Institute for Policy Research
Northwestern University
Working Paper Series

WP-14-19

The Generalizability of Survey Experiments*

Kevin J. Mullinix
Assistant Professor
Department of Government and Justice Studies
Appalachian State University

Thomas J. Leeper
Assistant Professor in Political Behaviour
London School of Economics and Political Science

James N. Druckman
Payson S. Wild Professor of Political Science and IPR Fellow
Northwestern University

Jeremy Freese
Professor of Sociology
Stanford University

Version: October 13, 2015

DRAFT
Please do not quote or distribute withoutpermission.

Keywords: survey experiments, sampling, causal inference

* The authors acknowledge support from a National Science Foundation grant for Time-Sharing Experiments in the Social Sciences (SES-1227179). Druckman and Freese are co-Principal Investigators of TESS, and Study 2 was designed and funded as a methodological component of their TESS grant. Study 1 includes data in part funded by an NSF Doctoral Dissertation Improvement Grant to Leeper (SES-1160156) and in part collected via a successful proposal to TESS by Mullinix and Leeper. Druckman and Freese were neither involved in Study 1 nor with any part of the review or approval of Mullinix and Leeper's TESS proposal (via recusal, given other existing collaborations). Only after data from both studies were collected did authors determine that the two studies were so complementary that it would be better to publish them together. The authors thank Lene Aarøe, Kevin Arceneaux, Christoph Arndt, Adam Berinsky, Emily Cochran Bech, Scott Clifford, Adrienne Hosek, Cindy Kam, Lasse Laustsen, Diana Mutz, Helene Helboe Pedersen, Richard Shafranek, Flori So, Rune Slothuus, Rune Stubager, Magdalena Wojcieszak, workshop participants at Southern Denmark University, and participants at The American Panel Survey Workshop at Washington University, St. Louis.

Abstract

Abstract: Survey experiments have become a central methodology across the social sciences. Researchers can combine experiments' causal power with the generalizability of population-based samples. Yet, due to the expense of population-based samples, much research relies on convenience samples (e.g., students, online opt-in samples). The emergence of affordable, but non-representative online samples has reinvigorated debates about the external validity of experiments. We conduct two studies of how experimental treatment effects obtained from convenience samples compare to effects produced by population samples. In Study 1, we compare effect estimates from four different types of convenience samples and a population-based sample. In Study 2, we analyze treatment effects obtained from 20 experiments implemented on a population-based sample and Amazon's Mechanical Turk. The results reveal considerable similarity between many treatment effects obtained from convenience and nationally representative population-based samples. While the results thus bolster confidence in the utility of convenience samples, we conclude with guidance for the use of a multitude of samples for advancing scientific knowledge.

Experiments have become increasingly common across the social sciences (Berger 2014; Druckman and Lupia 2012; Holt 2006; Kriss and Weber 2013; Morawski 1988). Of considerable appeal are survey experiments that "seek to establish causal relationships that are generalizable – that is, they try to maximize internal and external validity" (Barabas and Jerit 2010, 226). The ideal is that such studies afford clear causal inferences that generalize to a broad population.

For example, in one notable survey experiment, some respondents were randomly assigned to receive only information about the partisanship of the officials responsible for dealing with the aftermath of Hurricane Katrina (Malhotra and Kuo 2008). Others randomly received further descriptions of the officials' jobs. Those in the latter condition relied much less on partisanship in assessing blame for mishandling the response; thus, the influence of partisanship was mitigated when job responsibilities were provided. Given the data came from a representative sample of United States citizens, the researchers were able to sensibly generalize the results to this population.

Population-based survey experiments are experimental designs embedded within surveys that are "administered to a representative population sample" (Mutz 2011, 2; see also Nock and Guterbock 2010, 860). They have become an ostensible "gold standard" for generalizable causal inferences. Hundreds of population-based survey experiments have been carried out (Mutz 2011), and Sniderman (2011) refers to them as "the biggest change in survey research in a half century" (102).

A central challenge for population-based survey experiments, however, is their cost. Even a relatively brief survey on a population-based sample can cost more than $15,000. It is for this reason that many researchers continue to rely on cheaper convenience samples including those drawn from undergraduate students (Sears 1986), university staff (Kam, Wilking, and

Zechmeister 2007), social media sites (Broockman and Green 2013; Cassesse et al. 2013),[1] exit polls (Druckman 2004), and, perhaps most notably, Amazon's Mechanical Turk (MTurk). MTurk is an online crowdsourcing platform that has become widely used across the social sciences due its ease of use, low cost, and capacity to generate more heterogeneous samples than subject pools of students (see Berinsky, Huber, and Lenz 2012; Krupnikov and Levine 2014; Paolacci, Chandler, and Ipeirotis 2010). That said, MTurk is an opt-in sample, meaning that respondents self-select into participating rather than being drawn with known probability from a well-specified population, and, as such, MTurk and other convenience samples invariably differ from representative population samples in myriad, possibly unmeasured, ways.

Each of the aforementioned convenience samples is substantially cheaper than a population-based sample; however, do survey experiments using a convenience sample produce results that are similar to those conducted on a population-based sample?[2] That is, would we arrive at the same causal inference if a study were performed on a convenience sample versus on a population-based sample? A common concern is that the features of a given convenience sample may diverge from a representative population sample in ways that bias the estimated treatment effect. For instance, if the previously discussed Hurricane Katrina experiment was conducted on a convenience sample of strong partisans, the results likely would have differed. Isolating the presence of such biases is difficult since one can rarely, if ever, identify all the selection biases shaping the composition of a convenience sample.

[1] Survey research makes uses of other non-representative online platforms (Wang et al. 2015).

[2] This echoes a long-standing question about the generalizability of any convenience sample experiment, such as those conducted on "college sophomores" (Sears 1986). McDermott (2002, 334) notes that concerns about the sample are a "near obsession" (also see Gerber and Green 2008, 358; Gerring 2012, 271; Iyengar 1991, 21). It is for this reason that population-based survey experiments have been so alluring to social scientists; Mutz (2011) explains, "Critics over the years have often questioned the extent to which the usual subjects in social science experiments resemble broader, more diverse populations…. Population-based survey experiments offer a powerful means for research to respond to such critiques" (11).

Consequently, the extent to which varying types of convenience samples produce experimental treatment effects analogous to population-based surveys is an empirical question. Recent work has sought to compare samples (e.g. Berinsky, Huber, and Lenz 2012; Goodman, Cryder, and Cheema 2012; Horton, Rand, and Zeckhauser 2011; Krupnikov and Levine 2014; Paolacci, Chandler, and Ipeirotis 2010; Weinberg, Freese, and McElhattan 2014).[3] While these studies are impressive and telling, each includes only a small number of comparisons (e.g., three experiments) on a limited set of issues (e.g., three or four) and topics (e.g., question wording, framing) with few types of samples (e.g., three) at different points in time (e.g., data were collected on distinct samples far apart in time). Indeed, in one of the broader sample comparisons, Krupnikov and Levine (2014) conclude that their study with three samples (students, MTurk, and a population sample) is "only able scratch the surface" (78).

In what follows, we present two studies that offer one of the broadest sample comparisons to date. Study 1 involves three experiments on a population sample and four convenience samples implemented simultaneously. Study 2 presents results from 20 experiments implemented on a population sample and MTurk. Taken together, our data vastly expand the breadth of comparisons, issues, topics, and samples.

We find that the survey experiments we chose largely replicate with distinct samples (i.e., population and convenience samples). The implication is that convenience samples can play a fruitful role as research agendas progress; use of such samples do not appear to consistently generate false negatives, false positives, or inaccurate effect sizes. However, this does *not* mean that costly population samples can be abandoned. Population samples possess a number of

[3] See Huber, Hill, and Lenz (2012) for an argument for the validity of MTurk in a particular political science study. For related work on the implications of experimental samples and settings for causal inference, see Barabas and Jerit (2010); Coppock and Green (2015); Henrich, Heine, and Norenzayan (2010); Jerit, Barabas, and Clifford (2013); Klein et al. (2014); and Valentino, Traugott, and Hutchings (2002).

inherent properties that are lacking or unknowable in convenience samples. For instance, population samples facilitate the testing of heterogeneous treatment effects, particularly in cases where scholars lack a strong theory that identifies the nature of these effects *a priori*. Population-based survey experiments also serve as critical baseline of comparison for researchers seeking to assess the usefulness of ever changing convenience samples (e.g., does the validity of MTurk samples change as respondents continue to participate in literally hundreds of experiments?). Finally, while our results differ from other replication efforts (Open Science Collaboration 2015), it remains unclear just how often survey experiments, beyond the set we chose, replicate. We view our findings as part of an ongoing effort throughout the social sciences to identify the features of experiments that influence the likelihood of replicable and generalizable inferences.

Study 1

For both studies, the source of our population-based sample is the National Science Foundation funded Time-sharing Experiments for the Social Sciences (TESS) program (http://tessexperiments.org/; also see Franco et al. 2014). Since 2001, TESS has invited social scientists to submit proposals to implement population-based experiments. Proposals undergo peer-review and are fielded on a competitive basis. TESS offers graduate students and faculty the opportunity to field population-based experiments at no cost to the investigators themselves.

TESS makes use of what has become a central mode of survey data collection: the use of an ongoing panel of respondents who "declare they will cooperate for future data collection if selected" (Callegaro et al. 2014, 2-3). Specifically, TESS fields experiments using GfK's (formerly Knowledge Networks) online panel, which is based on a representative sample of the U.S. population. TESS data are particularly appealing because their panel is drawn from a probability-based sampling frame that covers 97% of the population (GfK 2013). This helps

ensure representation of minorities and low-income participants, who are often under-represented in non-probability panels.[4]

As explained, the central downside to the population-based sampling approach of TESS is cost: a typical TESS study costs more than $15,000 (with an average N of 1200 the cost per respondent is a bit less than $13.00). Moreover, while TESS offers a "free alternative" to investigators, the likelihood of being accepted to field a TESS survey experiment has become quite low. In 2013, for example, only 11.2% of submitted proposals were accepted; in 2014, 14.4% were accepted. The competitiveness of TESS and the high cost to scholars who want to collect population sample data themselves are likely primary reasons why researchers continue to rely on convenience samples.

In our first study, we implemented three experiments simultaneously on TESS and on 4 of the most common types of convenience samples used in political science. In this study, we focus on a single political science theory: framing. Framing theory has been used for the last quarter century to understand elite rhetoric and political debate (Entman 1993; Gamson and Modigliani 1989; Riker 1996). Experimental findings show that emphasizing particular elements

[4] There is some debate about the importance of having a probability-based panel sample as opposed to non-probability but representative opt-in panel samples (Baker et al. 2010). For their probability sample, GfK uses an established sampling method (presently address-based sampling), and then invites sampled persons to enter the panel, including providing free internet if necessary in exchange for participation (as well as payment for continued survey participation). Thus, nearly every unit in the population (e.g., the United States) has a known and non-zero probability of receiving an invitation to join the panel (Wright and Marsden 2010, 7). By contrast, non-probability population panel samples are often opt-in (Callegaro et al. 2014, 6), though methods of recruitment into the panel and individual studies can vary considerably. This includes highly sophisticated selection algorithms that generate a largely representative sample of populations (e.g., the United States). While a task force report from the American Association for Public Opinion Research states "Researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values... nonprobability samples are generally less accurate than probability samples" (Baker et al. 2010, 714; also see Callegaro et al. 2014, 6), there is debate about the need relative merits of the sampling approaches (e.g., Andrew Gelman and David Rothschild. "Modern Polling Needs Innovation, Not Traditionalism." *The Monkey Cage*. 4 August 2014.). That said, for our purposes, the important point about high quality opt-in samples is that 1) they are often prohibitively expensive for many researchers, not remarkably different from the cost of a TESS study (e.g., estimates we obtained suggested perhaps 30-50% cheaper), and 2) the methods used to create their panels and draw samples are not public information (Callegaro et al. 2014, 6). The question we address, then, would apply to any high quality opt-in survey experiment.

of a political issue alters citizens' preferences and behaviors (Chong and Druckman 2007a, b; Druckman 2001). A now classic example of a framing effect showed that when a newspaper editorial framed a hate group rally in terms of "free speech," readers placed more weight on "speech" considerations and ultimately became more tolerant of the rally (Nelson, Clawson, and Oxley 1997). Due to the wealth of experimental literature in this domain and its heavy reliance on convenience samples (Brady 2000; Klar, Robison, and Druckman 2013; Nelson, Clawson, and Oxley 1997), framing provides a propitious opportunity to explore the consequences of experimental samples for causal inferences.

In each of the three experiments respondents are exposed to one of two different arguments about a policy issue and then asked for their opinion on a seven-point scale (recoded to range from 0 to 1). Treatment effects are measured by the difference in support for each policy in each condition. In the first experiment, respondents are either simply told about the amount of student loan debt held in the United States or are given an argument that frames loan repayment as individuals' personal responsibility. They were then asked, "Do you oppose or support the proposal to forgive student loan debt?" ("Strongly oppose" to "Strongly support"). The second experiment followed from the canonical hate rally tolerance study, providing respondents with either a frame emphasizing free speech considerations or a control condition that simply described a "hypothetical" rally. Respondents were asked, "Do you think that the city should or should not allow the Aryan Nation to hold a rally?" ("Definitely should not allow" to "Definitely should allow"). The final experiment is similar to a recent partisan framing study about the DREAM Act; in this study we exposed respondents to either a "con" frame emphasizing the social burden imposed by immigrants or a no-information control condition (Druckman,

Peterson, and Slothuus 2013).[5] Participants were asked, "To what extent do you oppose or support the DREAM Act?" ("Strongly oppose" to "Strongly support").

The three experiments were implemented in the late fall of 2012 with five distinct (and widely used) samples.[6] The first was a TESS population-based sample. The other samples were convenience samples recruited using common recruitment strategies for political science experiments (Druckman et al. 2006). First, an online sample was recruited using MTurk, paying subjects $0.50 for participation (a la Berinsky et al. 2012). Second, a sample of university staff completed the experiment in-person at individual laptop stations, and were compensated $15 (a la Kam, Wilking, and Zechmeister 2007; Redlawsk, Civettini, and Emmerson 2010). Third, a convenience sample of university undergraduate students, who were compensated by course credit, completed the experiment in-person at individual laptop stations (a la Nelson, Clawson, and Oxley 1997). Last, a sample was recruited at polling places in Evanston, Illinois and Ann Arbor, Michigan after voting in the 2012 general election (a la Druckman 2004; Klar 2013). These respondents were offered $5, with the option of donating it to a charitable organization, to complete experiments via a paper-and-pencil form.

Though recruitment and compensation differ across these five samples, we employ the standard recruitment methods used for each type of sample for reasons of external validity. That is, when experiments are implemented with each of these samples using their typical procedures, what are the consequences for inferences? Holding recruitment and compensation constant

[5] The hate group rally and DREAM Act experiments had additional manipulations, but the similarity in treatment effects between samples is generally consistent across manipulations. Analyses of these additional conditions are shown in the Supplementary Materials.

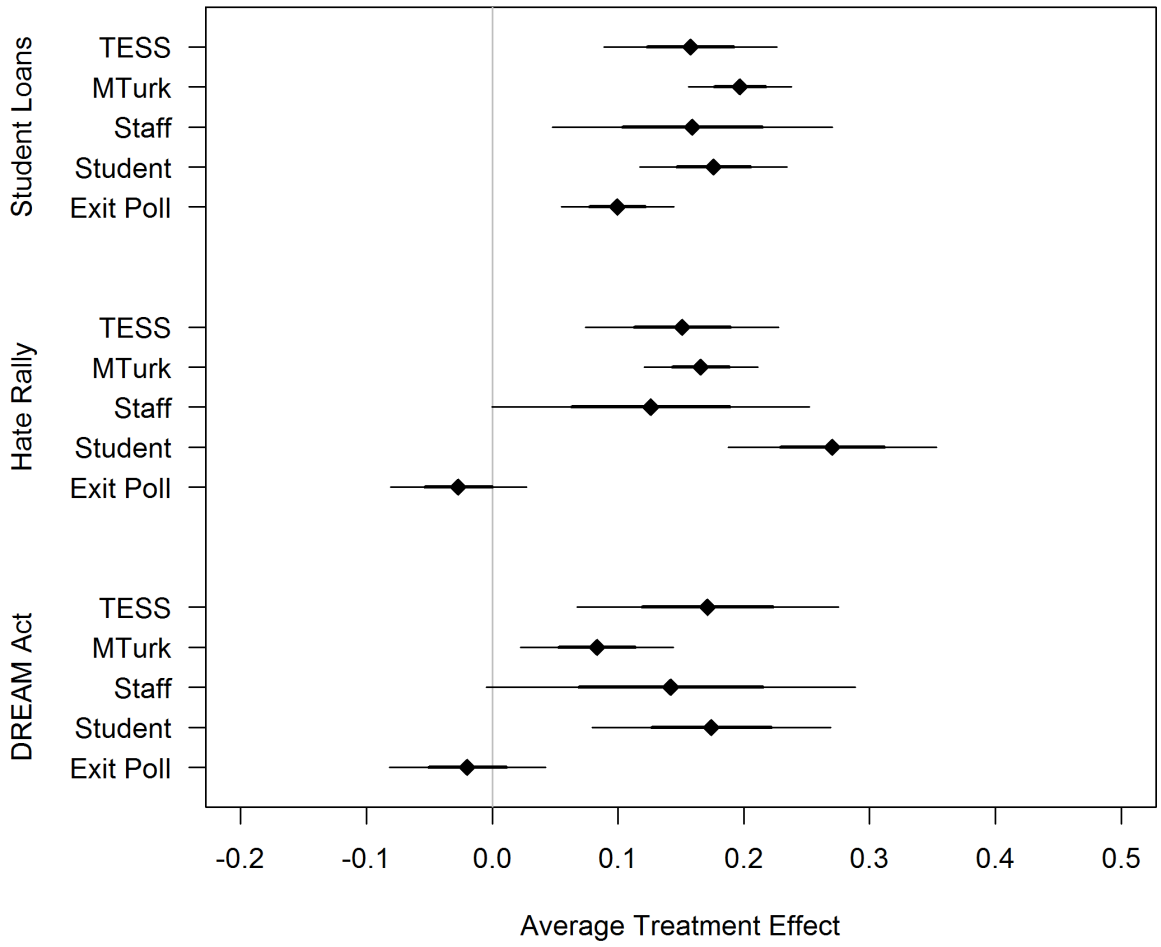
[6] Because Study 1 was executed during a presidential election period, we selected issues that were not receiving substantial attention in the campaign environment so as to avoid any potential contextual confounds. Additionally, research participants completed all three experiments. Consistent with similar framing research on multiple issues, order of experiments was held constant across samples (Druckman, Peterson, and Slothuus 2013).

across all samples would have limited utility because many of the convenience samples would no longer be implemented as they typically are.

The Appendix provides a demographic summary for each sample. The samples differ in age in predictable ways, but differences are not as pronounced on gender. Most of our convenience samples are as racially diverse as the TESS sample, with the Exit Poll supplying a high proportion of African American respondents and TESS under-representing Hispanics.

Due to probability sampling of participants from the U.S. population, the experimental effects drawn from the weighted TESS sample should provide unbiased estimates of treatment effects for the U.S. adult population as a whole. This is the typical approach with TESS data (e.g., weights are provided by GfK). In contrast, we do not weight the convenience samples since it is unconventional to do so (e.g., Berinsky et al. 2012; Druckman 2004; Kam et al. 2007). However, we will discuss the implications of weighting some convenience samples in Study 2. We compare average treatment effects (difference between treatment and control groups) from TESS (our representative baseline) to each of the convenience samples. Figure 1 shows the average treatment effect estimates from our three experiments with bars representing one and two standard errors of the mean-difference generated from a randomization-based permutation distribution. To simplify presentation of results, the direction of effects in the student loan and DREAM Act experiments have been reversed (control-treatment, rather than treatment-control).

Figure 1. Study 1 Results



Note: points are average treatment effects (difference between group means), and bars representing one and two standard errors for the mean-difference.

As expected, the treatment in the student loan forgiveness experiment has a statistically significant effect in the TESS sample. How well do the results from the convenience samples correspond to the TESS sample? Despite differences in the demographic composition of the samples, each convenience sample produces a treatment effect comparable to the TESS sample. That is, each of the convenience samples yields an estimated treatment effect in the same direction as the TESS sample estimate, that is statistically distinguishable from zero, and that is

also statistically indistinguishable from the TESS sample estimate according to a difference-in-difference estimator comparing the treatment-control group differences in each sample.

The results of the second experiment (on tolerance of a hate rally) closely mirror the results of the student loan experiment. The TESS sample yields a large, statistically significant effect of the treatment on support for the rally. The MTurk, university staff, and student samples all yield substantively and statistically similar effect estimates. The exit poll sample, however, yields an estimated effect statistically indistinguishable from zero and substantively pointing in the opposite direction of the TESS result (i.e., emphasizing free speech makes respondents less tolerant). This result appears to be due to very high level of tolerance for the rally in the control condition (i.e., a ceiling effect), possibly due to respondents having just exercised their voting rights moments before participating in the experiment (see Appendix for treatment group means).

The results for the third experiment again closely mirror those of the previous two experiments. As anticipated, TESS respondents exposed to a negative argument about immigration are less supportive of the DREAM Act than the control condition (recall Figure 1 shows a control-minus-treatment difference for this experiment). As in the second experiment, we find substantively and statistically similar results from the MTurk, staff, and student samples. Only the exit poll diverges from this pattern, but we have no definitive explanation, in this case, for this inconsistency.

In sum, all of the convenience samples (save the Election Day exit poll) consistently produce treatment effect estimates similar to TESS in terms of direction and significance. And in most instances, the effects were of a similar magnitude. The exit poll appears most problematic, only providing a comparable inference in the student loan experiment. Future work is needed to

assess whether differences in exit polls (if these results are typical of experiments embedded in exit polls) stem from the sample, context, or implementation technique. Nonetheless, overall, despite differences in demographic composition, the convenience samples – and in particular, student and MTurk – tend to provide substantively similar inferences about each of our treatments.

Yet, this study has limitations. First, it only examines three issues – all of which are built on framing theory. Thus, it is reasonable to ask to what extent the results generalize to other issues. Second, the samples differ in more than just their composition. For example, the university student and university staff samples were administered in-person on laptops whereas the TESS and MTurk samples were completed on-line. Also, the student sample was not financially compensated, but all the other samples were. These differences in implementation were done deliberately, as mentioned, so that each sample was recruited and implemented in a realistic manner, but it limits our ability to infer whether or not the composition of the samples is driving similarities and differences in treatment effects between samples. Finally, there were differences in sample sizes that impact the statistical power associated estimates for each sample.

Study 2

Study 2 complements Study 1 by addressing several of the aforementioned issues. First, we examine a much broader range of issues. Second, we focus on comparisons of the average treatment effects between MTurk samples and TESS population-based samples, so that the experiments can be implemented in an online mode in a maximally similar manner. Third, we conduct the experiments with large, comparably sized samples on both platforms. Note that, unlike study 1, where the TESS studies were newly implemented in concert with the other samples, here we rely on previously implemented TESS studies (for which again we apply the

relevant sampling weights as in Study 1), and compare them with newly implemented (unweighted) MTurk. While we could have compared the TESS sample directly to other convenience samples as we did in Study 1, we limited our focus to a single convenience sample (MTurk) in order to assess a larger number of issues in a manner that was feasible. MTurk is an increasingly popular avenue for experimental research across the social sciences (Bohannon 2011) and related research on the utility of the platform has been conducted but only with a small number of issues (Berinsky et al. 2012; Krupnikov and Levine 2014).[7]

We selected a total of 20 survey experiments that had been implemented using the TESS survey population sample platform. Ideally, in terms of selection of studies, we would have randomly sampled experiments from TESS archives, but this approach was not feasible for several reasons. First, TESS experiments with samples over 4000 respondents were not included. Second, experiments had to be able to be implemented in the survey software we used for the MTurk experiments (Qualtrics).[8] Third, many TESS experiments use subsamples of the population of one sort of another (e.g., Democrats, white respondents, respondents with children); we used only experiments intended to be fielded on the population-at-large. Finally, we restricted consideration to relatively recent TESS experiments for which we did not expect the treatment effect to be moderated by a precise time period (since we collected the MTurk data after the TESS data were collected). After eliminating potential experiments from the TESS archives based on these criteria, at the time of our implementation we were left with the 20 experiments shown in Table 1. As will be clear in our results, we did not select experiments

[7] There are two debates about internet panels that are beyond our purview here. First is whether a low response rate to a survey creates a problem for representativeness. Some studies suggest that response rate is orthogonal to representativeness and data quality (e.g., Keeter et al. 2006; Pew 2012); however, it is an ongoing question as internet panels continue to grow (see Steinmetz et al. 2014). Second, when it comes to any panel, although particularly opt-in panels, there is the question of whether there is an effect from participating in multiple surveys and/or whether the participants differ in their original motivation from non-participants (see Hillygus et al. 2014).

[8] A number of TESS studies require relatively complex programming by professionals at GfK. We were limited to studies that we were capable of programming ourselves in Qualtrics.

based on whether significant effects had been obtained using TESS, as this would bias comparisons because replications of experiments selected on statistical significance are expected to have a smaller average effect size than the original studies (Kraft 2008).

Table 1. Study 2 Experiments

Experiment Number	TESS Experiment Title	Lead TESS Principal Investigator
1	Onset and Offset Controllability in Perceptions and Reactions to Home Mortgage Foreclosures	Brandt, M.
2	To Do, to Have, or to Share? Valuing Experiences and Material Possessions by Involving Others	Caprariello, P.
3	Perceptions of Migration and Citizenship in the United States	Creighton, M.
4	Public Attitudes about Political Equality	Flavin, P.
5	Understanding How Policy Venue Influences Public Opinion	Gash, A.
6	Patient Responses to Medical Error Disclosure: Does Compensation Matter?"	Mello, M.
7	Informing the Public or Information Overload? The influence of school accountability data format on public satisfaction."	Jacobsen, R.
8	Terrorism Suspect Identity and Public Support for Controversial Detention and Interrogation Practices	Piazza, J.
9	Why Hillary Rodham Became Hillary Clinton: Consequences of Non-Traditional Last Name Choice in Marriage	Shafer, E.
10	Terrorist Threat: Overreactions, Underreactions, and Realistic Reactions	Thompson, S.
11	Environmental Values, Beliefs, and Behavior	Turaga, R.
12	The Reputational Consequences of International Law and Compliance	Wallace, G.
13	Unmasking Expressive Responses to Political Rumor Questions	Berinsky, A.
14	Social Desirability Bias	Kleykamp, M.
15	Smallpox Vaccine Recommendations: Is Trust a Shot in the Arm?	Parmer, J.
16	With God on Our Side	Converse, B.
17	Examining the Raced Fatherhood Premium	Denny, K.
18	The Mechanisms of Labor Market Discrimination	Pedulla, D.
19	An Experiment in the Measurement of Social and Economic Ideology	Jackson, N.
20	The Flexible Correction Model and Party Labels	Bergan, D.

The experiments address diverse phenomena such as perceptions of mortgage foreclosures, how policy venue impacts public opinion, and how the presentation of school accountability data impacts public satisfaction (see Supplementary Materials for details of each

experiment).[9] Testing across such a broad range of issues enables us to test whether some unexpected and/or unmeasured feature of the MTurk sample generates bias (e.g., Weinberg et al. 2014 note that some have suggested that people who seek out opportunities to participate in experiments online at sub-minimum wage rates may be unusual in various respects in terms of undocumented moderators).

We implemented the 20 experiments in ways that maximized assurance that differences stem from differences in samples, rather than differences in instrumentation. We used identical wording and virtually identical formatting. We also employed sample sizes that were as close as possible (given response rates) to TESS. As such, we obtained what is, to our knowledge, one of the largest pools of MTurk workers for social science experiments – over 9,500 unique Worker IDs across the 20 experiments. We paid about $.40 cents per respondent per experiment (see work on pay rates; Berinsky, Huber, and Lenz 2012).[10]

We focus analyses on the first post-stimuli dependent variable – since these variables are the primary focus of the experiments as proposed by the TESS investigators. We made comparisons between a control group and what clearly were the two main treatment groups for the experiment, or if no control group was included, between the conditions that clearly tested the main dimensions of interest. Four experiments only had two conditions, and as such, we only compare those two conditions.[11] By making simple group comparisons and focusing on only the

[9] Specifically, the number of experiments by the discipline of the lead investigator is as follows: eight from political science and public policy, six from sociology, three from psychology, one from communication, one from education, and one from law and public health.

[10] Most TESS experiments are implemented independently. We conducted analyses to determine whether fielding experiments independently on MTurk yielded different results from bundling multiple experiments into a single survey (with order randomized) to further reduce costs. Across four substantively distinct experiments, we found no evidence of a systematic effect of bundling (Supplementary Materials Figure S1), and so the remaining MTurk experiments were implemented using bundling. Although we tried to obtain similar sample sizes in MTurk and TESS, the use of bundling did result in some experiments with a larger sample size in MTurk.

[11] To ensure MTurk workers attended to the study task, we compared the percentage of correct respondents to three manipulation-check questions in two of our experiments (the only ones that included such checks in the original

first post-stimuli dependent variable, we are taking a uniform analytical approach in our assessment of these experiments. However, we emphasize that this may or may not be the analytical strategy employed the TESS Principal Investigators who designed these experiments. These investigators may have employed different analytical and modeling techniques or focused on different dependent variables.

Tables in the Appendix show the demographic data collected in our 20 experiments for both samples, and are consistent with previous research (e.g. Berinsky, Huber, and Lenz 2012). Among other differences, the MTurk respondents are younger and more educated than TESS respondents. The gender composition of the samples is quite similar.

Figure 2 shows the difference between group means for the control group and each experiment's first treatment group separately for the weighted TESS sample and the unweighted MTurk sample. Studies are sorted by magnitude of the effect size of the weighted TESS sample, which has been signed positive for all experiments (see Table 1 for topics of each experiment number, and Supplementary Materials for additional study details).

Figure 2 reveals that, generally, the two samples produce similar inferences with respect to the direction of the treatment effect and statistical significance. Indeed, fifteen of the twenty experiments produce the same inference. That is, when TESS produces a statistically significant treatment effect in a particular direction, a significant effect in the same direction is produced by MTurk; or, when there is a null effect in TESS there is a null effect in MTurk. Yet, there are five deviations from this overall pattern (Experiments 2, 11, 16, 17, 20). In these instances, there is a significant result in one sample, but a result statistically indistinguishable from zero in the other.

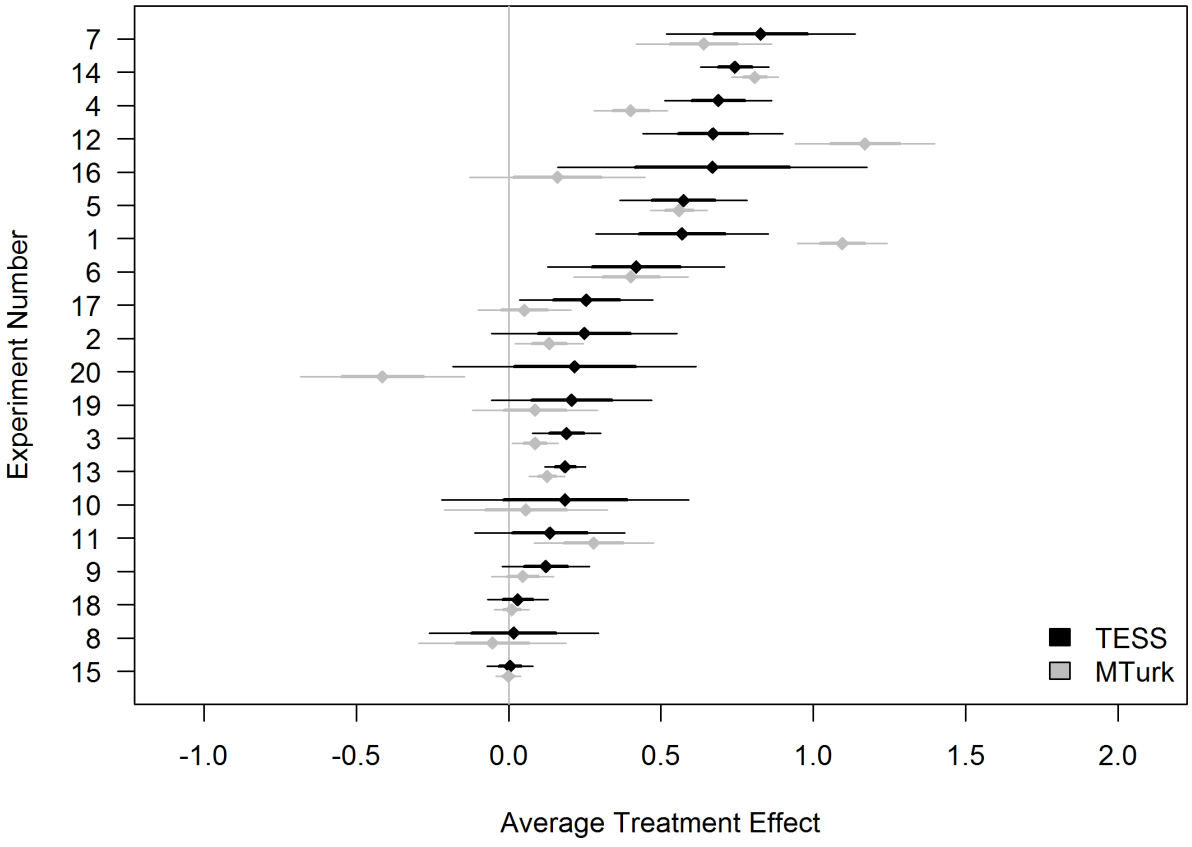
---

designs). The MTurk respondents were actually significantly more likely to answer the questions correctly than the TESS sample (also see Druckman and Kam 2011). Details are in Supplementary Materials Table S1. This finding is consistent with other research on the attention-levels of MTurk workers (Clifford and Jerit 2015; Weinberg et al. 2014). Although not employed here, Berinsky et al. (2014) have suggested that screener questions can be used to address concerns about attention levels in Mturk.

There is no clear pattern whereby one sample consistently produces the larger treatment effect. Importantly, there is not a single instance in which the samples produce significant effects in opposite directions.

We also compare magnitude of effects. An analysis of the difference in effect sizes between samples (i.e., a difference-in-differences) reveals that across the 20 experiments, in only 4 experiments (1, 4, 12, 20) do the samples generate statistically distinguishable effect sizes. In two cases, MTurk overestimates the treatment effect (1, 12), in one it underestimates the effect (4), and in only one (20) it yields a significant effect when the TESS sample indicated no effect.

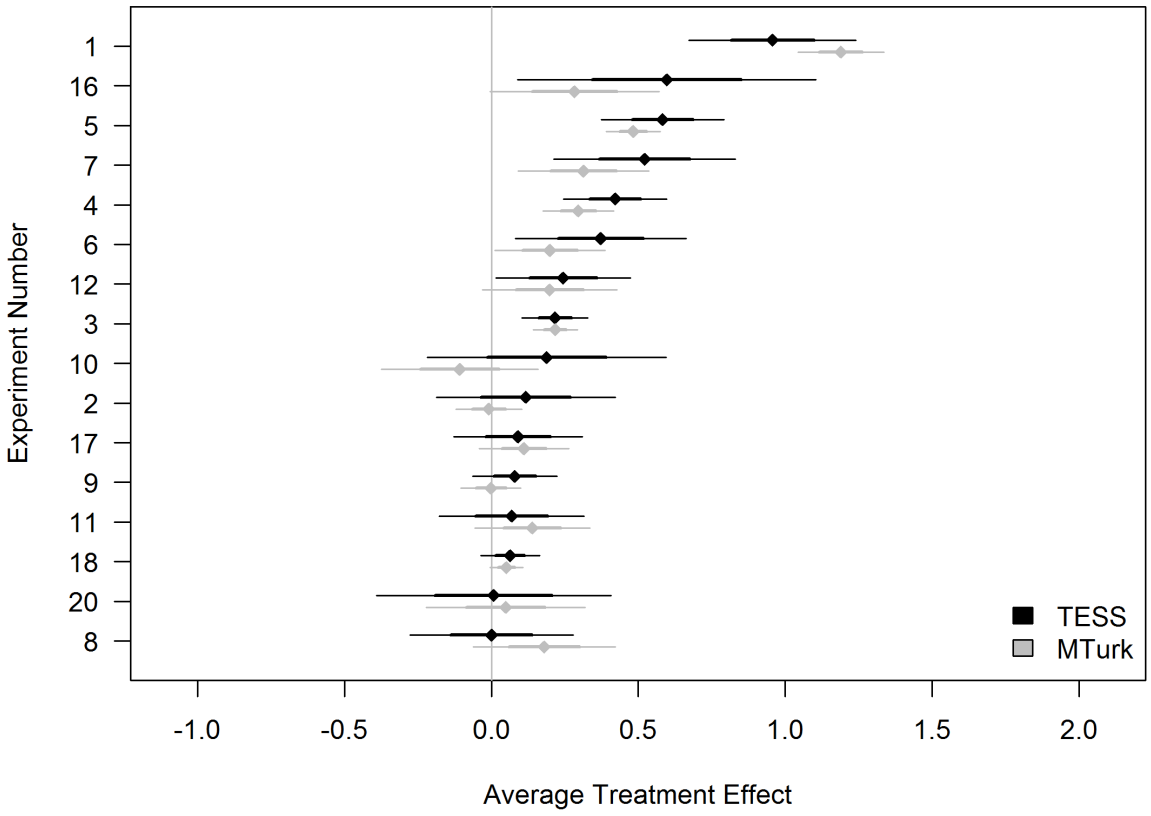
Figure 2: Control vs. Treatment Group 1



Note: points are average treatment effects (difference between control and treatment group means), and bars representing one and two standard errors for the mean-difference. Many of the experiments have multiple treatment groups. This figure focuses on the first treatment group.

These results are buttressed by Figure 3, which presents analyses of a second treatment group relative to control for the 16 (of 20) experiments that had a second treatment group. Again, the inferences with respect to the direction and statistical significance of treatment effects are quite similar between samples. Of the sixteen experiments, fourteen of the TESS treatment effects are replicated in MTurk in terms of direction and statistical significance. Only two experiments diverge from this overall pattern (Experiments 12, 18), but even these cases reflect one experiment barely exceeding the threshold of statistical significance while the other barely falls short of statistical significance. In none of the experiments is there a significant difference in the apparent effect size between samples.

Figure 3: Control vs. Treatment Group 2



Note: points are average treatment effects (difference between control and treatment group means), and bars representing one and two standard errors for the mean-difference. Many of the experiments have multiple treatment groups. This figure focuses on a second treatment group.

In sum, 29 (or 80.6%) of the 36 treatment effects in Figures 2 and 3 estimated from TESS are replicated by MTurk in the interpretation of the statistical significance and direction of treatment effects. Importantly, of the 7 experiments for which there is a significant effect in one sample, but a null result in the other, only one (Experiment 20) actually produced a significantly different effect size estimate (Gelman and Stern 2006). Across all tests, in no instance did the two samples produce significantly distinguishable effects in substantively opposite directions.

Although sample weighting is not the primary focus of this paper (i.e., we did not weight convenience samples because they are typically used without weights), we explored the possibility of weighting MTurk data using the same variables and data that GfK uses for its post-survey weighting.[12] The results are shown in Figures AF1 and AF2 in the Appendix (Figure AF3 shows results comparing treatment groups, where applicable). Results were decidedly mixed: for the seven treatment effects for which the samples differed in interpretation of statistical significance, the re-weighting of MTurk data eliminated two of these differences (11, 20), but exacerbated between-sample differences in two others (9, 19). Clearly more research is needed to understand the consequences of even basic weighting adjustments to improve the generalizability of causal inferences from convenience samples.

Discussion

As funding for social science decreases (Lupia 2014), technological improvements allow researchers to implement human subjects research at ever-lower costs. Novel types of convenience samples, such as MTurk, have been described as "social science for pennies" (Bohannon 2011). Indeed, although the actual costs varied slightly by experiment, a single study

[12] We weighted the MTurk data to the January 2014 Current Population Survey marginal distributions on sex, age, race, education, and region (variables used in the TESS weighting scheme) using iterative proportional fitting (raking). Note that weights in TESS data are a combination of sampling weights and post-survey weights.

in TESS costs about $15,000 while the same study was implemented with a comparable sample size on MTurk for about $500 (or even less in some of the other convenience samples). It is important to understand the implications of these alternative data collection approaches both to optimize resource allocation and to ensure progress of basic (e.g., Mutz 2011) and applied (e.g., Bloom 2005) research.

We find that, generally speaking, results from convenience samples provide estimates of causal effects comparable to those found on population-based samples. As mentioned, this differs somewhat from other broad replication efforts in neighboring disciplines (Open Science Collaboration 2015: 943). Varying replication rates may stem from an assortment of factors that produce treatment effect heterogeneity—such as the canonical dimensions of external validity sample, settings, treatments, and outcome measures (Shadish et al. 2002: 83), from uneven delineation or implementation of experimental protocol, or variation in topic/discipline. Clearly, more work is needed to identify conditions that influence experimental replicability (see, e.g., Hovland 1959; Barabas and Jerit 2010; Jerit, Barabas, and Clifford 2013; Coppock and Green 2015).

Of equal, if not greater importance, are what our findings suggest when it comes to using convenience samples in experimental research. Our results may be reassuring for those who have little choice but to rely on cheaper convenience samples; yet, one should *not* conclude that convenience samples are a wholesale or even partial substitute for population samples. For one, replications do not always succeed with different samples. Moreover, there are at least three reasons why population samples remain critical to social science experimentation. First, when one uses a convenience sample, its relationship to the population of interest is unknown and typically unknowable. Thus, one cannot assuredly conclude it generalizes, even if the

demographics of the sample seem to match the demographics of the larger population of interest (e.g., United States citizens) or if data are reweighted to match population distributions. There always exists the possibility that unmeasured features of the sample skew it from the population of interest. In cases where a given sample ostensibly matches the population of interest on key variables, it may still have problematic joint distribution properties. For example, relative to a population-based sample, a convenience sample may have similar percentages of older individuals and racial minorities, but may not match the population-based sample with respect to older minorities (Freese et al. 2015; Huff and Tingley 2015). These types of uncertainties inherent in convenience sample also vitiate their potential impact in some applied settings.

Second, experiments often have heterogeneous treatment effects such that the treatment effect is moderated by individual-level characteristics (e.g., the treatment effect differs among distinct subgroups of the sample; see Gerber and Green 2012: 310-311) or contextual variations (timing, geography, etc.). Recall the Hurricane Katrina experiment we described at the start of the paper—it could be that the treatment effect of offering officials' job descriptions lessened the impact of partisanship in opinion formation among weakly identified partisans but less so (or not at all) among strongly identified person. In this case, there is heterogeneity in the treatment effect depending on subgroups. If one has a well-developed theory about heterogeneous treatment effects, then convenience samples only become problematic when there is a lack of variance on the predicted moderator (e.g., the sample consists largely of strong partisan individuals) (Druckman and Kam 2011). Even with a theory in hand some convenience samples would be inappropriate such as a student sample where a moderator is age, a university staff sample where a moderator is education, or MTurk when a moderator is religion (i.e., MTurk samples tend to be substantially less religious than the general population).

Moreover, in reality, many areas of the social sciences have not developed such precise theories. Scholars have consequently begun to employ machine learning algorithms that automate the search for heterogeneous treatment effects (e.g., Green and Kern 2011, Egami and Imai 2015). In so doing, population samples have the unique advantage not only of containing substantial variance on the full range of population demographics, each of which could potentially moderate, but also of avoiding the joint distribution problem mentioned above

Third, the nature of convenience samples can change over time. This is particularly true of MTurk for which there is a growing concern that respondents have evolved to be less and less like respondents in other surveys (even survey panels).[13] Rand et al. (2014) report that in MTurk data collected between February 2011 and February 2013, the median MTurk respondent reported participation in 300 academic studies, 20 of which were in the last week; moreover, they note that, over the time period they studied, "the MTurk subject pool [had] transformed from naïve to highly experienced… [and this] makes it likely that subjects will be familiar more generally with experimental paradigms…" (4-5; also see Chandler, Mueller, and Paolacci 2014). Relatedly, it could be that MTurk respondents may differ in terms of fundamental motivation, based on how often they participate in surveys. Some participate strictly to earn money through piecework, and opt-in or randomly selected survey respondents, while others participate in survey experiments more for intrinsic rewards or other non-monetary reasons. The ethics of this difference in relationship between researcher and subject, and any possible empirical consequences thereof, merit further consideration (c.f. Dynamo, 2014). Notably, what is considered a fair incentive for study participation on MTurk is likely to change over-time and the particular rewards offered here may not be appropriate in the future. There are thus various reasons to closely monitor whether MTurk becomes less reliable in terms of replicating

[13] Research also suggests that the demographic composition of MTurk has evolved over time (Ross et al. 2010).

population-based experimental inferences. Researchers should also be cognizant of crowd-sourcing platforms beyond MTurk (Benoit et al. 2015).

One can only assess the implications of the changing nature of any convenience sample if there is a relevant population sample with which to compare. In short, population survey experiments serve as a critical baseline that allows researchers to assess the conditions under which convenience samples provide useful or misleading inferences. Indeed, we began by stating that assessing the validity of any convenience sample is an empirical question and going forward that will continue to be the case—and can only be evaluated with the continued wide-scale implementation of population-based survey experiments.[14]

In sum, convenience samples can play a fruitful role as research agendas progress. They are useful testing grounds for experimental social science. Yet, they *do not replace* the need for studies on population samples; rather, convenience samples serve as a place to begin to test hypotheses and explore whether they are falsified, which coheres with the Popperian approach to causation (Campbell 1969, 361). Our efforts highlight that scientific knowledge advances through replication rather than accepting or rejecting research based on sample-related heuristics. Convenience samples can lead to substantial progress in the social sciences, most acutely when researchers understand the conditions under which those samples are more or less likely to provide generalizable population inferences. This can best be done through theory and continued empirical comparisons across samples. As such, our findings contribute to more efficient and robust experimental social sciences that generate data for more studies by taking unreserved advantage of cost-effective ways of conducting studies when they are likely to provide a good

[14] Yet the validity of population-based samples must also be evaluated. With growing nonresponse rates and an almost universal reliance on empanelled respondents, it is increasingly difficult to claim purely design-based population inferences from any sample. Such challenges highlight the need in all survey-based research of thinking through and justifying design and analytic decisions if the inferential goal is to make claims about a given population as a whole.

reflection of population estimates. An inexpensive and high quality platform for implementing survey experiments not only reduces the cost of traditional experiments, but allows researchers to explore more complex and over-time designs (Ahler 2014; Fowler and Margolis 2014). In so doing, we can more judiciously save the strengths of population-based samples for projects with the strongest justification that the extra expense is needed for accurate inference.

Works Cited

Ahler, Douglas J. 2014. "Self-Fulfilling Misperceptions of Public Polarization." *The Journal of Politics* 76(3): 607-620.

Baker, Reg, et al. 2010. "Research Synthesis: AAPOR Report on Online Panels." *Public Opinion Quarterly* 74(4): 1-71.

Barabas, Jason and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104 (May): 226-242.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2015. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review*: Forthcoming.

Berger, Arthur Asa. 2014. *Media and Communications Research Methods: An Introduction to Qualitative and Quantitative Approaches*. Los Angeles: Sage Publication, Inc.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (Summer): 351-368.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58(3): 739-753.

Bloom, Howard S. 2005. *Learning More from Social Experiments*. New York: Russell Sage Foundation.

Bohannon, John. 2011. "Social Science for Pennies." *Science* 334 (October): 307.

Brady, Henry E. 2000. "Contributions of Survey Research to Political Science." *PS: Political Science & Politics* 33(1): 47-57.

Broockman, David E., and Donald P. Green. 2013. "Do Online Advertisements Increase Political Candidates' Name Recognition or Favorability? Evidence from Randomized Field Experiments." *Political Behavior* 36(2): 263-289.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6(1): 3-5.

Callegaro, Mario, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas. 2014. "Online Panel Research: History, Concepts, Applications, and a Look at the Future." In Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas, eds., *Online Panel Research: A Data Quality Perspective*. West Sussex, United Kingdom: John Wiley & Sons Ltd.

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

Cassese, Erin C., Leonie Huddy, Todd K. Hartman, Liliana Mason, and Christopher R. Weber. 2013. "Socially Mediated Internet Surveys: Recruiting Participants for Online Experiments." *PS: Political Science and Politics* 46(4): 1-10.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaiveté Among Amazon Mechanical Turk Workers: Consequences and Solution for Behavioral Researchers." *Behavior Research Methods* 46(1): 112-130.

Chong, Dennis and James N. Druckman. 2007a. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101(4): 637-655.
Chong, Dennis and James N. Druckman. 2007b. "Framing Theory." *Annual Review of Political Science* 10(1): 103-126.
Clifford, Scott and Jennifer Jerit. 2015. "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.
Coppock, Alexander and Donald P. Green. 2015. "Assessing the Correspondence Between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research." *Political Science Research and Methods* 3(1): 113-131.
Druckman, James N. 2001. "The Implications of Framing Effects for Citizen Competence." *Political Behavior* 23(3): 225-256.
Druckman, James N. 2004. "Priming the Vote: Campaign Effects in a US Senate Election." *Political Psychology* 25: 577-594.
Druckman, James N. and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base'." In *Cambridge Handbook of Experimental Political Science*, eds. J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. New York: Cambridge University Press, 41-57.
Druckman James N., and Arthur Lupia. 2012. "Experimenting with Politics." *Science* 335 (March): 1177-1179.
Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100 (November): 627-635.
Druckman, James N., Erik Peterson, and Rune Slothuus. 2013. "How Elite Partisan Polarization Affects Public Opinion Formation." *American Political Science Review* 107(1): 57-79.
Dynamo. 2014. "Guidelines for Academic Requesters." Retrieved 6 October 2015 from http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters.
Egami, Naoki, and Kosuke Imai. 2015. "Causal Interaction in High-Dimension." Working paper.
Entman, Robert M. 1993. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication* 43(4): 51-58.
Fowler, Anthony and Michele Margolis. 2014. "The Political Consequences of Uninformed Voters." *Electoral Studies* 34: 100-110.
Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (August): 1502-1505.
Freese, Jeremy, Adam Howat, Kevin J. Mullinix, and James N. Druckman. 2015. "Limitations of Screening Methods to Obtain Representative Samples Using Online Labor Markets." Working Paper, Northwestern University
Gamson, William A., and Andre Modigiliani. 1989. "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach." *American Journal of Sociology* 95(1): 1-37.
Gelman, Andrew, and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant." *The American Statistician* 60(4): 328-331.
Gerber, Alan S. and Donald P. Green. 2008. "Field Experiments and Natural Experiments." In *Oxford Handbook of Political Methodology*, eds. J.M. Box-Steffensmeier, H. E. Brady, and D. Collier. New York: Oxford University Press, 357-381.

Gerber, Alan S. and Donald P. Green. 2011. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton & Company.
Gerring, John 2012. *Social Science Methodology: A Unified Framework*. New York: Cambridge University Press.
GfK. 2013. "Knowledge Panel Design Summary." http://www.gfk.com/Documents/GfK-KnowledgePanel-Design-Summary.pdf
Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema. 2012. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26: 213-224.
Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491-511.
Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (April): 61-83.
Hillygus, D. Sunshine, Natalie Jackson, and McKenzie Young. 2014. "Professional Respondents in Nonprobability Online Panels." In Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas, eds., *Online Panel Research: A Data Quality Perspetive*. West Sussex, United Kingdom: John Wiley & Sons Ltd.
Holt, Charles A. 2006. *Markets, Games, and Strategic Behavior: Recipes for Interactive Learning*. Addison-Wesley.
Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14: 399-425.
Hovland, Carl I. 1959. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change." *The American Psychologist* 14: 8-17.
Huber, Gregory A., Seth J. Hill, and Gabriel S. Lenz. 2012. "Sources of Bias in Retrospective Decision-Making: Experimental Evidence of Voters' Limitations in Controlling Incumbents." *American Political Science Review* 106(4): 720-741.
Huff, Connor, and Dustin Tingley. 2015. "'Who are these people?' Evaluating the demographic characteristics and political preferences of MTurk survey respondents." Research & Politics 2(3). DOI: 10.1177/2053168015604648.
Iyengar, Shanto. 1991. *Is Anyone Responsible? How Television Frames Political Issues*. Chicago, IL: The University of Chicago Press.
Jerit, Jennifer, Jason Barabas, and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly* 77 (1): 256-282.
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(4): 415-440.
Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70(5): 759–779.
Klar, Samara. 2013. "The Influence of Competing Identity Primes on Political Preferences." *Journal of Politics* 75(4): 1108-1124.
Klar, Samara, Joshua Robison, and James N. Druckman. 2013. "Political Dynamics of Framing." In *New Directions in Media and Politics*, ed. Travis N. Ridout. New York: Routledge, 173-192.

Klein, Richard A., et al. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Social Psychology* 45: 142-152.
Kraft Peter. 2008. "Curses—Winner's and Otherwise—in Genetic Epidemiology." *Epidemiology* 19 (September): 649-651.
Kriss, Peter H., and Roberto Weber. 2013. "Organizational Formation and Change: Lessons from Economic Laboratory Experiments." In *Handbook of Economic Organization: Integrating Economic and Organizational Theory*, ed. A. Grandori. Northampton: Edward Elgar Publishing Limited, 245-272.
Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1 (Spring): 59-80.
Lupia, Arthur. 2014. "The 2013 Ithiel de Sola Pool Lecture: What is the Value of Social Science? Challengers for Researchers and Government Funders." *PS: Political Science & Politics* 47 (January): 1-7.
Malhotra, Neil and Alexander G. Kuo. 2008. "Attributing Blame: The Public's Response to Hurricane Katrina." *The Journal of Politics* 70:1, 120-135.
McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis*, 10: 325-342.
Morawski, Jill G. 1988. *The Rise of Experimentation in American Psychology*. New Haven: Yale University Press.
Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.
Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91(3): 567-583.
Nock Steven L, and Thomas M. Guterbock. 2010. "Survey experiments." In Handbook of Survey Research, eds Marsden PV, Wright JD (Emerald, UK), 837-864.
Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349: 943.
Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making*, 5(August): 411-419.
Pew. 2012. *Assessing the Representativeness of Public Opinion Surveys*. www.peoplepress.org.
Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen Wurzbacher, Martin A. Nowak, and Joshua D. Greene. 2014. "Social Heuristics Shape Intuitive Cooperation." *Nature Communications* 5: 1-12.
Redlawsk, David P., Andrew J. Civettini, and Karen M. Emmerson. 2010. "The Affective Tipping Point: Do Motivated Reasoners Ever 'Get It'?" *Political Psychology* 31: 563-593.
Riker, William H. 1996. *The Strategy of Rhetoric: Campaigning for the American Constitution*. New Haven, CT: Yale University Press.
Ross, Joel, Lily Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. "Who are the Crowdworkers? Shifting Demographics in Amazon Mechanical Turk. In *CHI EA 2010*, New York: ACM Press, 2863-2872.
Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology*. 51: 515-530.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.

Sniderman, Paul. 2011. The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation." In *Cambridge Handbook of Experimental Political Science*, eds. J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. New York: Cambridge University Press, 102-114.

Steinmetz, Stephanie, Annamaria Bianchi, Kea Tijdens, and Silvia Biffignandi. 2014. "Improving Web Surveys Quality: Potentials and Constraints of Propensity Score Adjustments." In Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas, eds., *Online Panel Research: A Data Quality Perspetive*. West Sussex, United Kingdom: John Wiley & Sons Ltd.

Valentino, Nicholas A., Michael W. Traugott, and Vincent L. Hutchings. 2002. "Group Cues and Ideological Constraint: A Replication of Political Advertising Effects Studies in the Lab and in the Field." *Political Communicaton* 19(1): 29-48.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-representative Polls." *International Journal of Forecasting*, in press.

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

Wright, James D., and Peter V Marsden. 2010. "Survey Research and Social Science: History, Current Practice, and Future Prospects." In *Handbook of Survey Research*, eds PV Marsden, JD Wright. Emerald, UK, 3-26.

Appendix

Study 1 Student Loans Experiment Treatment Group Means, Effects, and Sample Sizes

	Treatment	Control	Effect	SE	N
Exit Poll	0.70	0.60	0.10	0.02	739
Student	0.72	0.54	0.18	0.03	292
Staff	0.68	0.52	0.16	0.06	128
MTurk	0.69	0.49	0.20	0.02	1009
TESS	0.50	0.34	0.16	0.03	593

Study 1 Hate Rally Experiment Treatment Group Means, Effects, and Sample Sizes

	Treatment	Control	Effect	SE	N
Exit Poll	0.60	0.63	-0.03	0.03	739
Student	0.69	0.42	0.27	0.04	292
Staff	0.64	0.52	0.13	0.06	128
MTurk	0.68	0.52	0.17	0.02	1005
TESS	0.59	0.44	0.15	0.04	593

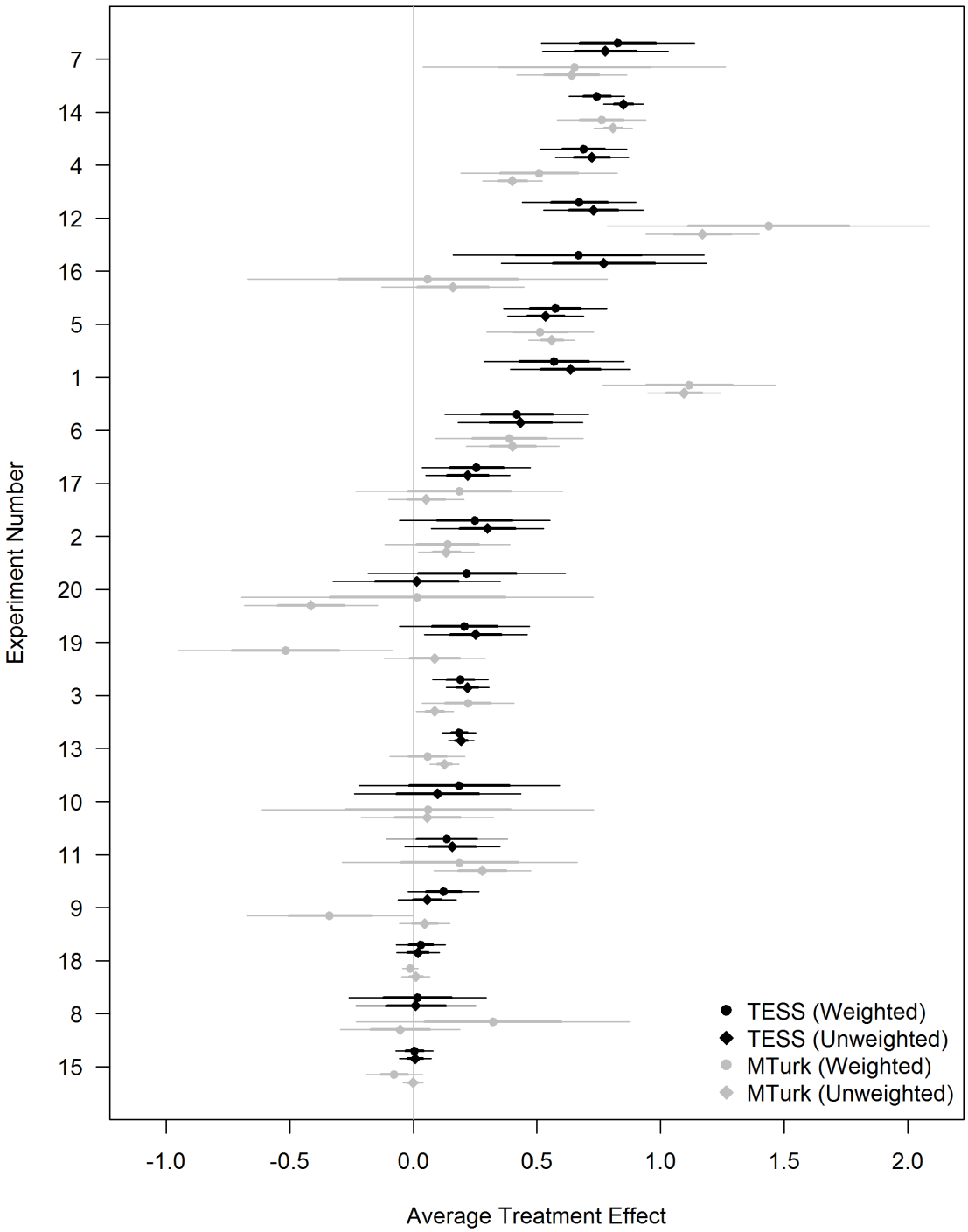
Study 1 DREAM Act Experiment Treatment Group Means, Effects, and Sample Sizes

	Treatment	Control	Effect	SE	N
Exit Poll	0.82	0.84	-0.02	0.03	301
Student	0.87	0.69	0.17	0.05	110
Staff	0.75	0.60	0.14	0.07	54
MTurk	0.66	0.58	0.08	0.03	404
TESS	0.67	0.50	0.17	0.05	133

Study 1 Demographics

	Female (%)	18-24 (%)	25-34 (%)	35-50 (%)	51-65 (%)	65+ (%)	White, Non-Hispanic (%)	Black, Non-Hispanic (%)	Hispanic (%)
TESS	51.10	9.27	15.35	22.77	33.73	18.89	77.91	5.56	0.00
Exit Poll	60.77	36.45	26.81	36.75	0.00	0.00	67.61	12.96	1.62
Student	56.36	99.65	0.35	0.00	0.00	0.00	64.38	5.14	7.19
Staff	50.79	33.06	46.28	20.66	0.00	0.00	60.16	6.25	2.34
MTurk	41.67	38.60	42.04	19.35	0.00	0.00	75.98	6.45	4.98

Figure AF1: Control vs. Treatment Group 1



Note: Points are average treatment effects (difference between control and treatment group means), and bars represent one and two standard errors for the mean-difference. Figure is sorted

by the magnitude of the effect size of the weighted TESS sample, which has been signed positive for all experiments. Weighting of the MTurk sample is based raking to the January 2014 Current Population Survey estimates of the U.S. household population, using a method analogous to that used by GfK to weight their samples. The larger error bars for the weighted MTurk sample are due to missingness on key demographic variables used in the weighting process; no imputation has been used.

Study 2 Treatment Group 1 Treatment Group Means, Effects, and Sample Sizes (TESS Weighted and TESS Unweighted)

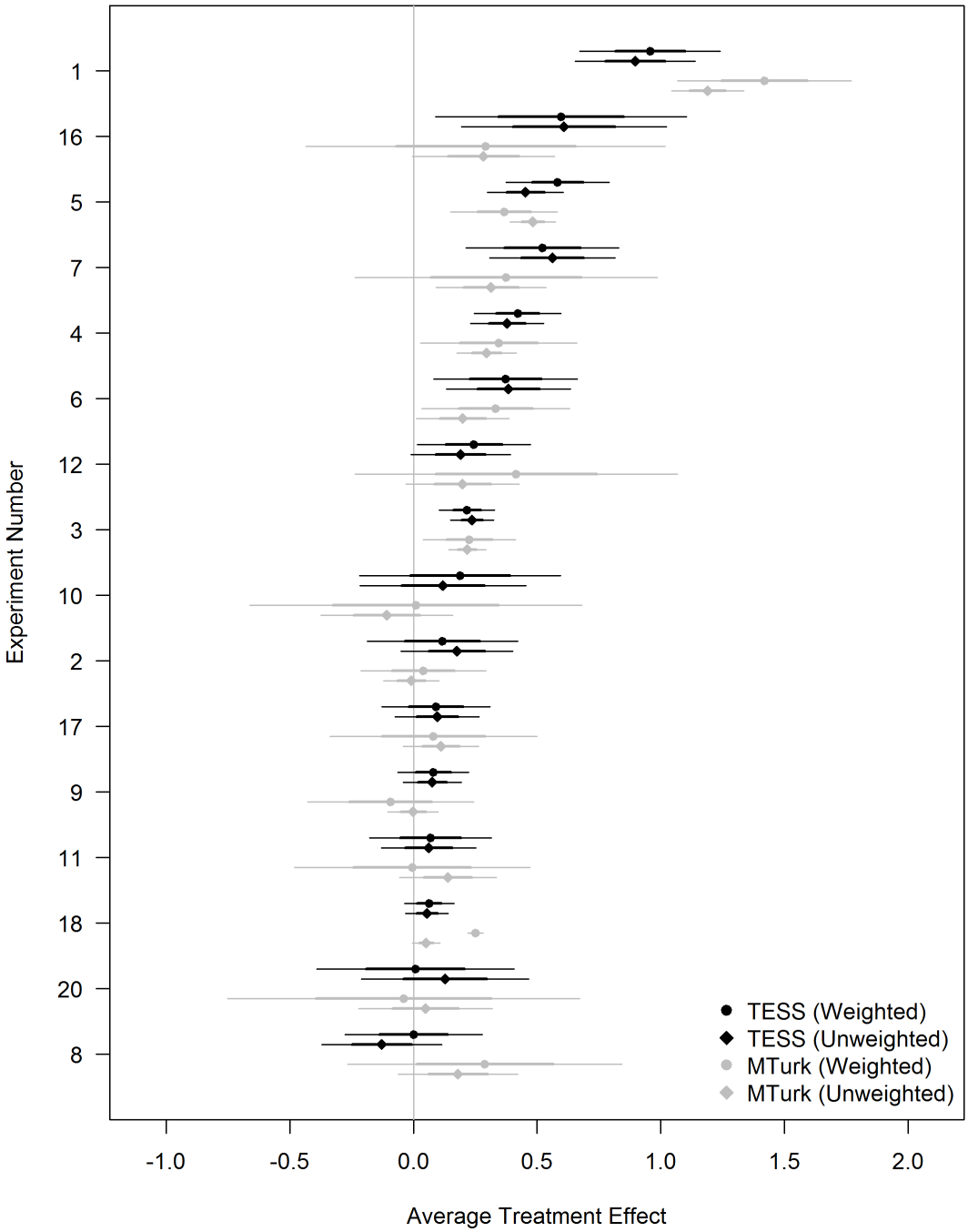
	Control	Treatment	Effect	N	Control	Treatment	Effect	N	DID (SE)
1	5.77	5.20	-0.57	625	5.90	5.27	-0.64	625	-0.53 (0.16)
2	3.48	3.23	-0.25	399	3.54	3.24	-0.30	399	0.12 (0.16)
3	1.88	2.07	0.19	1606	1.90	2.12	0.22	1606	-0.10 (0.07)
4	2.59	1.91	-0.69	770	2.61	1.89	-0.72	770	0.29 (0.11)
5	2.36	1.78	-0.57	496	2.35	1.81	-0.53	496	0.01 (0.11)
6	3.21	2.79	-0.42	271	3.17	2.74	-0.43	271	0.02 (0.17)
7	4.46	5.29	0.83	542	4.49	5.26	0.78	542	-0.19 (0.19)
8	3.51	3.53	0.02	443	3.44	3.45	0.01	443	-0.07 (0.18)
9	3.02	2.90	-0.12	870	2.97	2.91	-0.06	870	0.08 (0.09)
10	4.16	3.97	-0.18	400	4.15	4.05	-0.10	400	0.13 (0.24)
11	2.84	2.98	0.14	497	2.80	2.96	0.16	497	0.14 (0.16)
12	3.47	2.80	-0.67	467	3.48	2.75	-0.73	467	-0.50 (0.16)
13	2.05	2.24	0.18	3551	2.06	2.26	0.19	3551	-0.06 (0.04)
14	3.63	2.89	-0.74	2731	3.77	2.92	-0.85	2731	-0.07 (0.07)
15	0.85	0.85	0.00	519	0.84	0.85	0.01	519	-0.01 (0.04)
16	3.57	4.24	0.67	508	3.59	4.36	0.77	508	-0.51 (0.29)
17	3.74	4.00	0.25	293	3.72	3.94	0.22	293	-0.20 (0.13)
18	0.85	0.88	0.03	274	0.84	0.86	0.02	274	-0.02 (0.06)
19	4.15	4.36	0.21	982	4.24	4.49	0.25	982	-0.12 (0.17)
20	2.85	2.64	-0.22	396	2.68	2.67	-0.01	396	0.63 (0.24)

Note: DID is the difference-in-differences estimate between the Weighted TESS effect and the Unweighted TESS effect, as reported in the main body text of the paper. The standard error for the DID estimate is generated from a 5000-iteration permutation test.

Study 2 Treatment Group 1 Treatment Group Means, Effects, and Sample Sizes (MTurk Weighted and MTurk Unweighted)

	Control	Treatment	Effect	N	Control	Treatment	Effect	N
1	6.01	4.89	-1.12	1415	5.93	4.84	-1.10	1572
2	3.67	3.53	-0.14	1140	3.62	3.49	-0.13	1282
3	1.79	2.01	0.22	1323	1.79	1.88	0.09	1473
4	2.17	1.67	-0.51	885	2.02	1.62	-0.40	1003
5	2.28	1.77	-0.51	1350	2.29	1.73	-0.56	1519
6	3.23	2.85	-0.39	331	3.19	2.79	-0.40	369
7	4.63	5.28	0.65	441	4.86	5.50	0.64	485
8	3.67	3.99	0.32	358	3.74	3.68	-0.05	412
9	2.87	3.21	0.34	738	3.02	2.97	-0.05	840
10	3.82	3.76	-0.06	585	3.52	3.47	-0.06	670
11	2.75	2.93	0.19	595	2.60	2.87	0.28	682
12	3.73	2.29	-1.44	396	3.60	2.43	-1.17	454
13	2.22	2.28	0.06	1536	2.17	2.30	0.13	1740
14	3.79	3.03	-0.76	1822	3.78	2.97	-0.81	2045
15	0.88	0.80	-0.08	928	0.88	0.88	-0.00	1058
16	3.06	3.12	0.06	801	2.64	2.80	0.16	893
17	3.95	4.14	0.19	273	3.99	4.04	0.05	301
18	0.99	0.97	-0.01	319	0.92	0.93	0.01	346
19	3.56	3.04	-0.52	910	3.24	3.32	0.09	999
20	2.74	2.73	-0.02	532	2.89	3.31	0.41	587

Figure AF2: Control vs. Treatment Group 2



Note: Points are average treatment effects (difference between control and treatment group means), and bars represent one and two standard errors for the mean-difference. Figure is sorted by the magnitude of the effect size of the weighted TESS sample, which has been signed positive for all experiments. Weighting of the MTurk sample is based raking to the January 2014 Current Population Survey estimates of the U.S. household population, using a method analogous to that used by GfK to weight their samples. The larger error bars for the weighted MTurk sample are due to missingness on key demographic variables used in the weighting process; no imputation has been used.

Study 2 Treatment Group 2 Treatment Group Means, Effects, and Sample Sizes (TESS Weighted and MTurk Unweighted)

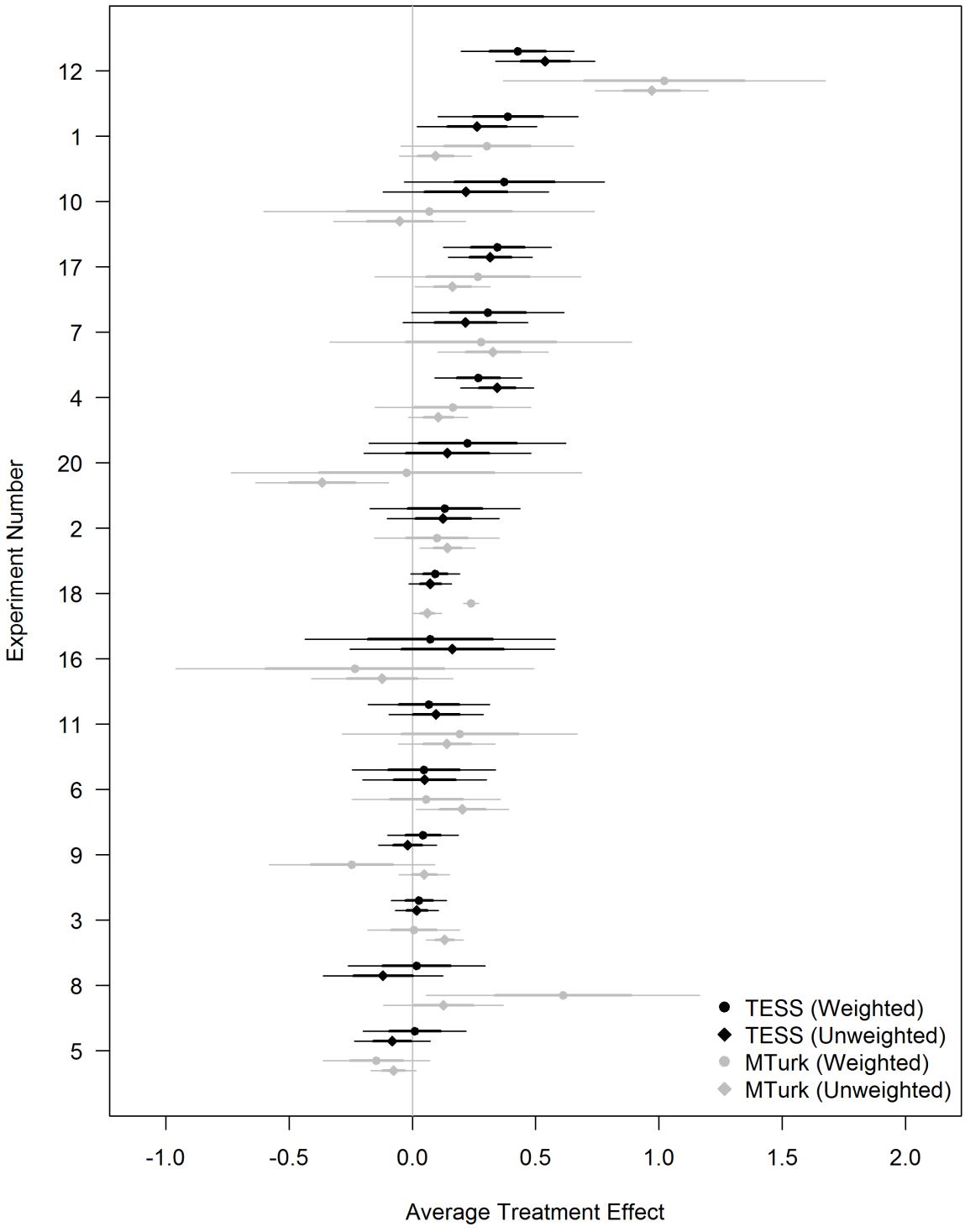
	Control	Treatment	Effect	N	Control	Treatment	Effect	N	DID (SE)
1	5.77	4.82	-0.96	611	5.93	4.74	-1.19	1549	-0.23 (0.16)
2	3.48	3.37	-0.12	402	3.62	3.63	0.01	1320	0.13 (0.16)
3	1.88	2.09	0.22	1576	1.79	2.01	0.22	1508	0.00 (0.07)
4	2.59	2.17	-0.42	790	2.02	1.73	-0.30	1028	0.13 (0.11)
5	2.36	1.77	-0.58	493	2.29	1.81	-0.48	1521	0.10 (0.11)
6	3.21	2.84	-0.37	256	3.19	2.99	-0.20	351	0.17 (0.17)
7	4.46	4.98	0.52	561	4.86	5.17	0.31	495	-0.21 (0.19)
8	3.51	3.51	-0.00	434	3.74	3.56	-0.18	404	-0.18 (0.18)
9	3.02	2.94	-0.08	855	3.02	3.02	0.00	874	0.08 (0.09)
10	4.16	4.34	0.19	389	3.52	3.42	-0.11	682	-0.30 (0.24)
11	2.84	2.91	0.07	507	2.60	2.73	0.14	659	0.07 (0.16)
12	3.47	3.23	-0.24	474	3.60	3.40	-0.20	461	0.05 (0.16)
13	2.05	--	--	1794	2.17	--	--	854	
14	3.63	--	--	1362	3.78	--	--	997	
15	0.85	--	--	260	0.88	--	--	536	
16	3.57	4.17	0.60	508	2.64	2.92	0.28	907	-0.31 (0.29)
17	3.74	3.65	-0.09	289	3.99	3.88	-0.11	300	-0.02 (0.13)
18	0.85	0.79	-0.06	290	0.92	0.87	-0.05	343	0.01 (0.06)
19	4.15	--	--	496	3.24	--	--	528	
20	2.85	2.86	0.01	403	2.89	2.94	0.05	606	0.04 (0.24)

Note: DID is the difference-in-differences estimate between the Weighted TESS effect and the Unweighted MTurk effect, as reported in the main body text of the paper. The standard error for the DID estimate is generated from a 5000-iteration permutation test.

Study 2 Treatment Group 2 Treatment Group Means, Effects, and Sample Sizes (MTurk Weighted and MTurk Unweighted)

	Control	Treatment	Effect	N	Control	Treatment	Effect	N
1	6.01	4.59	-1.42	1393	5.93	4.74	-1.19	1549
2	3.67	3.63	-0.04	1161	3.62	3.63	0.01	1320
3	1.79	2.02	0.23	1343	1.79	2.01	0.22	1508
4	2.17	1.83	-0.34	903	2.02	1.73	-0.30	1028
5	2.28	1.92	-0.37	1360	2.29	1.81	-0.48	1521
6	3.23	2.90	-0.33	310	3.19	2.99	-0.20	351
7	4.63	5.01	0.37	439	4.86	5.17	0.31	495
8	3.67	3.38	-0.29	363	3.74	3.56	-0.18	404
9	2.87	2.96	0.09	774	3.02	3.02	0.00	874
10	3.82	3.83	0.01	587	3.52	3.42	-0.11	682
11	2.75	2.74	-0.01	581	2.60	2.73	0.14	659
12	3.73	3.31	-0.42	403	3.60	3.40	-0.20	461
13	2.22	--	--	745	2.17	--	--	854
14	3.79	--	--	881	3.78	--	--	997
15	0.88	--	--	468	0.88	--	--	536
16	3.06	3.36	0.29	826	2.64	2.92	0.28	907
17	3.95	3.87	-0.08	265	3.99	3.88	-0.11	300
18	0.99	0.74	-0.25	314	0.92	0.87	-0.05	343
19	3.56	--	--	482	3.24	--	--	528
20	2.74	2.70	-0.04	551	2.89	2.94	0.05	606

Figure AF3: Treatment Group 2 vs. Treatment Group 1



Study 2 Treatment Group 2 versus Treatment Group 1 Means, Effects, and Sample Sizes (TESS Weighted and TESS Unweighted)

	Treat. 1	Treat 2.	Effect	N	Control	Treatment	Effect	N	DID (SE)
1	5.20	4.82	-0.39	612	4.84	4.74	-0.09	1549	0.29 (0.16)
2	3.23	3.37	0.13	371	3.49	3.63	0.14	1306	0.01 (0.16)
3	2.07	2.09	0.03	1584	1.88	2.01	0.13	1531	0.10 (0.07)
4	1.91	2.17	0.27	790	1.62	1.73	0.10	1051	-0.16 (0.11)
5	1.78	1.77	-0.01	501	1.73	1.81	0.08	1558	0.09 (0.11)
6	2.79	2.84	0.05	263	2.79	2.99	0.20	346	0.16 (0.17)
7	5.29	4.98	-0.31	549	5.50	5.17	-0.33	522	-0.02 (0.19)
8	3.53	3.51	-0.02	461	3.68	3.56	-0.13	414	-0.11 (0.18)
9	2.90	2.94	0.04	867	2.97	3.02	0.05	844	0.01 (0.09)
10	3.97	4.34	0.37	385	3.47	3.42	-0.05	708	-0.42 (0.24)
11	2.98	2.91	-0.07	524	2.87	2.73	-0.14	693	-0.07 (0.16)
12	2.80	3.23	0.43	475	2.43	3.40	0.97	411	0.54 (0.16)
13	2.24	--	--	1757	2.30	--	--	886	
14	2.89	--	--	1369	2.97	--	--	1048	
15	0.85	--	--	259	0.88	--	--	522	
16	4.24	4.17	-0.07	494	2.80	2.92	0.12	894	0.19 (0.29)
17	4.00	3.65	-0.34	278	4.04	3.88	-0.16	321	0.18 (0.13)
18	0.88	0.79	-0.09	280	0.93	0.87	-0.06	339	0.03 (0.06)
19	4.36	--	--	486	3.32	--	--	471	
20	2.64	2.86	0.22	407	3.31	2.94	-0.37	619	-0.59 (0.24)

Note: DID is the difference-in-differences estimate between the Weighted TESS effect and the Unweighted TESS effect. The standard error for the DID estimate is generated from a 5000-iteration permutation test.

Study 2 Treatment Group 2 versus Treatment Group 1 Means, Effects, and Sample Sizes (MTurk Weighted and MTurk Unweighted)

	Treat. 1	Treat. 2	Effect	N	Control	Treatment	Effect	N
1	4.89	4.59	-0.30	1408	4.84	4.74	-0.09	1549
2	3.53	3.63	0.10	1151	3.49	3.63	0.14	1306
3	2.01	2.02	0.01	1362	1.88	2.01	0.13	1531
4	1.67	1.83	0.16	942	1.62	1.73	0.10	1051
5	1.77	1.92	0.15	1398	1.73	1.81	0.08	1558
6	2.85	2.90	0.06	309	2.79	2.99	0.20	346
7	5.28	5.01	-0.28	468	5.50	5.17	-0.33	522
8	3.99	3.38	-0.61	371	3.68	3.56	-0.13	414
9	3.21	2.96	-0.25	752	2.97	3.02	0.05	844
10	3.76	3.83	0.07	628	3.47	3.42	-0.05	708
11	2.93	2.74	-0.19	602	2.87	2.73	-0.14	693
12	2.29	3.31	1.02	359	2.43	3.40	0.97	411
13	2.28	--	--	791	2.30	--	--	886
14	3.03	--	--	941	2.97	--	--	1048
15	0.80	--	--	460	0.88	--	--	522
16	3.12	3.36	0.23	805	2.80	2.92	0.12	894
17	4.14	3.87	-0.27	286	4.04	3.88	-0.16	321
18	0.97	0.74	-0.24	305	0.93	0.87	-0.06	339
19	3.04	--	--	428	3.32	--	--	471
20	2.73	2.70	-0.02	571	3.31	2.94	-0.37	619

Study 2 Demographics (Sex and Age)

	TESS Female (%)	MTurk Female (%)	TESS 18-29 (%)	MTurk 18-29 (%)	TESS 30-44 (%)	MTurk 30-44 (%)	TESS 45-59 (%)	MTurk 45-59 (%)	TESS 60+ (%)	MTurk 60+ (%)
1	51.02	49.50	12.08	48.77	22.78	34.05	33.22	13.55	31.92	3.63
2	47.27	50.12	17.94	50.26	22.42	34.52	26.30	11.72	33.33	3.51
3	53.34	49.50	14.71	48.77	22.31	34.05	33.43	13.55	29.54	3.63
4	49.08	50.12	14.09	50.26	22.38	34.52	31.76	11.72	31.76	3.51
5	48.73	49.50	15.75	48.77	26.32	34.05	31.02	13.55	26.91	3.63
6	52.51	49.50	16.71	48.77	24.72	34.05	29.97	13.55	28.60	3.63
7	50.05	48.82	17.82	48.89	23.40	34.69	29.97	12.99	28.80	3.44
8	48.55	48.82	17.27	48.89	24.41	34.69	29.87	12.99	28.46	3.44
9	51.74	50.12	17.79	50.26	25.46	34.52	29.82	11.72	26.93	3.51
10	50.25	52.37	15.57	44.58	23.35	36.44	31.30	14.81	29.78	4.17
11	50.78	52.37	16.93	44.58	25.84	36.44	27.65	14.81	29.59	4.17
12	48.55	50.12	15.43	50.26	22.98	34.52	29.67	11.72	31.92	3.51
13	48.92	50.12	15.12	50.26	22.95	34.52	32.25	11.72	29.68	3.51
14	51.88	49.50	14.17	48.77	23.15	34.05	30.83	13.55	31.84	3.63
15	47.79	52.37	16.51	44.58	24.76	36.44	30.52	14.81	28.21	4.17
16	50.34	46.44	16.88	50.85	29.54	34.81	29.93	10.64	23.65	3.70
17	51.41	46.44	15.87	50.85	32.60	34.81	41.26	10.64	10.27	3.70
18	48.83	46.44	15.14	50.85	23.74	34.81	30.49	10.64	30.63	3.70
19	52.71	48.82	16.97	48.89	24.01	34.69	30.42	12.99	28.61	3.44
20	49.83	46.44	16.75	50.85	23.30	34.81	28.69	10.64	31.26	3.70
CPS	51.79		21.39		25.38		26.94		26.29	

Study 2 Demographics (Race and Ethnicity)

	TESS White, Non-Hispanic (%)	MTurk White, Non-Hispanic (%)	TESS Black, Non-Hispanic (%)	MTurk Black, Non-Hispanic (%)	TESS Hispanic (%)	MTurk Hispanic (%)
1	75.59	82.74	6.45	6.22	11.27	1.12
2	74.91	81.74	8.85	5.52	9.21	1.37
3	72.52	82.74	8.91	6.22	10.91	1.12
4	75.93	81.74	8.14	5.52	8.49	1.37
5	75.54	82.74	9.78	6.22	9.69	1.12
6	74.95	82.74	9.52	6.22	9.61	1.12
7	-	81.51	-	6.76	-	1.44
8	75.42	81.51	7.84	6.76	9.78	1.44
9	73.26	81.74	9.61	5.52	10.57	1.37
10	77.16	81.70	6.94	6.39	9.81	1.68
11	77.00	81.70	7.49	6.39	10.72	1.68
12	76.51	81.74	8.43	5.52	8.84	1.37
13	74.89	81.74	9.23	5.52	10.09	1.37
14	76.75	82.74	8.28	6.22	7.99	1.12
15	71.79	81.70	9.98	6.39	11.13	1.68
16	77.43	81.92	9.42	5.64	7.16	1.52
17	74.61	81.92	9.46	5.64	9.29	1.52
18	72.85	81.92	10.73	5.64	9.88	1.52
19	75.27	81.51	8.84	6.76	8.94	1.44
20	72.97	81.92	8.37	5.64	10.03	1.52
CPS	79.07		12.34		0.13	

Study 2 Demographics (Education)

	TESS <HS (%)	MTurk <HS (%)	TESS HS (%)	MTurk HS (%)	TESS Some College (%)	MTurk Some College (%)	TESS Bachelor+ (%)	MTurk Bachelor+ (%)
1	11.84	1.16	33.80	9.20	24.98	42.60	29.39	47.04
2	10.30	1.25	28.00	9.79	28.24	43.62	33.45	45.34
3	13.10	1.16	30.82	9.20	28.15	42.60	27.93	47.04
4	7.44	1.25	29.68	9.79	29.48	43.62	33.40	45.34
5	13.41	1.16	34.15	9.20	25.24	42.60	27.20	47.04
6	11.27	1.16	30.59	9.20	28.74	42.60	29.40	47.04
7	10.62	1.26	32.04	11.12	28.08	43.18	29.25	44.44
8	10.57	1.26	28.72	11.12	27.93	43.18	32.78	44.44
9	11.30	1.25	32.91	9.79	26.35	43.62	29.44	45.34
10	8.80	1.14	29.78	10.58	27.92	40.01	33.50	48.28
11	12.02	1.14	32.69	10.58	28.04	40.01	27.26	48.28
12	8.54	1.25	30.22	9.79	27.48	43.62	33.77	45.34
13	7.89	1.25	29.23	9.79	29.52	43.62	33.35	45.34
14	7.10	1.16	27.18	9.20	30.71	42.60	35.01	47.04
15	9.40	1.14	32.05	10.58	28.98	40.01	29.56	48.28
16	14.03	1.07	30.03	9.54	28.66	42.07	27.28	47.32
17	5.25	1.07	23.49	9.54	29.37	42.07	41.89	47.32
18	10.95	1.07	29.92	9.54	28.00	42.07	31.13	47.32
19	13.18	1.26	32.40	11.12	26.08	43.18	28.34	44.44
20	10.95	1.07	32.17	9.54	27.78	42.07	29.10	47.32
CPS	12.41		49.06		9.23		29.30	