Cambridge Handbook of Experimental Political Science

Edited by

JAMES N. DRUCKMAN

Northwestern University

DONALD P. GREEN

Yale University

JAMES H. KUKLINSKI

University of Illinois at Urbana–Champaign

ARTHUR LUPIA

University of Michigan

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521174558

First published 2011

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Cambridge Handbook of Experimental Political Science / [edited by] James N. Druckman,
Donald P. Green, James H. Kuklinski, Arthur Lupia.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-19212-5 (hardback) – ISBN 978-0-521-17455-8 (paperback)
1. Political science – Methodology. 2. Political science – Research.
3. Political science – Experiments. I. Druckman, James N., 1971– II. Title.
JA71.C325 2011
320.072 – dc22 2010044869

ISBN 978-0-521-19212-5 Hardback
ISBN 978-0-521-17455-8 Paperback

ıgley, Jonathan A.
tto, and Dominic
ɔnoamine Oxidase
Behavioral Aggres-
." Proceedings of the
106: 2118–23.
anath, and Ernest
Shocks and Civil
ariable Approach."
12: 725–53.
ience to Authority.

Scalable Protocols
eld Experiments."

illiam J. Dickson.
orker. Cambridge,
ss.
ıction to Experi-
book of Experimen-
Kagel and Alvin
ıceton University

D. Cook, and
02. Experimental
ns for Generalized
ghton-Mifflin.
ackie. 1995. Social

, Nina G. Curley,
F. Camerer, and
Thinking Like a
Individuals' Loss
ational Academy of

P. Green. 2011.
ation in Political
American Journal

ahneman. 1974.
: Heuristics and

Approaches to
American Behav-

CHAPTER 4

Students as Experimental Participants

A Defense of the "Narrow Data Base"

James N. Druckman and Cindy D. Kam

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. 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 generalizability, 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, 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 that such situations are relatively limited; any convenience sample poses a problem *only* when the size of an experimental treatment effect depends on a characteristic on which the convenience sample has virtually no variance. 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, the burden of proof – articulating and demonstrating that student subjects pose an inferential problem – should lie with critics rather than experimenters.

1. The "Problem" of Using Student Subjects

Although internal validity may be the sine qua non of experiments, most researchers use experiments to make *generalizable* causal

We thank Kevin Arceneaux, Don Green, Jim Kuklinski, Peter Loewen, and Diana Mutz for helpful advice, and Samara Klar and Thomas Leeper for research assistance.

inferences (Shadish, Cook, and Campbell 2002, 18–20). For example, suppose one implements a laboratory study with students and finds a causal connection between the experimental treatment (e.g., a media story about welfare) and an outcome of interest (e.g., support for welfare). An obvious question is whether the relationship found in the study exists within a heterogeneous population, in various contexts (e.g., a large media marketplace), and over time. This is an issue 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) explains, "External validity . . . tend[s] to preoccupy critics of experiments. This near obsession . . . tend[s] to be used to dismiss experiments" (334).

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."[1] Many political scientists simply assume that "a student sample lacks external generalizability" (Kam, Wilking, and Zechmeister 2007, 421). Gerber and Green (2008) explain, "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 2008, 358). In short, social scientists in general and political scientists in particular view student subjects as a major hindrance to drawing inferences from experimental studies.

Assessing the downside of using student subjects has particular current relevance. First, many political science experiments use student subjects; for example, Kam et al. report that from 1990 through 2006, one fourth of experimental articles in general political science journals relied on student subjects, whereas more than seventy percent did so in more specialized journals (Kam et al. 2007, 419–20; see also Druckman et al. 2006). 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 nonstudent participants. 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 recruit nonstudent 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.

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). 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.[2]

Moreover, when it comes to generalization from a series of studies, the goal is to generalize across *multiple* dimensions. External

1 Through 2008, Sears' (1986) article was 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.

2 This is consistent with a Popperian approach to causation that suggests causal hypotheses are never confirmed and evidence accumulates via multiple tests, even if these tests have limitations. Campbell (1969) 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" (361).

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 that, consequently, they can moderate causal relationships. One powerful recent example comes from the political communication literature; a number of experiments, using both student *and* nonstudent subjects, show that when exposed to political communications (e.g., in a laboratory), individuals' opinions often reflect the content of those communications (e.g., Kinder 1998; Chong and Druckman 2007b). The bulk of this work, however, ignores the contextual reality that people outside 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 (i.e., the effects 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.[3]

Timing also matters. Results from experiments implemented at one time may not hold at other times, given the nature of world events. Gaines, Kuklinski, and Quirk (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 in this volume). 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 (i.e., respondents who regularly read newspaper coverage of the campaign) in the economic information condition did not significantly differ from attentive individuals in the control group.[4] The noneffect 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 noneffect lacked external validity, not because of the sample but because it failed to account for the timing of the treatment and what had occurred prior to that time (also see Slothuus 2009).[5]

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).

In short, external validity does *not* simply refer to whether a specific study, if rerun on a different sample, would provide the same results. It refers more generally to whether "conceptually equivalent" (Anderson and Bushman 1997, 21) 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,

3 A related example comes from Barabas and Jerit's (2010) study, which compares the impact of communications in a survey experiment against analogous dynamics that occurred in actual news coverage. They find that the survey experiment vastly overstated the effect, particularly among certain subgroups. 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 misestimates the impact of communications on public opinion.

4 For reasons explained in his paper, Druckman (2009) also focuses on individuals more likely to have formed prior opinions about the casino.

5 Another relevant timing issue concerns the duration of any experimental treatment effect (see, e.g., Gaines et al. 2007; Gerber et al. 2007).

the "to what" refers to behavior as observed outside the study, but this is not always the case. Experiments have different purposes; Roth (1995) identifies three nonexclusive roles that experiments can play: *search for facts*, *speaking to theorists*, or *whispering in the ears of princes*, (22) which facilitates "the dialogue between experimenters and policy-makers" (see also Guala 2005, 141–60). These types likely differ in the target of generalization. Of particular relevance is that theory-oriented experiments are typically 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) explains, "The experiment should be judged by the lessons it teaches about the theory and not by its similarity with what nature might have happened to have created" (906). This echoes Mook's (1983) argument that much experimental work is aimed at developing and/or testing a theory, not at establishing generalizability. Experiments that are designed to demonstrate "what can happen" (e.g., Milgram 1963; Zimbardo 1973) can still be useful, even if they do not mimic everyday life.[6] In many instances, the nature of the subjects in the experiments are of minimal relevance, particularly given experimental efforts to ensure that their preferences and/or motivations match those in the theory (e.g., see Dickson's chapter in this volume).

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 a 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. Although this is best done over a series of studies, we acknowledge the need to assess the strengths of a particular study with respect to external validity. Individual studies can be evaluated in at least two ways (Aronson and Carlsmith 1968; Aronson, Brewer, and Carlsmith 1985; Aronson, Wilson, and Brewer 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, Brewer, and Carlsmith 1985, 485). Second, mundane realism concerns "the extent to which events occurring in the research setting are likely to occur in the normal course of the subjects' lives, that is, in the 'real world'" (Aronson et al. 1985, 485).[7]

Much debate about samples focuses on mundane realism. When student subjects do not match the population to which a causal inference is intended, many conclude that the study has low external validity. Emphasis on mundane realism, however, is misplaced (e.g., 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., Dickhaut, Livingstone, and Watson 1972, 477; Liyanarachchi 2007, 56).[8] In contrast, at worst, low levels

6 Aronson, Wilson, and Brewer (1998) 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" (132).

7 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; Simon 1963, 1979, 475–76; also see MacDonald 2003).

8 By "seriously," we mean analogous to how individuals treat the same stimuli in the settings to which one hopes to generalize (and not necessarily "serious" in a technical sense). We do not further discuss steps that can be taken to ensure experimental realism because this moves into the realm of other design issues

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) 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) (57).[9] That said, beyond superficiality, we recognize that student subjects – although having no ostensibly relevant connection with experimental realism[10] – may limit mundane realism that constrains generalizations of a particular study. This occurs when characteristics of the subjects affect the nature of the causal relationship being generalized. When this happens, and with what consequences, are questions to which we now turn.

2. Statistical Framework

In this section, we examine the use of student samples from a statistical point of view. This allows us to specify the conditions under which student samples might constrain causal generalization (in the case of a single experiment). Our focus, as in most political science analyses of experimental data, is on the magnitude of the effect of some experimental treatment, T, on an attitudinal or behavioral dependent measure, y. Suppose, strictly for presentational purposes, we are interested in the effect of a persuasive communication (T) on a subject's poststimulus policy opinion (y). 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.[11] Suppose the true data-generating process features a homogeneous treatment effect:

$$y_i = \beta_0 + \beta_T T_i + \varepsilon_i. \tag{1}$$

Assuming that all assumptions of the classical linear regression model are met, the ordinary least squares estimate for β_T is unbiased, consistent, and efficient.[12] The results derived from estimation on a given sample would be fully generalizable to those that would result from estimation on any other sample.

Specific samples will differ in their distributions of individual covariates. Continuing with our running example, samples may differ in the distribution of *attitude crystallization* (i.e., an attitude is increasingly crystallized when it is stronger and more stable).[13] Student samples may yield a disproportionately large group of subjects who are low in crystallization. A random sample from the general population might generate a group that is normally distributed and centered at the

(e.g., subject payments, incentives; see Dickson's chapter in this volume).

9 Berkowitz and Donnerstein (1982) explain, "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 generalizability of an experiment's outcome than does the sample's demographic representatives or the setting's surface realism" (249).

10 This claim is in need of empirical evaluation because it may be that students are more compliant, and this may have an impact on realism.

11 For ease of exposition, our example only has one treatment group. The lessons easily extend to multiple treatment groups.

12 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, $\mathbf{X}$). Under random assignment, the expected covariance between the treatment and $\mathbf{X}$ is zero. Hence, if we were to estimate the model without $\mathbf{X}$, then omitted variable bias would technically not be an issue. If the data-generating process does include $\mathbf{X}$, and even though we might not have an omitted variable bias problem, including $\mathbf{X}$ in the model may still be advisable. Inclusion of relevant covariates (i.e., 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 $\mathrm{Cov}(\mathbf{X},T) = 0$. In any given sample, $\mathrm{Cov}(\mathbf{X},T)$ 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) 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" (172; see also Bowers' chapter in this volume).

13 This example is inspired by Sears' (1986) discussion of "Uncrystallized Attitudes."

middle of the range. A sample from politically active individuals (e.g., conventioneers) might result in a group that is disproportionately high in crystallization.[14]

Consider the following samples with varying distributions on attitude crystallization. In all cases, $n = 200$, and treatment is randomly assigned to half the cases. Attitude crystallization ranges from 0 (low) to 1 (high). Consider a "student sample," where ninety percent of the sample is at a value of "0" and ten percent of the sample is at a value of "1." Consider a "random sample," where the sample is normally distributed and centered on 0.5, with standard deviation of 0.165. Finally, consider a "conventioneers sample," where ten percent of the sample is at a value of "0" and ninety percent of the sample is at a value of "1."[15]

Suppose the true treatment effect (β_T) takes a value of "4." We set up a Monte Carlo experiment 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 4.1.

The results demonstrate that when the true data-generating process produces a single treatment effect, estimates from *any* sample 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 entire population), then any convenience sample should produce an unbiased estimate of that single treatment effect, and, thus, the results from any convenience sample should generalize easily to any other group.*

Suppose, however, the "true" underlying data-generating process contains a heterogeneous treatment effect: that is, the effect of the treatment is moderated[16] by individual-level characteristics. The size of the treatment effect might depend on some 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).

As one method of overcoming this issue, a researcher can randomly sample experimental subjects. By doing so, 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. (Shadish et al. 2002, 91)*

Although random sampling has advantages for external validity, Shadish et al. note that "it is so rarely feasible in experiments" (91). 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 say a bit more about this possibility later in the chapter. 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 the challenges a researcher

14 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.

15 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.

16 See Baron and Kenny (1986) for the distinction between moderation and mediation. 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. 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 in this volume.

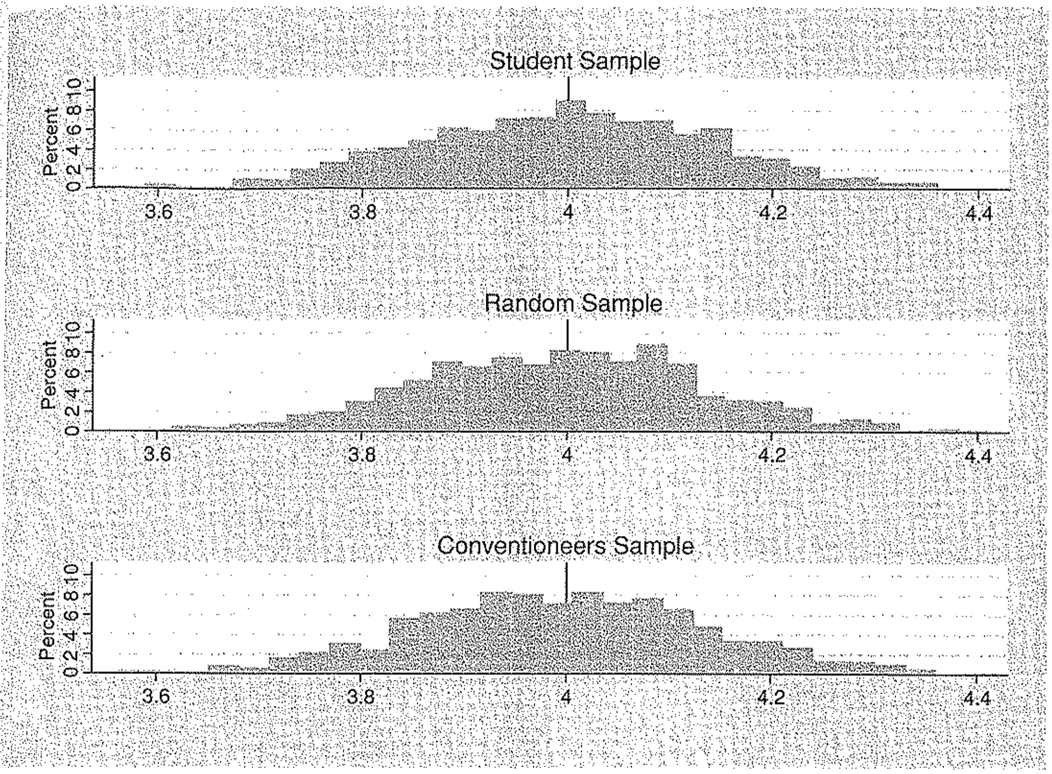

Figure 4.1. Sampling Distribution of b_T, Single Treatment Effect
Note: 1,000 iterations, estimated using Equation (1).

using a convenience sample might face in this framework.

We revise our data-generating process to reflect a heterogeneous treatment effect by taking Equation (1) and modeling how some individual-level characteristic, Z (e.g., attitude crystallization), influences the magnitude of the treatment effect:

$$\beta_1 = \gamma_{10} + \gamma_{11} Z_i. \tag{2}$$

We also theorize that Z might influence the intercept:

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

Substituting into Equation (1),

$$\begin{aligned} y_i &= (\gamma_{00} + \gamma_{01} Z_i) + (\gamma_{10} + \gamma_{11} Z_i)\, T_i + \varepsilon_i \\ y_i &= \gamma_{00} + \gamma_{01} Z_i + \gamma_{10} T_i + \gamma_{11} Z_i^* T_i + \varepsilon_i. \end{aligned} \tag{3}$$

If our sample includes sufficient variance on this moderator, and we have ex ante theorized that the treatment effect depends on 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 also misestimate the on-average effect depending on what specific range of Z is present in our sample.

The question of generalizing treatment effects reduces to asking whether there is a single treatment effect or a set of treatment effects, the size of which depends on some (set of) covariate(s). Note that this is 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 wonder whether the causal relationship detected in a student sample experiment would be systematically different from the causal relationship in the general population.

Revisiting our running example, suppose we believe that a subject's level of attitude crystallization (Z) influences the effect of a persuasive communication (T) on a subject's poststimulus policy opinion (y). The more crystallized someone's attitude is, the smaller the treatment effect should be, whereas the less crystallized someone'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:

$$\gamma_{00} = 0$$
$$\gamma_{01} = 0$$
$$\gamma_{10} = 5$$
$$\gamma_{11} = -5$$

Let Z, attitude crystallization, range 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.

First, consider what happens when we estimate Equation (1), the simple (but theoretically incorrect, given that it fails to model the moderating effect) model that looks for the "average" treatment effect. We estimated this model 1,000 times, each time drawing a new ε term. We repeated this process for each of the three samples. The results appear in Figure 4.2.

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 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 the case of a conventioneers sample, where high levels of crystallization depress the treatment effect, we systematically underestimate the treatment effect, relative to the estimates obtained from the general population.

We have obtained three different results across the samples because we estimated a model based on Equation (1). Equation (1) should only be estimated when the data-generating process produces a *single* treatment effect: the value of β_1. However, 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 in Figure 4.2 provide the "average" treatment effect, which depends directly on the mean value of Z within a given sample: $5 - 5^*E(Z)$.

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, 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 (i.e., the treatment depends on an individual-level moderator), then we should embed this *theoretical model* into our *statistical model*.

To do so, 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 samples. The results appear in Figure 4.3.

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 β_T and β_{TZ}, regardless of which sample is used. We uncover unbiased point estimates even where only ten percent of the sample provides key variation on Z (student sample and conventioneers sample).

Second, 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

n (1). Equation (1)
ed when the data-
uces a *single* treat-
f β_1. However, we
iated Equation (1)
rating process pro-
nt effects (governed
$5Z_i$). The sampling
2 provide the "aver-
ich depends directly
thin a given sample:

one sample more
sults from another
l. (2002) note, con-
n a random sample
e" treatment effect;
he results from the
more desirable than
er two convenience
hree sets of results
uncture between the
and the true data-
ve have a theoreti-
the data-generating
(i.e., the treatment
al-level moderator),
this *theoretical model*

to Equation (3) and
oo times, each time
Ve repeated this pro-
of the three samples.
gure 4.3.
sampling distribu-
ed on the same value,
ibutions for b_{TZ} are
same value, −5. In
3) produces *unbiased*
d β_{TZ}, regardless of
Ve uncover unbiased
ere only ten percent
y variation on Z (stu-
ioneers sample).
read of the sampling
the most certainty
sample and substan-
the random sample
sample. The greater
the student sample
mass of the sample

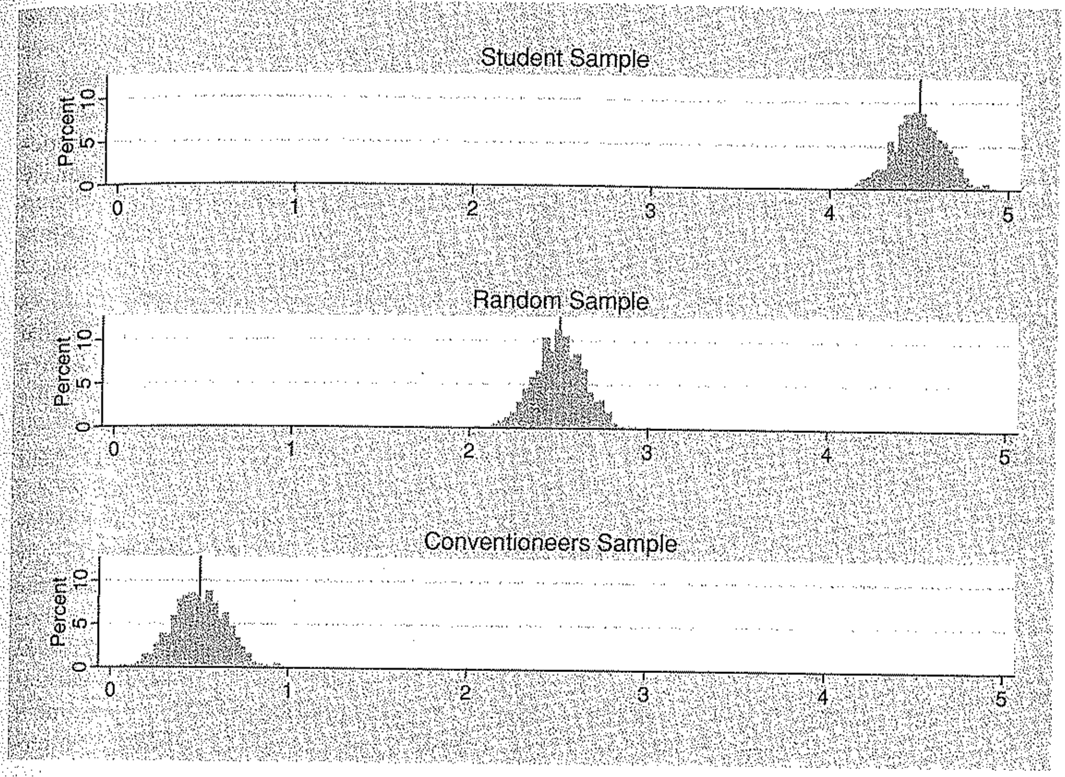

Figure 4.2. Sampling Distribution of b_T, Heterogeneous Treatment Effects
Note: 1,000 iterations, estimated using Equation (1).

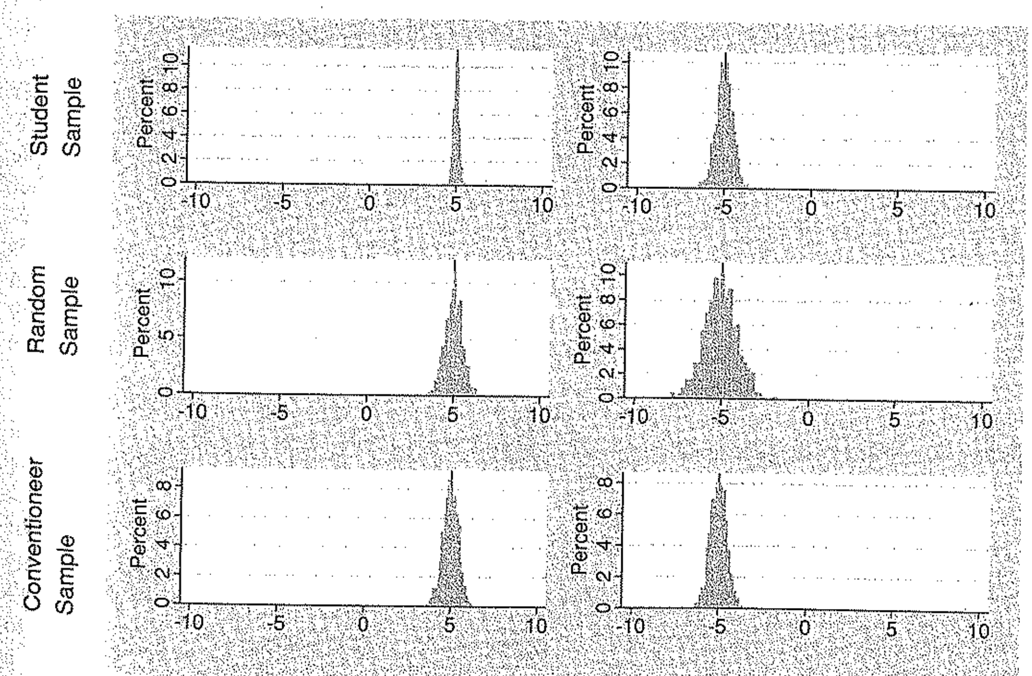

Figure 4.3. Sampling Distributions of b_T and b_{TZ}, Heterogeneous Treatment Effects
Note: 1,000 iterations, estimated using Equation (3).

that is located at 0 in the student sample (because the point estimate for β_T, the uninteracted term in Equation (3), represents the effect of T when Z takes on the value of 0).

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.[17] Moreover, we are still able to uncover the interactive treatment effect because these samples still contain some variation across values of Z.

How much variation in Z is sufficient? As long as Z varies to any degree in the sample, the estimates for b_T and b_{TZ} will be *unbiased*. Being "right on average" may be of little comfort if the degree of uncertainty around the point estimate is large. If Z does not vary very much in a given sample, then the estimated standard error for b_{TZ} will be large. But concerns about uncertainty are "run of the mill" 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 effect is theorized, then researchers should explicitly theorize how the treatment effect should vary along a specific (set of) covariate(s), and researchers can thereby estimate such relationships as 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 generalizability of student samples to offer more constructive, more theoretically oriented critiques that reflect the possibility that student samples may be problematic if the magnitude and direction of the treatment effect depend on a particular (set of) covariate(s) that are peculiarly distributed within a student sample.

In sum, we have identified three distinct situations. First, in the homogeneous case – where the data-generating process produces a single treatment effect – we showed that 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 (indeed, any convenience sample) may misestimate the average treatment effect if the sample is nonrepresentative on the particular covariate Z. However, in this case, even a representative sample would misspecify the treatment 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 nonrepresentative on the covariate Z, will misestimate 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. An implication is that the burden, to some extent, falls on an experiment's critic to identify the moderating factor and demonstrate that it lacks variance in an experiment's sample.

3. Contrasting Student Samples with Other Samples

We argue that a given sample constitutes only one – and arguably not the critical one – of many considerations when it comes to assessing external validity. Furthermore, a student sample only creates a problem when a researcher 1) fails to model a contingent

17 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 on the underlying distribution of Z in a given sample.

causal effect (when there is an underlying heterogeneous treatment effect), and 2) the students differ from the target population with regard to the distribution of the moderating variable. This situation, which we acknowledge does occur with nontrivial 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 will occur.

Kam (2005) offers some telling evidence comparing student and nonstudent samples on two variables that can affect information processing: political awareness and need for cognition. She collected data from a student sample using the same items that are used in the American National Election Study's (ANES's) representative sample of adult citizens. She found that the distributions for both variables in the student sample closely resemble those in the 2000 ANES. This near-identical match in distribution allowed Kam to generalize more broadly results from an experiment on party cues that she ran with the student subjects.

Kam (2005) focuses on awareness and need for cognition because these variables plausibly moderate the impact of party cues; as explained, in comparing student and nonstudent samples, one should focus on possible differences that *are relevant to the study in question*. Of course, one may nonetheless wonder whether students differ in other ways that *could* matter (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).

These data consist of telephone and web interviews with 2,232 individuals aged 15 years and older living in the continental United States. We limited the analysis to individuals aged 18 years and older. We selected all ostensibly politically relevant predispositions available in the data[18] and then compared individuals currently enrolled in college against the general population. The Web Appendix[19] contains question wording for each item.

As we can see from Table 4.1, in most cases, the difference in means for students and the nonstudent general population are indistinguishable from zero. Students and the nonstudent general population are, on average, indistinguishable when it comes to partisanship (we find this for partisan direction and intensity), ideology, importance of religion, belief in limited government, views about homosexuality as a way of life, contributions of immigrants to society, social trust, degree of following and discussing politics, and overall media use. Students are distinguishable from nonstudents in religious attendance, level of political information as measured in this dataset,[20] and specific types of media use. Overall, however, we are impressed by just how similar students are to the nonstudent general population on key covariates often of interest to political scientists.

In cases where samples differ on variables that are theorized to influence the size and direction of the treatment effect, the researcher should, as we note previously, model the interaction. The researcher might also 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 (Friedman and Sunder 1994, 39–40).[21] The

18 We did this before looking at whether there were differences between students and the nonstudent general population sample; that is, we did not selectively choose variables.

19 Available at http://faculty.wcas.northwestern.edu/~jnd260/publications.html.

20 The measure of political information in this dataset is quite different from that typically found in the ANES; it is heavier on institutional items and relies more on recall than recognition.

21 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).

Table 4.1: Comparison of Students versus Nonstudent General Population

	Students	*Nonstudent General Population*	p *Value*
Partisanship	0.47 (0.02)	0.45 (0.01)	*ns*
Ideology	0.50 (0.01)	0.52 (0.01)	*ns*
Religious attendance	0.56 (0.02)	0.50 (0.01)	<.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 (0–6 correct)	2.53 (0.11)	1.84 (0.07)	<.01
Newspaper use (0–7 days)	2.73 (0.14)	2.79 (0.11)	*ns*
National television news (0–7 days)	3.28 (0.15)	3.63 (0.10)	<.05
News radio (0–7 days)	2.47 (0.16)	2.68 (0.11)	*ns*
Web news (0–7 days)	3.13 (0.16)	2.18 (0.10)	<.01
Overall media use	2.90 (0.09)	2.83 (0.06)	*ns*

Notes: ns, not significant. Weighted analysis. Means with standard errors in parentheses. See the Web Appendix for variable coding and question text.
Source: Lopez, Mark Hugo, Peter Levine, Deborah Both, Abby Kiesa, Emily Kirby, and Karlo Marcelo. 2006. "The 2006 Civic and Political Health of the Nation Survey: A Detailed Look at How Youth Participate in Politics and Communities." October. Retrieved from www.civicyouth.org/PopUps/2006_CPHS_Report_update.pdf (October 30, 2010).

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 on a student sample subject pool can sometimes make it *more difficult* 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). 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 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.[22]

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.[23] 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 causal relationship discovered in the sample to differ systematically from what would be found in the general population. Researchers then need to elaborate on 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.

4. Conclusion

As mentioned, political scientists are guilty of a "near obsession" with external validity (McDermott 2002, 334) and this obsession focuses nearly entirely on a single dimension of external validity – who is studied. Our goal in this chapter 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:

- 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 because these are typically 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 the following:
  - If the heterogeneous effect is theorized, then the sample only matters if there is virtually no variance on the moderator. If there is even scant variance, then the treatment effect will not only be correctly estimated but may also be estimated with greater confidence. The suitability of a given sample can be

22 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 (e.g., Druckman 2001).

23 Convenience samples might be chosen to represent groups that are high and low on a particular covariate of interest. This purposive sampling might yield more rewards than using a less informative random sample.

assessed (e.g., empirical variance can be analyzed).

- The range of heterogeneous, nontheorized cases may be much smaller than is often believed. Indeed, when it comes to a host of politically relevant variables, student samples do not significantly differ from nonstudent samples.
- There are cases where student samples are desirable because they facilitate causal tests or make for more challenging assessments.

Our argument has a number of practical implications. First, we urge researchers to attend more to the potential moderating effects of the other dimensions of generalizability: context, time, and conceptualization. The past 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). Sniderman, Brody, and Tetlock (1991) elaborate that "the conventional survey interview, though well equipped to assess variations among individuals, is poorly equipped to assess variation across situations" (265). Unlike most controlled lab settings, researchers using survey experiments have limited ability to introduce contextual variations.

Second, we encourage the use of dual samples of students and nonstudents. 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 compare samples (e.g., Gordon et al. 1986; James and Sonner 2001; Peterson 2001; Mintz, Redd, and Vedlitz 2006; Depositario et al. 2009; Henrich, Heine, and Norenzayan 2009), although sometimes reporting differences, rarely explore the nature of the differences.[24] 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, which may be more amenable to using nonstudent subjects. Although we recognize the benefits of using survey and/or field experiments, we should not be overly sanguine about their advantages. For example, the control available in laboratory experiments enables researchers to maximize experimental realism (e.g., by using induced value or simply by more closely monitoring the subjects). Similarly, there is less concern in laboratory settings about compliance × treatment interactions that become problematic in field experiments or spillover effects in survey experiments (Transue, Lee, and Aldrich 2009; also see Sinclair's chapter in this volume). The increased control offered by the laboratory setting often affords greater ability to manipulate context and time, which, we argue, deserves 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 nonresponse biases in surveys (see Groves and Peytcheva 2008) and the impact of using "professional" survey respondents that are common in many web-based panels.[25] In short, the nature of any particular sample needs to be assessed in

24 For example, Mintz, Redd, and Vedlitz (2006) implemented an experiment, with both students and military officers, about counterterrorism decision making. They found that 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. Mintz et al. conclude that "student samples are often inappropriate, as empirically they can lead to divergence in subject population results" (769). We would argue that this conclusion is premature. Although their results reveal differences, on average, between the samples, the authors leave unanswered why the differences exist. Mintz et al. speculate that the differences may stem from variations in expertise, age, accountability, and gender (769). A thorough understanding of the heterogeneity in the treatment effects (which, as explained, is the goal of any experiment) would, thus, require exploration of these moderators. 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 that would be found in the military sample.

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

light of various trade-offs, including consideration of an experiment's goal, costs of different approaches, and other dimensions of generalizability.

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 stimulate 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." Paper presented at the annual meeting of the International Society for Political Psychology, Paris.

Aronson, Elliot, Marilynn B. Brewer, and J. Merill Carlsmith. 1985. "Experimentation in Social Psychology." In *Handbook of Social Psychology*. 3rd ed., eds. Gardner Lindzey and Elliot Aronson. New York: Random House, 441–86.

Aronson, Elliot, and J. Merill Carlsmith. 1968. "Experimentation in Social Psychology." In *Handbook of Social Psychology*. 2nd ed., eds. Gardner Lindzey and Elliot Aronson. Reading, MA: Addison-Wesley, 1–79.

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

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

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: 1173–82.

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

Berkowitz, Leonard, and Edward Donnerstein. 1982. "External Validity Is More Than Skin Deep: Some Answers to Criticisms of Laboratory Experiments." *American Psychologist* 37: 245–57.

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

Campbell, Donald T. 1969. "Prospective: Artifact and Control." In *Artifact in Behavioral Research*, eds. Robert Rosenthal and Robert Rosnow. New York: Academic Press, 351–82.

Chong, Dennis, and James N. Druckman. 2007a. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101: 637–55.

Chong, Dennis, and James N. Druckman. 2007b. "Framing Theory." *Annual Review of Political Science* 10: 103–26.

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

Depositario, Dinah Pura T., Rodolfo M. Nayga Jr., Ximing Wu, and Tiffany P. Laude. 2009. "Should Students Be Used as Subjects in Experimental Auctions?" *Economic Letters* 102: 122–24.

Dickhaut, John W., J. Leslie Livingstone, and David J. H. Watson. 1972. "On the Use of Surrogates in Behavioral Experimentation." *The Accounting Review* 47 (Suppl): 455–71.

Druckman, James N. 2001. "On the Limits of Framing Effects." *Journal of Politics* 63: 1041–66.

Druckman, James N. 2009. "Competing Frames in a Campaign." Unpublished paper, 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: 627–36.

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

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

Friedman, Milton. 1953. *Essays in Positive Economics.* Chicago: The University of Chicago 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 S., 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 *The Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press, 357–81.

Gordon, Michael E., 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: 167–89.

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

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2009. "The Weirdest People in the World: How Representative Are Experimental Findings from American University Students? What Do We Really Know about Human Psychology?" Unpublished paper, University of British Columbia.

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: 163–82.

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

Kam, Cindy D., and Robert J. Franzese, Jr. 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: 415–40.

Kinder, Donald R. 1998. "Communication and Opinion." *Annual Review of Political Science* 1: 167–97.

Kruglanski, Arie W. 1975. "The Human Subject in the Psychology Experiment: Fact and Artifact." In *Advances in Experimental Social Psychology*, ed. Leonard Berkowitz. New York: Academic Press, 101–47.

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

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

MacDonald, Paul. 2003. "Useful Fiction or Miracle Maker: The Competing Epistemological Foundations of Rational Choice Theory." *American Political Science Review* 97: 551–65.

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

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology* 67: 371–78.

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: 757–76.

Mook, Douglas G. 1983. "In Defense of External Invalidity." *American Psychologist* 38: 379–87.

Morton, Rebecca B., and Kenneth C. Williams. 2008. "Experimentation in Political Science." In *The Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press, 339–56.

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: 450–61.

Plott, Charles R. 1991. "Will Economics Become an Experimental Science?" *Southern Economic Journal* 57: 901–19.

Roth, Alvin E. 1995. "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*, eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press, 3–109.

Sears, David O. 1986. "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: 515–30.

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

Simon, Herbert A. 1963. "Problems of Methodology Discussion." *American Economic Review Proceedings* 53: 229–31.

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

Slothuus, Rune. 2009. "The Political Logic of Party Cues in Opinion Formation." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago.

Sniderman, Paul M., Richard A. Brody, and Philip E. Tetlock. 1991. *Reasoning and Choice: Explorations in Political Psychology*. Cambridge: Cambridge University Press.

Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In *Studies in Public Opinion*, eds. Willem E. Saris and Paul M. Sniderman. Princeton, NJ: Princeton University Press, 133–65.

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.

Transue, John E., Daniel J. Lee, and John H. Aldrich. 2009. "Treatment Spillover Effects across Survey Experiments." *Political Analysis* 17: 143–61.

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

Zimbardo, Phillip. 1973. "A Pirandellian Prison." *New York Times Magazine*, April 8.