitled to vote in federal elections: residents (0), taxpayers (0), legal residents (0), citizens (1). (Q.102)
- Please name one of the president's cabinet secretaries and identify the department they represent (Q.103a). [Open-ended response, coded 1 for naming cabinet member and department; and 0 otherwise]
- Please name another one of the president's cabinet secretaries and identify the department they represent (Q.103b). [Open-ended response, coded 1 for naming cabinet member and department; and 0 otherwise]
- Five countries have permanent seats on the security council of the United Nations. Which of these countries can you name? (Q.104_1) [First open-ended response, coded 1 for naming France, PRC (China), Russian Federation, United Kingdom, or United States; and 0 otherwise]

Newspaper use
(Count: 0 to 7)
Over the past seven days, please tell me on how many days you read a newspaper? (Q.56)

National TV news
(Count: 0 to 7)
Over the past seven days, please tell me on how many days you watched the national news on television? (Q.58)

News radio
(Count: 0 to 7)
Over the past seven days, please tell me on how many days you listened to the news on the radio? (Q.59)

Web news
(Count: 0 to 7)
Over the past seven days, please tell me on how many days you read news on the internet? (Q.60)

Overall media use
Average of newspaper, national TV, news radio, web news variables.