Cheating on Political Knowledge Questions in Online Surveys:

An Assessment of the Problem and Solutions*

Scott Clifford
University of Houston

&

Jennifer Jerit
Stony Brook University

Accepted for publication in *Public Opinion Quarterly*

Abstract

Survey researchers worry about the quality of data produced by online surveys. One concern is that respondents might cheat on performance questions, such as political knowledge, invalidating their responses. Yet, existing evidence is unclear about the prevalence of cheating, and scholars lack a validated method for coping with the problem. In this paper, we demonstrate that such cheating behavior varies considerably by sample and provide some evidence that it is motivated by self-deceptive enhancement. We experimentally test a variety of methods for reducing cheating and find that common methods, such as timers, are not the most effective approach. By contrast, a commitment mechanism, in which respondents affirm their choice not to cheat, is more efficacious. Although cheating in online surveys can distort estimates of knowledge and decrease the validity of the measure, there are methods for coping with this problem.

* The authors thank the following people for helpful comments and suggestions on earlier versions of this paper: Craig Burnett, John Bullock, Mona Kleinberg, Yanna Krupnikov, Thomas Leeper, Adam Seth Levine, and Hillary Shulman.

Across the social sciences, survey research is increasingly being conducted online because this mode is cost-effective and convenient. Many prestigious organizations are adding online components (e.g., the American National Election Studies) and some companies operate completely online (e.g. YouGov/Polimetrix, Knowledge Networks/GfK). Yet important issues regarding data quality arise when data are collected remotely. Measures of political knowledge obtained from online surveys may be compromised if respondents use an external source to answer these items (e.g., look up the answers on the internet). This is an important concern because political knowledge is a central construct in political science, communications, and related fields.[1] As a crucial individual difference variable, political awareness influences attention to and reception of elite discourse (Zaller 1992), the use of heuristics (Lau and Redlawsk 2001), and other aspects of information processing such as effortful thinking, motivated reasoning, and value tradeoffs (e.g., Kam 2005; Taber and Lodge 2006, Jacoby 2006).

In this paper, we document how often respondents seek outside assistance while answering political knowledge questions online. Rates of self-reported cheating vary considerably across commonly-used populations, from single digits in MTurk samples to over 40 percent in some student samples.[2] We also examine several approaches to reducing cheating in

[1] We focus on survey-based measures of knowledge because, as Mondak observes, "there is compelling evidence that political awareness is best represented with data from survey batteries that measure factual knowledge" (2001, 224; also see Delli Carpini and Keeter 1996).

[2] We use the word "cheating" to be consistent with past research (e.g., Burnett 2015; Shulman et al. 2014; Jensen and Thomsen 2014), and employ the term to describe a specific kind of respondent behavior—namely, consulting an external source when answering web-based

experiments across a variety of samples, and find that there is substantial variation in the effectiveness of different approaches. A simple request not to cheat has only modest effects, while the use of commitment language (Clifford and Jerit 2015; Krosnick 1991) is more useful. Additionally, the commitment mechanism leads to the largest reduction in knowledge scores among respondents who are the most likely to cheat. More generally, our analyses indicate that the manipulations are most effective among student populations, who are most inclined to cheat, and least effective among MTurk respondents, who are least inclined to cheat. Overall, cheating can distort estimates of political knowledge and decrease the validity of the measure, but there are methods for decreasing the incidence of this behavior.

The Debate over Cheating in Online Questionnaires

There is a lively debate about how often respondents seek outside assistance on knowledge questions in online surveys. Several scholars question the frequency of cheating (e.g., noting the tendency of respondents to satisfice), and some evidence seems to support this belief (Gooch and Vavreck 2015). In one study, for example, difficult questions were not answered correctly at rates better than chance, which suggests that people are not looking up answers (Berinsky, Huber, and Lenz 2012, 359; also see Munzert and Selb n.d.).[3] Other researchers

knowledge questions. Our discussion focuses on students, workers on Amazon's Mechanical Turk (MTurk) platform, and panelists in online surveys such as YouGov because these are commonly-used subject populations in social science research (Krupnikov and Levine 2014).

[3] Gooch and Vavreck (2015) report low rates of cheating, but their study took place in a lab-like setting, making it different than the typical online survey in which respondents participate at a time and place of their own choosing.

acknowledge that political knowledge scores are higher in online samples, but they maintain that the difference is due to sample composition rather than cheating (Ansolabehere and Schaffner 2014). Collectively, these studies imply that cheating on web-based political knowledge questions is uncommon.

However, a different conclusion emerges from a mode experiment in which student subjects were randomized to take a questionnaire online or in a lab (Clifford and Jerit 2014). In that study, online participants had significantly higher political knowledge scores than those completing the survey in a lab. Due to the random assignment of survey mode, the Clifford and Jerit (2014) study provides some of the clearest evidence of cheating in online surveys. Additionally, the analysis revealed that cheating harms the descriptive validity of political knowledge scales (Luskin and Bullock 2011), as well as the criterion validity of the measure. Both findings support our assumption that cheating can impair the quality of political knowledge scales. Other studies using student (Burnett 2015; Shulman and Boster 2014) and adult subjects (Strabac and Aalberg 2010; Friker et al. 2005) also find higher political knowledge scores online compared to telephone or paper-and-pencil surveys. Cheating appears to be prevalent in nationally representative samples as well. When respondents are directly asked whether they looked up answers, cheating rates among national samples range from 13 percent (Shulman et al. 2014) to 22 percent (Jensen and Thomsen 2014). Taken together, the evidence regarding cheating suggests enough of a problem to warrant further examination of its prevalence and ways to reduce this behavior.

The Motivations for Cheating in an Anonymous Online Survey

We contribute to this literature by developing an account of why a person might cheat on knowledge questions in the first place. It is often argued that online surveys produce low levels

of social desirability bias (e.g., Holbrook and Krosnick 2010), resulting in the conclusion that respondents have little incentive to consult external sources on web-based knowledge questions. However, there is an important distinction between two forms of socially desirable responding: impression management (IM), which is a "*deliberate* tendency to over report desirable behaviors and under report undesirable ones" and self-deceptive enhancement (SDE), which refers to the "tendency to give honestly believed but overly positive reports about oneself" (Booth-Kewley et al. 2007, 464, emphasis in original). Notably, SDE occurs on attributes involving social and intellectual competence (Paulhus and John 1998) and leads to over-claiming of knowledge across a wide variety of domains (Paulhus et al. 2003). Thus, while IM pressures may indeed be low in an online survey (though see Krupnikov, Piston, and Bauer 2016), SDE is *higher* due to the greater opportunity to self-enhance in an online setting (Booth-Kewley et al. 2007).

But what does it mean for respondents to self-enhance in the area of political knowledge? A recent line of research demonstrates that the internet allows people to "offload" the burden of remembering facts and that offloading takes place without much (if any) conscious awareness (Sparrow, Liu, and Wegner 2011; Ward 2013). Consequently, people feel as if they "know" the information simply because they know where to find it. This research has shown, for example, that subjects who were instructed to access the internet while answering trivia questions took credit for their higher scores, rating themselves as more knowledgeable and having higher cognitive self-esteem than subjects who were not instructed to access the internet (Ward 2013; Fisher, Goddu, and Keil 2015). The implication is that as the internet becomes the primary way people acquire information about the world, they fail to differentiate data that is personally known from that obtained online.

These various lines of research imply that online surveys are ripe for distortions in political knowledge. The internet provides the opportunity for people to deceive themselves into believing they are knowledgeable, while self-deceptive enhancement provides the motive (Shulman and Boster 2014). Yet the effects of SDE might differ by subject population due to variation in intrinsic interests and material incentives. In particular, research has shown that the drive for self-enhancement is strongest when an individual values the topic (Brown 2012). If student subjects, who are typically recruited from social science courses, place greater value on being politically knowledgeable than other respondents, they may have greater desire to self-enhance on knowledge questions.[4] The opportunity cost of looking up answers may also affect the incidence of cheating. Compared to students and online panelists, MTurk respondents face the greatest opportunity cost from cheating due to the unique structure of that platform (Chandler, Mueller, and Paolacci 2014). Time spent looking up answers to knowledge questions represents foregone earnings from completing other tasks. In contrast, students and online panelists typically participate at their leisure, are limited in the number of surveys they can take, and do not forfeit potential earnings if they spend a moment looking up answers to questions.

The preceding discussion suggests that cheating rates vary considerably by sample, just as response styles such as survey satisficing (Hauser and Schwarz n.d.) and social desirability (Krupnikov, Piston, and Bauer 2015) vary across populations. This variation may explain why cheating on web-based knowledge questions appears common in some studies (e.g., Clifford and Jerit 2014), but less common in others (Ansolabehere and Schaffner 2015; Gooch and Vavreck

[4] Indeed, in the data we report below, student subjects had a tendency to place more importance on being politically knowledgably than campus staff ($p < .10$).

2015; Shulman et al. 2014). Looking up answers should be the most prevalent among student subjects because they face low opportunity costs and a stronger motivation to self-enhance on studies about politics.[5] In contrast, cheating should be rare among MTurk respondents because they face high opportunity costs for doing so. Adult subjects and members of online panels fall in between students and MTurk respondents on both dimensions, and thus should also fall in between the two in terms of in their tendency to consult external sources on web-based political knowledge questions.

Given the nascent status of the literature, another purpose of our study is to examine the effectiveness of various instruction sets in reducing cheating behaviors. The treatments are described in the next section, but in brief, we used instruction sets that were suggested by the existing literature or were being employed by major survey organizations. Remarkably, certain practices, such as the inclusion of timers on political knowledge questions, have never been systematically evaluated despite their use by many researchers and polling firms.[6] Based on the preceding discussion, we expect that the treatments will be most efficacious among respondents who are inclined to cheat. Consequently, the treatments should vary in effectiveness both within and across samples: they should be most effective among people who are high in SDE and students subjects, and least effective within MTurk samples.

[5] Additionally, student subjects may be likely to offload the burden of remembering facts because many of them came of age in the internet era (Kleinberg and Lau 2014).

[6] Shulman and Boster (2014, 187) and Jensen and Thomsen (2014, 3353) both encourage future researchers to examine how to reduce cheating on knowledge questions in online surveys.

Data and Measures

To test our expectations, we conducted ten studies across populations commonly used in social science research. Table 1 summarizes the key features of each study. Four samples were drawn from undergraduate subject pools at two universities in different regions of the country, with sample sizes of 84, 845, 271, and 66. Student 1 and Student 4 were conducted at Stony Brook University, while Student 2 and Student 3 were run at the University of Houston.

<Table 1 here >

Additionally, we collected data from several adult samples, including a study of campus staff (Kam, Wilking, and Zechmeister 2007) that was administered in parallel with Student 4 at Stony Brook University (N = 59), an original survey conducted by YouGov (N = 1,000), and a team module on the 2014 Cooperative Congressional Election Study (CCES; N = 1,000). The CCES data piggybacked on an unrelated study, and there we examine cheating on a five-item Wordsum battery, which measures vocabulary rather than political knowledge. The Wordsum measure is often used as a proxy for intelligence (e.g., Gooch and Vavreck 2015) and should create the same self-enhancement dynamics as political knowledge (Burnett 2015; Shulman et al. 2014). Finally, we have data from three separate MTurk samples that were part of unrelated studies (MTurk 1–3; N = 500, 505, 300). Several of the studies included experimental manipulations designed to reduce cheating which we discuss in detail below.

In each study political knowledge was measured with batteries of 5-12 questions, all but one of which were in multiple choice format.[7] Details on question wording are provided in

[7] With the exception of MTurk 2 (N = 505), there was no explicit "Don't Know" (DK) option (see Miller and Orr [2008] for discussion).

Appendices B and C, but all of the scales included questions about institutions, politicians, and policy-specific facts (Barabas, Jerit, Pollock, and Rainey 2015), and they are representative of knowledge scales used by survey researchers.[8] Immediately following the knowledge items, respondents were asked whether they had looked up any answers. The question, shown below, was designed to be as forgiving as possible to decrease misreporting (e.g., Peter and Valkenburg 2011):

> Many people struggle to remember facts, even when they know them, and so they get help remembering. When you were answering the factual knowledge questions, did you get help from any other source, such as the internet or another person? (Please be honest, this is for statistical purposes only.)

Naturally, our estimate of cheating may be conservative (i.e., a *low* estimate) if some subjects under report cheating behavior. However, we show that multiple pieces of evidence, including knowledge scores and time spent on the questions, converge with this measure.

[8] The scales vary in terms of length and apparent difficulty, as judged by percent correct on the scale (which ranges from 44% to 77% correct). There also were differences in the topics of the surveys in which the knowledge scales were embedded. Any of these factors might influence the motivation to cheat, but as we report below, we observe the expected patterns despite this variation. Moreover, data from Study 4 suggests that scale differences do not seem to affect rates of self-reported cheating. In that study, students were randomized to receive one of two knowledge batteries. One knowledge scale was substantially more difficult, as evidenced by lower scores ($p < .01$, $d = .34$). Yet respondents receiving the more difficult scale were no more likely to report cheating ($p = .56$).

Prevalence of Cheating across of Samples

Figure 1 displays the rates of self-reported cheating by sample, excluding data from experimental treatment conditions.[9] Among the student samples, rates of self-reported cheating are relatively high, ranging from 24% to 41%. In contrast, self-reported cheating is much lower in the MTurk samples, ranging from 4% to 7%. The disparity between students and MTurk subjects is consistent with the notion that the financial structure of MTurk provides an incentive not to look up answers.[10] As expected, the reported cheating rates in our adult samples are in between the rates for students and MTurk subjects. In both the campus staff and the YouGov samples, 14% of respondents reported cheating. Similarly, 13% reported cheating on the five-item Wordsum battery in the 2014 Cooperative Congressional Election Study (CCES).

<Figure 1 here>

Other metrics (time spent on the question, knowledge score; shown in Table A1 of Appendix F) support our claim about variable motives and opportunities for cheating across samples. Self-reported cheaters spent significantly longer answering the knowledge questions, with standardized effect sizes (Cohen's *d*) ranging from medium to large, $d = .63$ to 1.27. This is the expected pattern if respondents are seeking outside assistance on knowledge questions (cf. Jensen and Thomsen 2014). At the same time, there was considerable variation in the median time per question across sample types, with students taking more time than YouGov respondents

[9] Sample sizes may be smaller than those reported in Table 1 due to item non-response on the cheating question.

[10] In later analyses we examine whether MTurk subjects are more likely to under report cheating for fear of not being compensated.

and MTurk workers. This difference is consistent with the claim that student subjects face both lower opportunity costs and a stronger motivation to cheat on political knowledge questions than other types of respondents. Notably, even among students, the average time per question was low, with the median student who did not report cheating averaging 13 seconds per question and the median student who did report cheating averaging 27 seconds. Thus, our data cast doubt on the efficacy of timers for reducing cheating since the time limits are often longer than 30 seconds (see also Jensen and Thomsen 2014).

When it comes to knowledge scores among respondents who self-report cheating, there is a tendency for cheaters to have higher knowledge scores, but this difference is statistically significant only in four larger studies (see Table A1 in Appendix F for details).[11] Among respondents in the YouGov sample, people who reported cheating scored significantly *lower* than those who did not report cheating ($p < .05$). This suggests that the relationship cheating and knowledge scores may be complex. Cheating seems most likely to take place among respondents who do not know (or who are not certain of) the correct answer. The key issue is whether these respondents overcome their knowledge deficit by consulting the internet. Data across our ten studies indicates that while cheaters frequently score higher than non-cheaters, sometimes they do not.

The Effect of Interventions Designed to Reduce Cheating

In this section, we examine whether changes in survey design reduce cheating, beginning with a series of experiments conducted on student samples. We focus on students because this is the subject population that is most likely to cheat. Upon identifying the most effective

[11] Those studies include students, online panelists from the CCES, and MTurk workers.

treatments, we examine the effects of these manipulations among a large national sample and an MTurk sample.

In our first experiment (Student 1), all subjects received a standard set of instructions to the knowledge section, but half were randomly assigned to also receive a statement asking them not to look up answers during the survey ("Direct Request"), an approach used in the 2012 ANES Time Series (see also Berinsky, Huber, and Lenz 2012).[12] In the control condition, 23.9% reported cheating, while 18.4% reported cheating in the treatment condition, a difference that is not statistically significant ($p = .54$). Knowledge in the treatment group is lower than in the control (3.7 vs. 3.9, respectively), but the difference is not statistically significant or substantively large ($p = .44$, $d = .17$).

In our second student study (Student 2) we explored three different manipulations: Timer, Commitment, and Forgiving Request. All respondents (including those in the control group) received the standard introduction (see note 12), while those in the treatment conditions also saw additional text described below. The Timer treatment placed a 30-second timer on each question and explained to subjects that the screen would automatically advance after 30 seconds (Ansolabehere and Schaffner 2014; Prior, Sood, and Khanna 2015; Bullock et al. 2015).

> *Timer:* "Please do NOT use outside sources like the Internet to search for the correct answer. You will have 30 seconds to answer each question."

[12] The language was as follows: "Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers, and the like. [*Please give your best guess and do NOT use outside sources like the Internet to search for the correct answer.*] "

The second treatment asked respondents whether they were willing to answer the knowledge questions without getting help and requested a yes or no response ("Commitment"; see Clifford and Jerit 2015 or Krosnick 1991). We expect this treatment to be effective because few people will reject a request from an interviewer, and subsequently they will be motivated to maintain consistency with their previous commitment (Cialdini et al. 1978; Cannell, Miller, and Oskenbeurg 1981).

> *Commitment:* "It is important to us that you do NOT use outside sources like the Internet to search for the correct answer. Will you answer the following questions without help from outside sources?" (Yes, No) [13]

The third and final treatment was a stronger version of the Direct Request from the Student 1 study. Like the Direct Request, the treatment instructed subjects not to look up answers, but it also included additional forgiving language intended to reduce social desirability pressures ("Forgiving Request"; e.g., Duff et al. 2007).

> *Forgiving Request:* "Many people will know the answers but won't be able to remember them at the moment. However, it is essential that we only measure what people are able to recall on their own. Thus, it is important to us that you do NOT use outside sources like the Internet to search for the correct answer."

The comparisons from the Student 2 sample are shown in Figure 2. On average, subjects in the control condition answered 5.1 questions out of 8 correctly, with 41% reporting cheating. Beginning with the Timer condition, subjects were significantly less likely to report cheating (16.4%, $p < .001$) and scored significantly lower on knowledge (4.3, $p < .001$, $d = .54$). One

[13] Only 5 out of 222 subjects (2.3%) answered "no" to the question posed by the treatment. All subjects are retained for analysis regardless of their response.

would expect this treatment (by design) to reduce time spent on the knowledge questions, and the results bear out this expectation ($p < .001$, $d = .71$).[14]

<Figure 2 here>

Turning to the Commitment condition, self-reported cheating was dramatically lower than in the control condition (9.5%, $p < .001$). In addition, knowledge scores (4.4; $p < .001$, $d = .49$) and time spent on the knowledge questions were significantly lower as well ($p = .015$, $d = .24$). Finally, we examine the Forgiving treatment, which was designed to be a stronger version of the Direct Request. Relative to the rate of cheating in the control condition (41.4%), the Forgiving treatment significantly decreased rates of self-reported cheating (22.4%, $p < .001$), knowledge scores (4.5; $p < .001$, $d = .40$), and time spent on the knowledge questions ($p = .004$, $d = .29$). However, this manipulation was less effective than both the Commitment and Timer conditions.

The three treatments from the Student 2 study significantly reduced cheating, knowledge scores, and time spent on the knowledge questions. The Timer and Commitment treatments led to the lowest levels of political knowledge, while the Commitment condition produced significantly lower levels of self-reported cheating ($p < .05$). This combination of results was unexpected, but it could have occurred if the Timer condition was less effective at reducing cheating *and* it interfered with non-cheaters' ability to answer the questions. Recall that even among those who reported cheating, the median subject spent less than 30 seconds answering each question in all of our samples and conditions. Thus, timers face the challenge of being short enough to prevent cheating, but not so short that they interfere with the regular response process.

[14] We use the log of the average time spent per question for all tests of reaction times.

Overall, the Commitment language appears to be the most effective method for minimizing cheating on political knowledge questions.

In our next study, Student 3, we sought to replicate our findings regarding the Commitment mechanism and examine whether the treatment has the strongest effects among respondents who are the most likely to cheat (i.e., those high in SDE). In this study, 271 students were randomly assigned to the Commitment or control condition. The study also included the Balanced Inventory of Desirable Responding (BIDR-40; Paulhus 1991), which measures both self-deceptive enhancement (SDE) and impression management (IM). In the BIDR-40, subjects are asked to rate how true 40 statements are about themselves (e.g., "I am a completely rational person"; see Appendix D for full scale). Each subscale (IM and SDE) is scored as the average agreement with each of the 20 corresponding items (Stöber, Dette, and Musch 2002).

To test whether the treatment had different effects across levels of SDE, we estimated a model in which knowledge was regressed on treatment status, SDE, IM, and interactions between each disposition and the treatment condition (full results shown in the Appendix F; Table A2). As expected, we find a negative interaction between SDE and the Commitment condition ($p = .07$), suggesting that the Commitment treatment was more effective among those high in SDE. This finding is consistent with our argument that these respondents are the most motivated to cheat.

Deterring Cheating in Adult Samples

While our manipulations proved effective at reducing cheating in student samples, adult respondents may respond differently to the treatments. We investigate this possibility by testing the two most effective treatments from the student studies, Timer and Commitment, in the adult

sample collected through YouGov.[15] Recall from Figure 1 that adult subjects were less likely than students to report cheating. This pattern is consistent with the claim that online panelists have weaker motives to cheat than student subjects. Yet the weaker motivation to cheat may make it more difficult to find a significant effect for language designed to reduce cheating among adult subjects. In the control condition of the YouGov study, the rate of self-reported cheating was 14%. The Timer condition reduced self-reported cheating to 8% ($p = .03$) while the Commitment condition reduced cheating to 6% ($p = .009$). When it comes to levels of knowledge the differences across conditions were more muted. On average, subjects in the control condition answered 5 out of 8 questions correctly. Knowledge scores did not significantly differ in either the Timer (4.8; $p = .31$) or Commitment conditions (5.0; $p = .67$), though both the Timer and Commitment conditions decreased time spent on the questions ($p < .001$, $p = .08$, respectively). Overall, both manipulations appeared less effective on the adult sample than the student samples, though both reduced self-reported cheating and time spent on the questions. Morever, as we show below, the Commitment condition also improved the validity of the political knowledge scale in the YouGov sample.

Although neither treatment had a main effect on political knowledge, the interventions might have an influence on the subgroup most inclined to cheat: those high in SDE. For this purpose, we included an abbreviated 16-item measure of SDE and IM (Bobbio and Manganelli

[15] We included an experiment involving the Commitment language in Student 4 and Campus Staff ($N = 66$ and $N = 59$, respectively). Despite our recruiting efforts, these studies were smaller than expected which in turn reduced statistical power. In both studies, the magnitude of treatment effects was consistent with our other studies, but the differences were not statistically significant.

2011) in the YouGov study. Similar to our analysis in the previous section, we predicted political knowledge as a function of the Timer and Commitment conditions, SDE, IM, and interactions between each condition and both SDE and IM (see Table A3 in Appendix F for full model results). We find suggestive evidence of the expected negative interaction between SDE and the Commitment condition ($p = .12$) but weaker evidence for an interaction between SDE and the Timer condition ($p = .36$) and between IM and either condition (all ps $> .76$). Overall, there are similarities across student and adult respondents although the results are weaker among the YouGov sample. This may stem from YouGov respondents being less motivated to cheat.

Deterring Cheating in MTurk Samples

We now turn to the question of whether our treatments are effective in MTurk samples. Earlier, we argued that the low rate of self-reported cheating among MTurk subjects (see Figure 1) is a function of opportunity costs that countervail the motivation to cheat (cf. Goodman, Cryder, Cheema 2013). An alternative explanation is that MTurk subjects are simply more likely to *under report* cheating for fear of having their pay rejected. If cheating is in fact low on MTurk (as we argue), the Commitment manipulation should have little effect on knowledge scores of MTurk respondents. To test this prediction, we randomly assigned subjects in our MTurk 3 sample ($N = 300$) to the Commitment condition or a control condition, followed by four knowledge questions. In line with our expectations, rates of self-reported cheating among MTurk subjects in the control condition were low, at 6.6%. This rate dropped to 1.5% in the treatment condition ($p < .05$). However, there was no significant difference in political knowledge scores across conditions ($p = .35$). This pattern bolsters the claim that cheating actually is rare among MTurk subjects. It also suggests that previous reports of low levels of cheating among MTurk

respondents who had been instructed not to cheat (Berinsky, Huber, and Lenz 2012) may be driven by the sample more than the instructions.

The Effect of Cheating on the Validity of Political Knowledge Measures

Previous research provides some evidence that cheating harms the criterion validity of knowledge measures in student and adult samples (Clifford and Jerit 2014; Jensen and Thomsen 2014).[16] Here we investigate the related question of whether cheating interventions *improve* the predictive validity of these measures. In the survey conducted by YouGov, we included several open-ended thought-listing questions. Based upon previous research, political knowledge should be positively related to the number of considerations a person is able to list (e.g., Zaller 1992; Zaller and Feldman 1992). Thus, if the Commitment language improves the validity of knowledge measures, there should be a positive and significant interaction between the treatment and political knowledge in a model predicting the number of thoughts listed. Analysis of predictive validity in the YouGov sample provides a difficult test for the Commitment mechanism, as the effects on self-reported cheating and knowledge scores were smaller than in student samples.

At the beginning of the YouGov study, respondents reported their opinion to a closed-ended question on gun control and then were asked to list the thoughts that came to mind as they answered the question. Following the thought-listing, respondents were asked what factors they

[16] We also have evidence from Student 1, which consisted of a two-wave panel in which students answered political knowledge questions in a lab and then answered different knowledge items in a follow-up survey online. The correlation between attention to politics and political knowledge is significantly lower in the online wave, relative to the lab wave (see Appendix G).

thought were responsible for the mass shootings that have occurred in the U.S. in recent years. There was a parallel set of questions on the topic of health care reform.[17] Once again, we expected that political knowledge would be associated with listing a larger number of considerations for both the thought-listing and attribution items (Gomez and Wilson 2001; Zaller 1992). A coder who was blind to our expectations and treatment assignment coded the number of distinct considerations listed by respondents in each of the four questions (Krippendorf's α = .85).[18] The four items were summed to form an index of thought-listing (Cronbach's α = .85).

We use a negative binomial regression model to predict the number of thoughts as a function of political knowledge scores, the treatment conditions, and interactions between knowledge and each condition.[19] Full model results are shown in Table A4 in Appendix F, but political knowledge is a strong predictor of holding more considerations ($p < .001$). As expected, however, there is a positive interaction between political knowledge and the Commitment condition, indicating that political knowledge is a stronger predictor of thought-listing in the Commitment condition than in the control condition ($p = .058$). In the Timer condition, the

[17] The fixed choice questions read as follows: "Do you favor or oppose stricter gun control laws?" and "Do you support or oppose the health care law passed by the President and Congress in 2010?" See Appendix D for question wording of the closed- and open-ended items.

[18] The reported Krippendorf's alpha comes from an intercoder reliability analysis based on a random subset of the data (n=60).

[19] We exclude inattentive respondents because they are less likely to give reliable responses to the SDE battery and are less responsive to experimental manipulations (Berinsky, Margolis, and Sances 2014; see Appendix E for details).

interaction between political knowledge is negative and statistically insignificant ($p = .86$). Moreover, a direct comparison between the Commitment and Timer conditions shows that the political knowledge scale has significantly higher predictive validity in the Commitment condition than in the Timer condition ($p < .05$). Overall, the results suggest that the Commitment condition, but not the Timer condition, improves the predictive validity of the political knowledge scale.

Are Cheating Interventions Off-Putting to Respondents?

Although interventions designed to reduce cheating are effective in most cases, researchers may worry that the treatments are off-putting. Timers may increase the stress on respondents while the commitment language may be interpreted as an accusation of dishonesty. We explored these concerns with two measures from the YouGov study. At the end of that survey respondents were asked to rate how interested they would be "in taking another survey like this one" on a five-point scale. We reasoned that respondents would be less interested if they found the treatments distasteful. In the control condition, the average interest in taking a similar survey in the future was 4.08, roughly corresponding with the "very interested" response option. Survey interest did not significantly differ in the Timer or Commitment conditions (ps > .50), which is at odds with the claim that these interventions upset respondents.

Additionally, we included an open-ended item asking if there was anything respondents disliked about the survey. Overall, 19% of the sample reported disliking something about the survey. Among that group, however, only 34 respondents (3% of the total sample) made a statement related to political knowledge. Several ($n = 19$) noted embarrassment over not being able to answer the questions, however, only three respondents referenced some aspect of the cheating manipulations. In the Timer condition, one person objected to not knowing when the 30

seconds was up and another complained that he could not take a break between questions (due to the automatic advance on the knowledge questions in that condition). The only dislike noted among respondents in the Commitment condition was, "The quiz part without looking anything up." Overall, respondents were more likely to dislike some feature of the knowledge questions in the control group (4.2%) compared to either the Timer (2.7%; $p = .28$) or the Commitment conditions (3.2%; $p = .47$).[20] Moreover, several people remarked that they were tempted to look up answers, or that they followed instructions in spite of their usual habit of searching for the answers. Comments such as these underscore the prevalence of cheating in online questionnaires and the importance of identifying this behavior in survey data.

Conclusion

Collectively, these studies demonstrate that respondents frequently look up the answers to political knowledge questions in online surveys and provide suggestive evidence that this behavior varies by sample. On MTurk, where respondents have a financial incentive to finish quickly, cheating rates seem relatively low. Among students, who face lower opportunity costs and may have a greater desire to self-enhance on political knowledge, cheating behavior is common. In this latter population, cheating rates are high enough to distort levels of political knowledge. Finally, in national samples, self-reported cheating rates were moderate, though we still found that a cheating intervention (i.e., the Commitment language) improved the validity of a knowledge scale.

[20] Overall, 56% of respondents explicitly stated that there was nothing they disliked about the survey, and this figure did not vary across conditions ($p = .59$).

Researchers and survey organizations have adopted practices such as time limits and instructions to discourage cheating, but have provided little evidence for their efficacy. According to our analyses, a direct request asking people not to cheat was the least effective of the techniques we examined, even when paired with forgiving language intended to reduce social desirability pressures. Timers were more effective, but they may interfere with honest subjects' ability to answer the questions. As a result, we recommend the commitment item—i.e., asking respondents whether they are willing to answer knowledge questions without help, and requiring a yes or no answer. This technique yielded the lowest levels of cheating and the highest predictive validity—all without the potential disadvantages of timers or other unexpected disadvantages regarding respondent goodwill.

We also provide evidence for the utility of simple self-reports for assessing the prevalence of cheating behavior. Though the measure is no doubt a low estimate, it generally corresponded with two other indicators of cheating behavior: knowledge scores and reaction time. As a result, we encourage researchers to employ a self-report measure when measuring political knowledge in online surveys as an *aggregate* measure of cheating behavior. Such information will help diagnose potential problems and enhance our understanding of cheating behavior. [21]

[21] At present, researchers do not have a definitive way to identify cheating. The willingness to admit to this behavior may be correlated with characteristics that are related to knowledge (e.g., engagement, interest). Although we believe a self-reported cheating question can be useful in gauging the extent of this behavior in the aggregate, more caution is warranted when using this item as a measure of cheating at the individual level.

While we have collected and analyzed a large amount of data, questions remain about the motives for cheating. We find some evidence that the motivation for this behavior varies by sample and by levels of self-deceptive enhancement, but it is unclear whether cheating varies by the type of knowledge question. Comparing the CCES and YouGov data, respondents are equally likely to cheat on political knowledge and vocabulary tests. Furthermore, among all of our political knowledge questions, item difficulty alone does not seem to drive cheating behavior (see note 8). Further research is needed to uncover how context, personality, and question-level factors interact to motivate cheating behavior.

The rise of online surveys has made data collection faster and more convenient than ever, particularly in combination with crowd-sourcing platforms and student samples. Although researchers have explored the generalizability of findings across different types of samples (Krupnikov and Levine 2014; Mullinix et al. 2015), there has been less attention to how differences in sample characteristics affect data quality, particularly when research is conducted online (Weinberg, Freese, and McElhattan 2014 is a notable exception). Our analyses suggest that financial incentives and personal motivations can have dramatic effects on data quality with regard to political knowledge. As researchers turn towards online research, and adopt new approaches to recruiting participants for research, it is increasingly important to understand how data quality is affected by the unique characteristics of different subject populations.

References

Ansolabehere Stephen, and Brian F. Schaffner. 2015. "Distractions: The Incidence and Consequences of Interruptions for Survey Respondents." *Journal of Survey Statistics and Methodology* 3: 1–24.

Ansolabehere, Stephen, and Brian Schaffner. 2014. "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison." *Political Analysis* 22(3):285–303.

Barabas, Jason, Jennifer Jerit, William Pollock, and Carlisle Rainey. 2015 "The Question(s) of Political Knowledge." *American Political Science Review* 108(4): 840–55.

Berinsky, Adam J., Michele Margolis, and Michael Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Internet Surveys." *American Journal of Political Science* 58 (3): 739–53.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Using Mechanical Turk as a Subject Recruitment Tool for Experimental Research." *Political Analysis* 20(3): 351–368.

Bobbio, Andrea, and Anna Maria Manganelli. 2011. "Measuring Social Desirability Responding: A Short Version of Paulhus' BIDR 6." *Testing, Psychometrics, Methodology* 18 (2): 117–135.

Booth-Kewley, Stephanie, Gerald E. Larson, and Dina K. Miyoshi. 2007. Social Desirability Effects on Computerized and Paper-and-Pencil Questionnaires. *Computers in Human Behavior* 23(4): 463–477.

Brown, Jonathan D. 2012. "Understanding the Better than Average Effect: Motives (Still) Matter." *Personality and Social Psychology Bulletin* 38(2): 209–219.

Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. 2015. "Partisan Bias in Factual Beliefs about the Economy." *Quarterly Journal of Political Science* 10: 519–78.

Burnett, Craig, M. 2015. "Exploring the Difference in Performance on Knowledge by Participants between Online and In-Person Modes." Working paper, University of North Carolina Wilmington.

Cannell, Charles F., Peter V. Miller, and Lois Oskenbeurg. 1981. "Research on Interviewing Techniques." Sociological Methodology 12: 389–437.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaiveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavioral Research* 46 (1): 112–130.

Cialdini, Robert B., John T. Cacioppo, Rodney Bassett, and John A. Miller. 1978. "Low Ball Procedure for Producing Compliance: Commitment then Cost." Journal of Personality and Social Psychology 36(5): 463–76.

Clifford, Scott, and Jennifer Jerit. 2014. "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies." *Journal of Experimental Political Science* 1(2): 120–131.

Clifford, Scott, and Jennifer Jerit. 2015. "Do Attempts to Improve Respondent Attention Increase Social Desirability Bias?" *Public Opinion Quarterly* 79(3): 790-802.

Delli Carpini, Michael X., and Scott Keeter. 1996. *What Americans Know about Politics and Why it Matters*. New Haven, CT: Yale University Press.

Duff, Brian, Michael J. Hanmer, Won-Ho Park, and Ismail K. White. 2007. "Good Excuses: Understanding Who Votes With An Improved Turnout Question." *Public Opinion Quarterly* 71(1): 67–90.

Fisher, Matthew, Mariel K. Goddu, and Frank C. Keil. 2015. "Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge." *Journal of Experimental Psychology: General* 144(3): 674–687.

Fricker, Scott, Mirta Galesic, Roger Tourangeau, and Ting Yan. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 69(3): 370–92.

Gomez, Brad T., and J. Matthew Wilson. 2001. Political Sophistication and Economic Voting in the American Electorate: A Theory of Heterogeneous Attribution." *American Journal of Political Science* 45 (October): 899–914.

Gooch, Andrew, and Lynn Vavreck. 2015. "How Face-to-Face Interviews and Cognitive Skill Affect Item Non-response: A Randomized Experiment Assigning Mode of Interview." Working paper, UCLA.

Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26: 213–224.

Hauser, David J., and Norbert Schwarz. N.d. "Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants." *Behavior Research Methods*. Forthcoming.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74(1): 37–67.

Jacoby, William G. 2006. "Value Choices and American Public Opinion." *American Journal of Political Science* 50(3):706–23.

Jensen, Carsten, and Jens Peter Frølund Thomsen 2014. "Self-Reported Cheating in Web Surveys on Political Knowledge." *Qual Quant* 48: 3343–3354.

Kam, Cindy. 2005. "Who Toes the Party Line? Cues, Values, and Individual Differences." *Political Behavior* 27(2): 163–82.

Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Database': Another Convenience Sample for Experimental Research." *Political Behavior* 29 (December): 415–40.

Kleinberg, Mona and Richard R. Lau. 2014. "Cognitive Misers 2.0 – Does Internet Access and the 24/7 Availability of Political Knowledge Create a Disincentive to Committing Information to Memory?" Paper presented at annual meeting of American Political Science Association, Washington, DC.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psycholo*gy 5(3): 213–36.

Krupnikov, Yanna, Spencer Piston, and Nichole Bauer. 2016. "Saving Face: Identifying Voter Responses to Black Candidates and Female Candidates." *Political Psychology* 37 (2): 253–73.

Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1: 59–80.

Lau, Richard R., and David P. Redlawsk. 2001. "Advantages and Disadvantages of Cognitive Heuristics in Political Decision Making." *American Journal of Political Science* 45(4):951–71.

Luskin, Robert C., and John G. Bullock. 2011. "'Don't Know' Means 'Don't Know': DK Responses and the Public's Level of Political Knowledge." *Journal of Politics* 73 (2): 547–57.

Miller, Melissa K., and Shannon K. Orr. 2008. "Experimenting with a 'Third Way' in Political Knowledge Estimation." *Public Opinion Quarterly* 72 (3): 768–780.

Mondak, Jeffery J. 2001. "Developing Valid Knowledge Scales." *American Journal of Political Science* 45(1):224–238.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2): 109–138.

Munzert, Simon, and Peter Selb. N.d. "Measuring Political Knowledge in Web-Based Surveys: An Experimental Validation of Visual Versus Verbal Instruments." *Social Science Computer Review*. Forthcoming.

Paulhus, Delroy L., and Oliver P. John. 1998. Egoistic and Moralistic Bias in Self-Perceptions: the Interplay of Self-Deceptive Styles with Basic Traits and Motives. *Journal of Personality* 66: 1024–1060.

Paulhus, Delroy L., P. D. Harms, M. Nadine Bruce, and Daria C. Lysy. 2003. "The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability." *Journal of Personality and Social Psychology* 84(4): 890–904.

Peter, Jochen and Patti M. Valkenburg 2011. "The Impact of 'Forgiving' Introductions on the Reporting of Sensitive Behavior in Surveys: The Role of Social Desirability Response Style and Developmental Status." *Public Opinion Quarterly* 75(4): 779–787.

Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. "You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions." *Quarterly Journal of Political Science* 10: 489–518.

Shulman, Hillary C., and Franklin J. Boster. 2014. "Effect of Test-Taking Venue and Response Format on Political Knowledge Tests." *Communication Methods and Measures* 8(3): 177–189.

Shulman, Hillary C., Franklin J. Boster, Christopher Carpenter, and Allison Shaw. 2014. "Why Do Students Completing a Political Knowledge Test Score Higher Online than in the Clasroom? A Series of Studies." Working paper, North Central College.

Sparrow, Betsy, Jenny Lui, and Daniel M. Wegner. 2011. "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips." *Science* 333: 776–78.

Stöber, Joachim, Dorothea E Dette, and Jochen Musch. 2002. "Comparing Continuous and Dichotomous Scoring of the Balanced Inventory of Desirable Responding." *Journal of Personality Assessment* 78(2): 370–89.

Strabac, Zan, and Toril Aalberg. 2011. "Measuring Political Knowledge in Telephone and Web Surveys: A Cross-National Comparison." *Social Science Computer Review* 29(2): 175–92.

Taber, Charles and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50: 755–769.

Ward, Adrian F. 2013. "Supernormal: How the Internet is Changing Our Memories and Our Minds." *Psychological Inquiry* 24: 341–348.

Weinberg, Jill D., Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample." *Sociological Science* 1: 292–310.

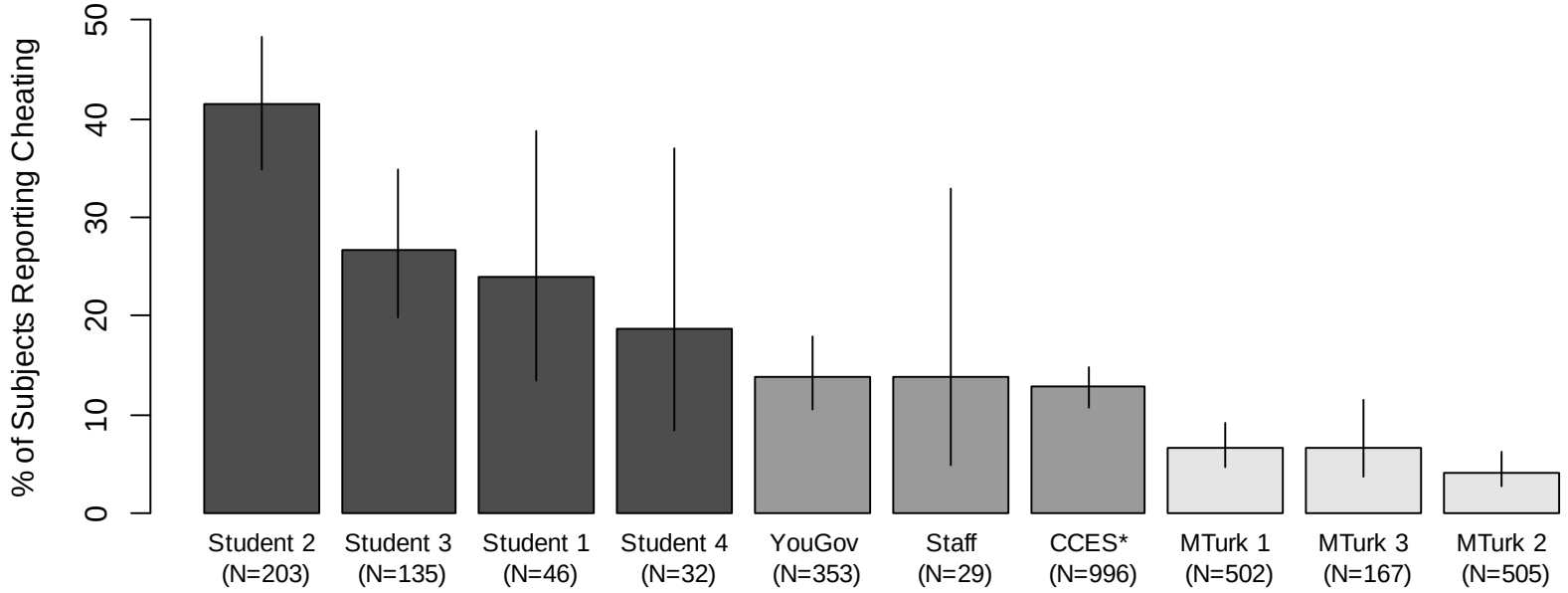
Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Table 1. Summary of Data Sources

	Sample Size (Control Group)	Approximate Date of Study	Included Experimental Manipulation(s)	Number of Knowledge Qs
Student 1	84 (46)	May, 2014	Direct Request	5
Student 2	845 (203)	November, 2014	Forgiving, Timer, Commitment	8
Student 3	271 (135)	March, 2015	Commitment	8
Student 4	66 (35)	April, 2015	Commitment	4
Campus Staff	59 (30)	May, 2015	Commitment	4
YouGov	1000 (354)	December, 2015	Timer, Commitment	8
CCES	1000 (NA)	November, 2014	None	5
MTurk 1	500 (NA)	July, 2014	None	6
MTurk 2	505 (NA)	June, 2015	None	12
MTurk 3	300 (167)	April, 2015	Commitment	4

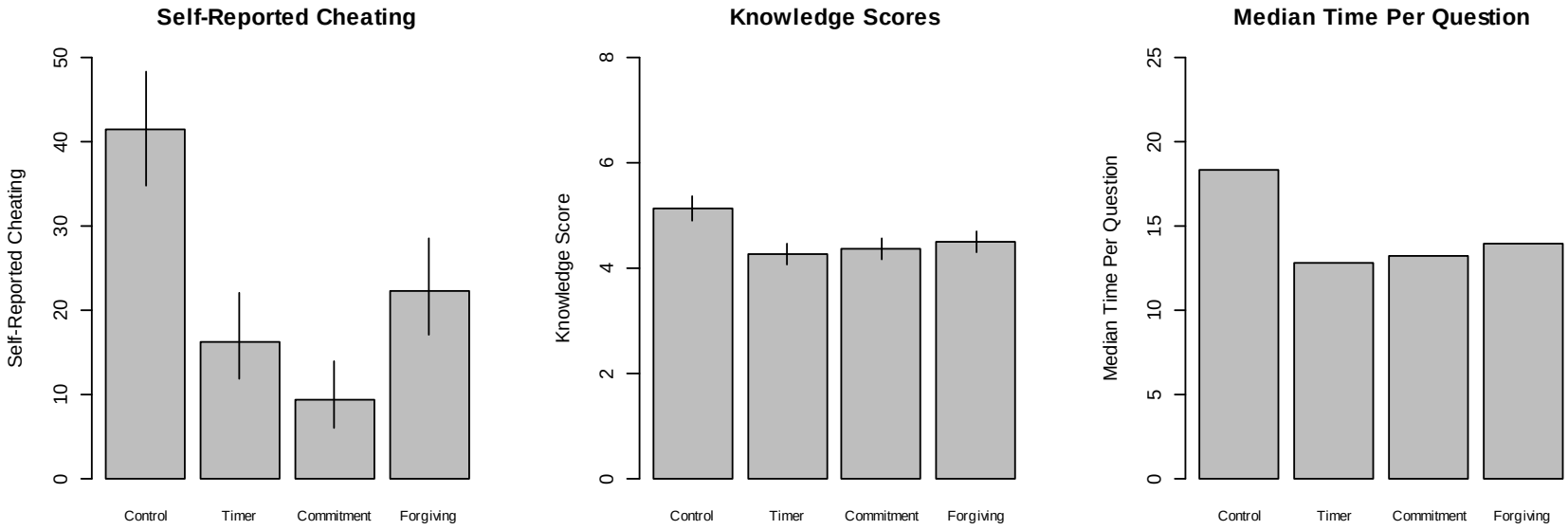
Note: The data reported in Student 1, Student 4 and Campus Staff were from the second wave of two-wave studies. See Appendix for details on the administration of all studies.

Figure 1. Rates of Self-Reported Cheating on Knowledge Questions in Online Surveys



Note: Column height represents the percentage of self-reported cheating on knowledge questions among student, staff, and MTurk samples (lines represent 95% confidence intervals). Studies arranged in descending order by self-reported cheating rates. See Table A1 for details. CCES cheating rates come from 5-item Wordsum measure. Data from control conditions only. Sample size may be smaller than reported in Table 1 due to item non-response.

Figure 2. A Test of Three Methods to Reduce Cheating



Note: Column height represents the percentage of subjects self-reporting cheating (left), knowledge scores, measured as number of questions correct out of eight items (middle), and the median response time per question (right). Data comes from Student 2 sample (N = 845).