Advance Access publication July 14, 2016

Political Analysis (2016) 24:356–373
doi:10.1093/pan/mpw016

Survey Experiments with Google Consumer Surveys: Promise and Pitfalls for Academic Research in Social Science**

Lie Philip Santoso
Department of Political Science, Rice University, Houston, TX 77005
e-mail: ls42@rice.edu (corresponding author)

Robert Stein
Department of Political Science, Rice University, Houston, TX 77005
e-mail: stein@rice.edu

Randy Stevenson
Department of Political Science, Rice University, Houston, TX 77005
e-mail: randystevenson@rice.edu

Edited by Jonathan Katz

In this article, we evaluate the usefulness of Google Consumer Surveys (GCS) as a low-cost tool for doing rigorous social scientific work. We find that its relative strengths and weaknesses make it most useful to researchers who attempt to identify causality through randomization to treatment groups rather than selection on observables. This finding stems, in part, from the fact that the real cost advantage of GCS over other alternatives is limited to short surveys with a small number of questions. Based on our replication of four canonical social scientific experiments and one study of treatment heterogeneity, we find that the platform can be used effectively to achieve balance across treatment groups, explore treatment heterogeneity, include manipulation checks, and that the provided inferred demographics may be sufficiently sound for weighting and explorations of heterogeneity. Crucially, we successfully managed to replicate the usual directional finding in each experiment. Overall, GCS is likely to be a useful platform for survey experimentalists.

1 Introduction

Google Consumer Surveys (GCS) is a relative newcomer to the rapidly expanding market for online surveying and is dramatically less expensive than its competitors for short surveys (under ten questions). Such cost savings have made GCS popular among many organizations interested in quick, inexpensive surveying; however, it is not yet clear if GCS has a role to play in high-quality social scientific work. In this article, we examine the case for and against using GCS in rigorous social scientific work aimed at making accurate causal inferences. To preview, we argue that GCS's relative strengths and weaknesses make it most useful to social scientists who are attempting to identify causal effects through randomization to treatment groups rather than selection on observables (i.e., controlling for a long list of potential confounders). Thus, we focus our evaluation of the method on its performance as an experimental platform.[1] Specifically, we ask: How can assignment to treatment groups be randomized using GCS's survey construction and delivery mechanism? Does that assignment achieve balance across treatment groups? How reliably are treatments

Authors' note: Replication code and data are available at the Political Analysis Dataverse (Santoso, Stein, and Stevenson 2016) while the Supplementary materials for this article are available on the Political Analysis Web site. We would also like to thank Google Inc. for allowing us to ask some of the questions reported here free of charge.

[1] Such experiments might be simple pre-tests of different question-wording options or even more elaborate survey experiments (e.g., list experiments or framing and preference reversal experiments) in which the intention is to make inferences to a larger population and even to explore response heterogeneity.

**Several of the authors' corrections were not made in the originally published version of this article. This version now includes all of the authors' corrections. The publisher regrets this error.

actually delivered to respondents? Can researchers use such short surveys to reliably explore treatment heterogeneity and so explore causal mechanisms? And, can inferences drawn from such experiments reasonably be generalized beyond the sample?

Given the peculiarities of the GCS methodology, the answers to these questions are not obvious, and so researchers may either be hesitant to use the service or simply do so without any assurance that it is an appropriate vehicle for rigorous work. To combat this, we provide evidence-based answers to these questions by attempting to replicate a number of canonical survey experiments (including some that explore response heterogeneity).[2] In general, we find that random assignment to treatments is achievable using the GCS platform; that while GCS's sampling methodology heightens the problem of reliably delivering treatments to respondents, it is possible to combat this problem in straightforward ways; and that the ten-question survey is adequate for carefully exploring heterogeneity—though here we raise an important qualification concerning the way GCS collects and reports demographic information about respondents.

In the rest of this article, we elaborate briefly on GCS's cost structure and methodology, explain why we think these features make GCS more useful to researchers attempting to identify causal effects using random assignment (survey experiments) rather than selection on observables, point out a number of challenges (and solutions) to using the platform to run survey experiments, and evaluate the success of the platform in replicating a number of canonical social science experiments. Besides our general conclusion that GCS will often be useful for conducting survey experiments, we make a number of other useful discoveries: we provide some initial evidence that GCS's unique methodology leads to mildly attenuated estimates of experimental effects; that the size of this attenuation varies in predictable ways with treatment question complexity; and that it can be detected and efficiently ameliorated with minimally intrusive (and minimally costly) manipulation checks. In addition, we provide a comparison of the usefulness of respondents' "implied" demographic information (provided for free by Google) to their self-reported demographics for the purposes of exploring effect heterogeneity and achieving representative samples either by stratified sampling (with its implications for achieving treatment group balance) or post-stratification weighting. Our conclusions, while generally positive (at least for gender and age), are qualified: they show that both implied and self-reported demographics lead, in our examples, to directionally consistent conclusions of broadly similar magnitudes, but that this correspondence is not so high that it can put to rest doubts caused by the relatively high degree of noise in implied demographics, especially for variables beyond gender and age.

2 The Cost Structure of GCS

Figure 1 demonstrates that GCS, as currently priced, is markedly less expensive than its leading competitors for short surveys: Conducting a ten-question survey using GCS is a third the price of both Toluna QuickSurveys and SSI's least expensive product (their DIY surveys), less than a quarter of the price of SSI's standard 5-min survey, and many times less than YouGov or GFK's omnibus surveys, which are by far their least expensive options.[3]

[2]Some might argue that it is too early for such a study, since few published studies have so far used the platform. We disagree. The cost benefits will almost certainly attract some researchers, so it seems better to get ahead of these questions rather than to retrospectively criticize early adopters.

[3]GCS only charges 10 cents per respondent for the first question, but charges $1.00 per respondent for two to ten questions. Toluna charges 35 cents per question per respondent (up to a maximum of twenty-five questions). The most affordable traditional panel service that we could find is YouGov's Omnibus Survey of U.S. adults, in which a researcher's questions are placed on a regularly scheduled omnibus survey along with those of other clients. One must buy a minimum of 1000 respondents and there is a $300 setup fee. After that, the cost is $500 per question, so a marginal cost of $0.80 per respondent for the first question, $0.50 for each subsequent one. Other omnibus services are somewhat more expensive (e.g., GFK's KnowledgePanel omnibus is $750 per question). Non-omnibus solutions are significantly more expensive. Of course, prices for all these surveys (including GCS) are subject to change, but for GCS's cost advantage relative to its competitors (for a ten-question survey) to disappear, it would have to raise (or its closest competitors lower) prices by 300%.

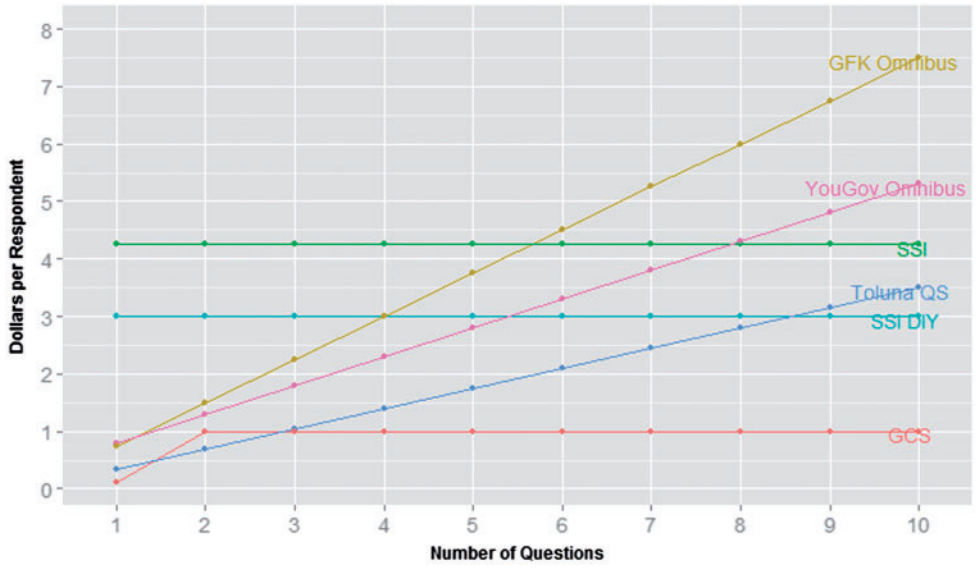

Fig. 1 Relative costs of short surveys.

3 The Promise of GCS as a Platform for Survey Experimentation

We assume that most social scientists are interested in using their data to learn about the causal connections between some set of treatment variables and some set of outcomes. As such, each must adopt some identification strategy that allows them to argue that the associations they estimate from the data have a causal interpretation (Morgan and Winship 2007). By far the most common identification strategy is selection on observables, which requires that one control for all variables that connect the treatment and outcome variables in ways that confound the relationship (Pearl 2011). This can often mean controlling for a large number of variables, and so the financial incentives of GCS may make it less likely to be useful for researchers using this strategy. Alternatively, if one uses random assignment of the treatment as one's identification strategy, then the need to control for a long list of potential confounders can often be avoided. That said, the agenda of most survey researchers surely goes beyond simply delivering a treatment and examining its overall effect. For example, treatment heterogeneity is an important concern of many survey experimenters, and exploring it is often the key to accessing causal mechanisms. In general, this will require fielding surveys that are longer than what is needed to deliver a treatment.

Thus, the first question we raise about the suitability of GCS as a vehicle for survey experimentation is somewhat mundane, but practically important. Is a survey limited to ten questions long enough to be useful to a large number of survey experimentalists?[4] It is easy to point to examples of published work in which the work could have been accomplished with such a survey (e.g., Kane, Craig, and Wald 2004; Streb et al. 2008; Holbrook and Krosnick 2010; Green and Kern 2012; Huber and Paris 2013). Indeed, we replicate some of these studies below. However, a more systematic analysis is possible by looking at the data available from an NSF-supported effort to make survey experimentation more widely available to researchers, the Time Sharing Experiments in the Social Sciences (TESS) initiative. First, in 2012 TESS initiated a special program for short experimental studies, where they defined a "short study" as "a survey experiment that is no more than three items (i.e., questions/stimuli) in length (not including the standard demographic variables that are delivered as part of every TESS study)." Thus far, more than 20 studies in this program have been funded (data on unfunded proposals for the short program were not available). More generally, when we look across all 592 proposed survey experiments submitted to TESS's regular

[4]Much of the attention to GCS's cost structure has focused on the ultra-cheap single-question survey (which is only $0.10 per respondent). However, since few social scientists (even experimentalists) will find a one-question survey useful, we focus on the two to ten-question formats.

program since April 2011, approximately 32% proposed ten questions or fewer (for full distribution of the number of question units requested, see Section 1 of the Supplementary Materials).[5] This evidence makes it clear that survey experimentalists can make use of short surveys. Thus, if GCS can meet the standards of rigorous social scientific work, it has the potential to provide scholars without access to large research budgets the opportunity to make progress in their research that would not otherwise be possible. Even for well-funded scholars, GCS could provide an affordable way to do a wider variety of short pretests before committing to a longer instrument.

That said, GCS's ability to deliver inexpensive, short surveys comes primarily from its reliance on a new sampling methodology that dispenses with the need to recruit and maintain a large panel of potential respondents. Further, as we explain below, survey experimentalists interested in using GCS should be concerned about the consequences of this platform for four fundamental questions:

1. How and to what extent can random assignment of respondents to treatment groups be achieved?
2. To what extent can treatments be reliably delivered to respondents?
3. How and to what extent can GCS be used to explore treatment effect heterogeneity?
4. To what extent are GCS samples representative and/or what can be done to make them representative?

Our answers to each of these rely on data from our replications of several canonical social science experiments and, after a brief description of GCS's methodology, we address each in turn.

4 GCS's Methodology

Figure 2 illustrates how GCS works. When individuals are browsing the Internet and attempt to access content provided by one of GCS's commercial partners (e.g., a news article), they may be confronted with a "survey wall." This wall shows the first question of the survey and gives the option to take the survey, ask for a different survey, or simply give up on getting to the desired content. GCS initially randomly assigns surveys to respondents, but subsequently uses information about the demographic characteristics of respondents to dynamically adjust the sample to target the age, gender, and geographic characteristics (state or region) of the Internet using population based on the most recent CPS *Internet Users Supplement*.[6] For surveys with a high level of non-response (which include all the surveys we conducted), this targeting process may not produce proportions of respondents that match the *CPS*, and so post-stratification weights (based on age, gender, and region when compared with the CPS *Internet Users Supplement*) are calculated and provided with the individual-level data.

Unlike other survey platforms, where respondents are part of an online panel, GCS has not collected extensive demographic and background information on respondents at an earlier date. Instead, GCS relies on "inferred demographics." This inferred information includes income, age, gender, parental status, and various variables based on location—which are derived either from the user's browsing history or her IP address.[7] Of course, not all of this information is available for all users, since it is possible for users to make their browsing history private and so prevent Google from inferring their demographic information. This is potentially (but not necessarily) problematic for the representativeness of the GCS sample, the accuracy of the post-stratification weights, and the balance of treatment groups in an experiment.

[5]Technically, the proposals asked for ten or fewer "units." These units are roughly equivalent to questions but are generalized to accommodate other sorts of non-question stimuli. A full description of what a unit is can be found at http://www.tessexperiments.org/limits.html.

[6]This demographic information is inferred from users' browsing history. Below, we evaluate the accuracy of these inferences and discuss their implications.

[7]Gender and age group can be inferred from the types of pages users visited in the Google Display network using the DoubleClick Cookie. The DoubleClick Cookie is a way for Google to provide ads that are more relevant to the respondents by using a unique identifier called a cookie ID to record users' most recent searches, previous interactions with advertisers, ads, and visits to advertisers' Web sites. Likewise, Google uses users, IP addresses to locate their nearest city and estimate the income and urban density.

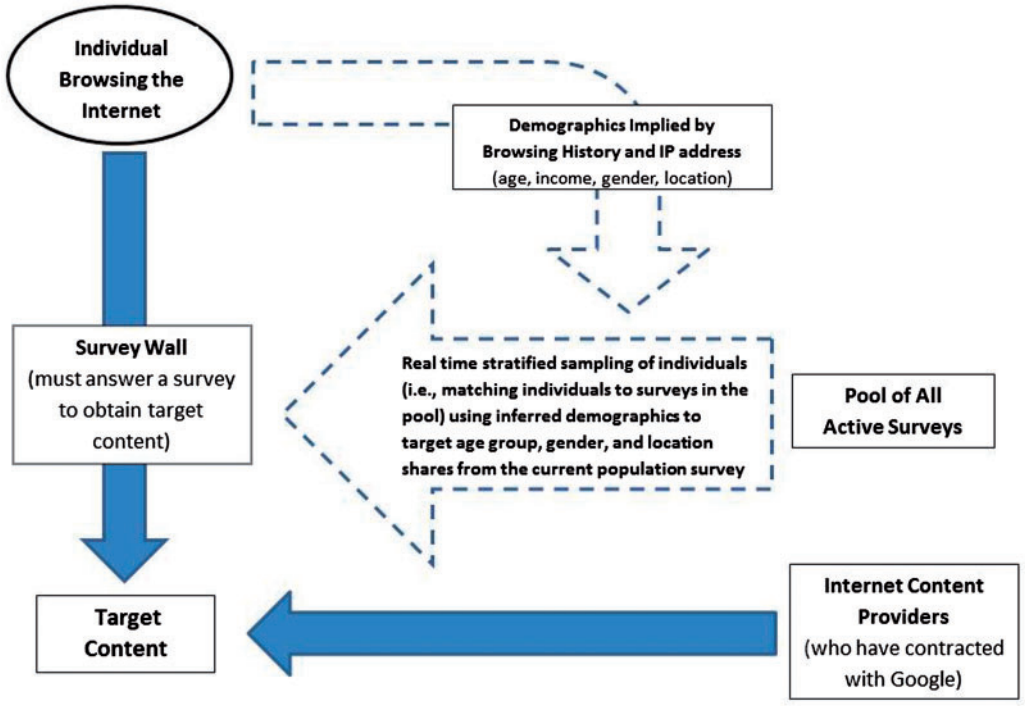

Fig. 2 Google consumer surveys "survey wall" methodology.

In Section 2 of the Supplementary Materials, we include an original analysis of how well GCS's inferred demographics compare to self-reported values of these same variables. The conclusion of that analysis is mixed. Clearly, the use of inferred demographics introduces noise into the data, but the amount of noise appears to be relatively modest for variables like age and gender.[8] Specifically, we found—in line with a similar analysis by Pew in 2012—that 75–80% of respondents reported the same gender as that inferred by Google; about 50% reported the same age category (with 75;–80% reporting the same or an adjacent category); and, about 20% of respondents reported income that matched Google's inferred income (with only about 55% within one income category).[9] In addition, Google could not infer anything about age or gender for 28% and 25% of our respondents, respectively.[10] Finally, the analyses in Supplementary Tables A2.1–A2.3 show very clearly that differences between self-reports and inferred demographics, as well as patterns of missingness in inferred demographics, are unrelated to non-response. Thus, we see no *prima facie* case for concluding that GCS's inferred demographics (at least on age, gender, and geography) are so out of line with other information that they should be categorically rejected. In the analyses below, we use them but also examine the consequences of doing so (compared with using self-reports) both for post-stratification weighting and for exploration of effect heterogeneity.[11]

[8]Further, there is no way to know from our analysis whether the implied demographics are any more inaccurate than self-reports.

[9]Pew's 2012 study did not examine income.

[10]This is based on our ten-question survey described below. Rates are similar for our other surveys and also match the rates reported in Tanenbaum et al. (2013). Inference on location variables is almost universal and for income is unknown in only 1% of cases. This appears to be because Google uses geographic information and corresponding geographic income distribution data (rather than browsing history) to infer income. That said, the actual inference process is somewhat of a black-box and is not as simple as reporting the average income for the respondent's geography since for 19% of cases, GCS reports income as "prefers not to say"—which suggests the inferred category for any individual is some assignment of census category data for the geography to the individual. Our efforts to clarify how income is inferred with Google were unsuccessful.

[11]Even if one is worried that the inaccuracy of GCS' inferred demographics will compromise post-sample adjustments or hinder exploration of response heterogeneity, one can always simply replacing them with self-reported ones—given that the GCS 'cost structure likely provides sufficient "room" to affordably include them in a ten-question survey.

5 The Challenges of GCS as a Platform for Survey Experimentation

All of this suggests that GCS may be particularly useful to survey experimenters. However, there are two challenges that one might face when implementing survey experiments in this platform: (1) assigning treatments randomly to achieve balance between treatment groups; and (2) assuring that respondents actually receive the treatments to which they were assigned. To evaluate how well GCS meets these challenges, we attempt to replicate several canonical social science experiments. First, we conducted two versions of the classic welfare framing experiment first published by Rasinski (1989): one in a one-question survey, and one in a ten-question survey. Second, we used three different one-question surveys to replicate Tversky and Kahneman's (1981) famous "Asian Disease" experiment as well as two list experiments measuring support for a female president and for immigration to the United States. The details of each experiment are provided in Section 3 of the Supplementary Materials, so we provide only a brief description here.

6 Benchmark Political Science Experiments

6.1 *Welfare Experiment*

This is a replication of Rasinski (1989), Green and Kern (2012), Berinsky et al. (2012), and Huber and Paris (2013), each of which asked respondents whether too much, too little, or about the right amount was being spent on either "welfare" or "assistance to the poor." All of these studies have shown that support for government spending depends greatly on whether the program is called "assistance to the poor" or "welfare."

6.2 *Asian Disease Experiment*

This is a replication of a frequently employed, and relatively complex, framing experiment known as the "Asian Disease Problem." This experiment was first reported in Tversky and Kahneman (1981) using a student sample, but has been replicated many times using different subject pools (Takemura 1994; Kuhberger 1995; Jou, Shanteau, and Harris 1996; Druckman 2001; Berinsky et al. 2012). In this experiment, respondents are given a hypothetical scenario in which a disease threatens a population and they are asked which of the two solutions they would choose.[12] The experiment consistently reveals large preference reversals for different frames (which are equivalent in expected value).

6.3 *List Experiments*

Next, we replicated two different "list experiments" that are designed to allow researchers to get more accurate responses to questions in which the respondent has an incentive to hide their true response. The first replication is the list experiment conducted by Janus (2010), who showed that there is much more resistance to immigration than individual respondents admit to in standard survey questions.[13] The second is a list experiment by Streb et al. (2008), who examine attitudes about a female president and again find much more resistance to the idea in the list experiment than in self-reports.

7 Random Assignment of Treatment and Balance

Unlike many other online survey services, GCS does not provide a facility for assigning a respondent taking a given survey to different versions of a question. Thus, to achieve differential assignment

[12]Since we cannot faithfully implement the exact wording of the experiment due to the word limitation imposed by GCS, we capture the necessary text in a picture format, which is then inserted into the question as a screen image.

[13]Again, the word limitation that GCS imposes forces us to shorten the question that Janus used in his experiment. While he asked: "Now I am going to read you three/four things that sometimes people oppose or are against. After I read all three/four, just tell me HOW MANY of them you oppose. I don't want to know which ones, just HOW MANY," our question was "How many of these items are you opposed to?"

one must first use GCS's survey construction tools to construct multiple versions of the survey (with differences only in the questions that define the treatment groups). Next, these surveys are served (simultaneously) to the pool of active respondents, making sure that the same person is prevented from taking both (which is possible to do in GCS). For example, if a researcher wanted to have 1000 respondents assigned to one of two questions, he or she would submit two surveys that differ only in that question and request that 500 respondents take each one.[14] Whether this assignment is random depends, of course, on how respondents are matched to surveys and how, once matched, they select into the survey. In GCS, there is (at any given time) an active pool of surveys in need of respondents. When a potential respondent encounters the survey wall, a survey is selected from this pool and served to the respondent. This selection is initially random; however, once some responses have come in, surveys that are not hitting demographic targets are routed to respondents with needed characteristics (e.g., females). As long as this matching process does not depend in any way on the differences between the treatment and control surveys, there is no reason to think that it would result in imbalance across the treatment and control samples. However, once matched with a survey, individuals can choose to take the survey or not (or drop out of it or not)—a choice that may well depend on the difference between the treatment and control surveys.[15]

In order to test if these assignment mechanisms achieve balance in treatment and control groups, we can examine whether balance was achieved in our replications for the inferred demographics provided.[16] In addition, in our ten-question survey we asked respondents to directly report their demographics and political attitudes (party identification and ideology), and so we can also present balance results for those variables.[17]

Overall, we found that GCS produced very tight balance across the treatment and control groups on both the inferred and self-reported demographic characteristics, as well as on attitudinal variables (Table 1).[18] Furthermore, as we show in Supplementary Table A4.1, we also achieve balance on inferred demographics for those respondents who choose to opt out of the survey.

8 The Potentially Coercive Nature of the Survey Wall Environment

As we have discussed above, one of the most significant differences between GCS and other Internet-based surveys is that it relies on the survey wall rather than a traditional opt-in panel to sample respondents. Specifically, this environment interrupts respondents' browsing activities with survey questions and "coerces" them to provide a response to survey questions before allowing them to proceed to their target content. Because such respondents have not agreed beforehand to cooperate with survey researchers, we should be quite sensitive to the extent in which they will not

[14]Since a treatment condition is delivered by making a separate survey for each treatment group, there is theoretically no limitation on the complexity of the design. For example, if one wanted a 2^3 factorial design, there would be eight treatment groups, each of which would deliver a survey with the appropriate combination of the three factors (which might be delivered in one question or three, depending on the design). That said, there is obviously a practical limitation imposed by the time it takes to construct and activate many surveys. For example, GCS would likely not be practically useful for implementing a conjoint design.

[15]The respondent can see the first question of the survey when they make the decision and so, if this is the critical treatment question (as it would be in a one-question survey), this choice to participate may well differ across treatment and control. In the case of our surveys, however, this did not happen (which we can know because GCS provided data on those who did not agree to participate and there was balance in both the responders and non-responders). Of course, in longer surveys one could guard against even the possibility of a potential first-question problem by choosing an innocuous first question.

[16]In Section 4 of the Supplementary Materials, we also test whether there is a balance in the inferred demographics for respondents who chose to opt in and out of surveys across control and treatment groups for GCS's sample.

[17]Senn (1994), Imai, King, and Stuart (2008), and Mutz and Pemantle (2013) warned against balance tests when one is sure the randomization procedure is sound. In our case, however, the veracity of the randomization procedure is exactly what we want to examine, and so this fits the narrow set of situations in which these authors argue such a test will be useful.

[18]We did formal balance tests for these variables across treatment and control groups using appropriate difference in proportions tests and found that the differences between the two groups were not statistically significant more often than would be expected by chance ($p > 0.05$).

https://doi.org/10.1093/pan/mpw016 Published online by Cambridge University Press

Table 1 Experiments' Sample

	Framing (1-qn)		*Framing (10-qn)*		*Asian disease*		*List (female president)*		*List (immigration)*			*Framing (10-qn)*	
Variables (Inferred)	C	T	C	T	C	T	C	T	C	T	*Variables (self-reported)*	C	T
Female	36	36	27	28	35	35	38	37	32	31	Female	39	42
Male	56	55	47	48	56	56	54	55	52	53	Male	61	58
18–24	14	16	17	18	17	17	15	17	12	14	18–24	20	21
25–34	16	15	21	23	15	14	18	15	13	14	25–34	19	20
35–44	14	14	14	15	14	14	13	13	12	13	35–44	14	15
45–54	15	15	14	14	14	15	17	17	15	16	45–54	19	17
55–64	20	19	22	18	18	18	15	16	16	16	55–64	18	15
65+	11	12	12	11	11	11	11	11	13	10	65+	11	11
$0–$24,999	8	7	8	9	6	6	6	7	7	7	$0–$24,999	22	24
$25K–$49,999	61	59	54	59	58	61	55	58	57	58	$25K–$49,999	19	21
$50K–$74,999	25	27	28	22	28	26	32	28	29	27	$50K–$74,999	21	17
$75K–$99,999	5	6	6	7	5	6	5	6	6	6	$75K–$99,999	13	12
$100,000+	2	1	4	3	2	1	1	1	1	1	$100,000+	25	26
Midwest	27	25	32	31	27	29	28	30	31	31	Less than High School	4	3
Northeast	16	17	21	22	18	16	17	18	18	18	Some High School	4	5
South	34	33	23	30	30	30	30	28	24	26	High School/GED	11	15
West	24	25	25	18	25	25	25	25	25	26	Some College	31	31
											Four-year College Degree	25	24
											Postgraduate	25	22
											Republican	31	32
											Independent	35	36
											Democrat	33	31
											Conservative	31	31
											Moderate	42	43
											Liberal	27	26

Note: Inferred demographics with unknown values for framing (10 qn) experiment are dropped.

be cooperative in answering questions.[19] One way this non-cooperation may be instantiated is for respondents to not respond, either by navigating away from the page or requesting another question (see the example question screen in Section 5 of the Supplementary Materials). Indeed, we do find that the average rates of non-response in our experiments (calculated based on everyone who was given the opportunity to participate and did not or who began the survey but did not complete it) is about 70–75% for our one-question surveys and about 94% for the longer ten-question survey (in Table A6.1 in Section 6 of the Supplementary Materials, we give the question by question retention rate). These rates are bigger than the typical rate of non-response that Shih and Fan (2008) report for web surveys (66%) and are certainly consistent with what we would expect from the incentives for non-cooperation in the survey wall environment.

Another way that respondents can fail to cooperate with the survey is to take it, but refuse to substantively engage. The advantage to the respondent of this strategy is that she can quickly finish the survey and thus can proceed to the desired content. One form of this behavior is simply to choose the "don't know" or "prefer not to say" option for most questions; and, as we saw for non-response, GCS appears to produce a larger number of such responses than other web-based platforms. Overall, we find the rate of "Don't Know" responses is about 25–42% among respondents who choose to answer the question (see the 2nd row of Supplementary Table A7.1 in Section 7 of the Supplementary Materials)—compared with the typical rate of 8.2% found in other web-based modes (e.g., DeRouvray and Couper 2002). Despite such large rates of both non-response and "Don't Know" responses, it is important to point out that the treatment assignment mechanism achieves balance across treatment and control groups for the remaining responders, the non-responders, and those who say "Don't Know) (see Section 7 of the Supplementary Materials for the full result)—which suggests that even this high level of overall non-cooperation is inconsequential for the goal of estimating the causal effect within the sample (we discuss the issue of making generalization outside of the sample below).

A more pernicious form of non-cooperation for experimental studies is not non-response or choosing the "Don't Know" option, but rather choosing substantive answer options without reading the questions. When respondents do this with respect to the treatment questions, it is equivalent to the treatment not being delivered. As Berinsky et al. (2014) have explained in detail, this kind of non-cooperation in survey experiments will tend to attenuate experimental treatment effects because in such situations treatment and control respondents essentially get the same (non-) treatment.

To examine this possibility—as well as the more general issue of whether GCS can replicate experimental benchmarks—Tables 2–4 present the results of our replications of four canonical social science experiments (along with examples of the estimated treatment effects from previous studies). Taken together, these results demonstrate that it is possible to replicate the direction and approximate size of the usually observed treatment effects using GCS. Importantly, however, in each case the estimates using GCS are smaller than all (or most) of the available benchmarks. Further, the size of this possible attenuation ranges from small (in the list experiments) to moderate (in the Asian Disease experiment) in rough correspondence to the complexity of the treatment question—reinforcing the possibility that the cause of the observed attenuation is respondent inattention to the treatment questions. Below, we explain these results in somewhat more detail and then explore the sources of the possible attenuation by including an explicit manipulation check in one of the experiments.

For the classic welfare framing experiment, Rasinski (1989) reported that in each year from 1984 to 1986, 20–25% of the (face-to-face interview) respondents from the General Social Surveys (GSSs) said that too little was being spent on welfare, while 63–65% said that too little was being spent on assistance to the poor. Berinsky et al. (2012) found a similar gap (38%) using

[19]Google is well aware of the possibility that respondents in the survey wall environment will lack motivation and so attempt to constrain the kinds of questions researchers may ask—ensuring that questions are short, simple, and (to some extent) uncontroversial. This can create challenges for researchers trying to build a survey experiment with a complex set of treatments. In Section 8 of the Supplementary Materials, we discuss these format limitation and some practical ways to work within them and some creative ways to circumvent them (when prudent to do so).

Table 2 Replication of welfare experiment

	GSS face to face 1986–2010; Rasinski (1989)	*Berinsky et al. MTurk (2012)*	*Paris and Huber (2013) MTurk and YouGov samples*	*GCS (1-question survey)*	*GCS (10-question survey)*
Assistance to the poor (%)	63–65	55	31–41	49	49
Welfare (%)	20–25	17	16–22	26	25
Difference (%)	35–40	38	15–19	23	24

Table 3 Asian disease experiments

	Tversky and Kahneman (1981) student sample	*Takemura (1994) student sample*	*Kuhberger (1995) student sample*	*Jou, Shanteau, and Harris (1996) student sample*	*Druckman (2001) student sample*	*Berinsky et al. (2012) MTurk sample*	*GCS*
Lives saved condition (%)	72%	80%	54%	65%	68%	74%	58%
Lives lost condition (%)	22%	20%	22%	20%	23%	38%	33%
Difference (%)	50%	60%	32%	45%	45%	36%	25%

Table 4 List experiments

	How many of these items do you oppose?			
	Sensitive item: "Cutting off immigration to the United States"		*Sensitive item: "A woman serving as president"*	
	Janus (2010) TESS Internet sample	GCS	Streb et al. (2008) Telephone sample	GCS
Mean number of Baseline items	1.77	2.35	2.16	2.3
Mean number of Treatment items	2.16	2.69	2.42	2.53
Difference	39% "Opposed to immigration"	34%	26% "Woman president makes them upset"	23%

MTurk samples. We replicated this experiment twice—once in a one-question survey and once as the only treatment question in a ten-question survey.[20] Table 2 reports the results. We found that the direction of the framing effect was the same as in prior studies (i.e., more respondents chose "too little" in the assistance to the poor version than in the welfare version) and that the size of the effect was only 23% for the one-question survey and 24% for the ten-question version of this experiment (both statistically different from zero at $p < 0.001$). On the one hand, these estimates are within the range of previous findings; on the other hand, of the previous results summarized in Table 2, these are only larger than two (those from the Paris and Huber studies using MTurk and YouGov) and are significantly smaller than the 35–40-percentage-point difference in the person

[20]It was the third question, after self-reported age and gender.

GSS samples and Berinsky's MTurk's samples.[21] Thus, a prudent conclusion may be that GCS may evidence some attenuation for this experiment, but if so it is relatively modest. That said, below we explore this question more carefully by introducing a manipulation check.

In their famous Asian Disease experiment (see Section 3.2 in the Supplementary Materials for detailed description), Tversky and Kahneman tested subjects, preference for risky versus certain choices and reported that 72% of their subjects picked the certain choice when the consequences were framed positively, but only 22% picked that choice when the framing was negative (a treatment effect of 50%). Subsequent replications have confirmed that result, with estimated treatment effects ranging from 32% to 60% (Table 3). In our replication using GCS, we again found the expected preference reversal: 58% of our respondents pick the certain choice in the positive framing, while only 33% choose the certain choice in the negative framing. However, the size of our estimated treatment effect (25%, $p < 0.001$) is lower than all the previously published estimates. Thus, like in the welfare experiment (but much more clearly in this case), estimates based on GCS's sample may be attenuated relative to previous studies that mostly used in-person samples. Given our suspicion that this potential attenuation results from respondents' inattention to treatments, it is also relevant that of all our experiments, this is the one with the most evidence of attenuation and is also the one with the most complicated treatment question.

Finally, we also replicated two list experiments (see Section 3.3 in the Supplementary Materials for detailed description) and found that 23% of the GCS's respondents oppose having a female president and 34% oppose immigration to the United States (Table 4). Similar to the previous experiments, we find a smaller effect using GCS than the benchmarks, though in this case the differences are much smaller (but statistically significant at $p < 0.01$).

Overall, the treatment effects we estimated using GCS were always in the expected direction and not wildly different in size from benchmarks. That said, they were almost always smaller than prior estimates. Further, this attenuation was most pronounced in the Asian Disease experiment, which was by far the most complicated treatment question.

9 A Closer Look at the Sources of Treatment Effect Attenuation in GCS

In order to examine whether the modestly attenuated treatment effects observed in each of our replications is really due to respondents not getting the intended manipulation, we included a manipulation check in the ten-question survey in which we replicated the welfare framing experiment. Specifically, we included a question immediately following the treatment question that asked respondents, "which of the following policy areas did we ask about in the previous question?" This was followed by a list of seven response options (e.g., Education, Homeland Security, and Agriculture), including "Welfare" or "Assistance to the Poor" as relevant; 75.3% of those who got the Welfare treatment and 75.6% who got the "Assistance to the Poor" treatment passed this check.

To examine whether the attenuation in our estimate of the treatment effect for this experiment can be attributed to inattention to the treatment question, we re-estimated the effect for those who passed the check and those who did not. The results are given in the first three rows of Table 5.

Clearly, the direction of the results is consistent with the hypothesized mechanism driving attenuation. The estimated effect for the respondents who failed the manipulation check is dramatically smaller (12%) than our original estimate (24%). However, since this is a relatively small group (only 18% of final sample), the impact on the overall estimate is modest. Specifically, if we confine the same to only those who passed the manipulation check, our estimate of the treatment effect increases from 24% to 28% (a statistically significant difference at $p < 0.05$). Clearly, then,

[21]Green and Kern (2012) also analyzed the 1985–2010 GSS data using a different estimation method but only reported results on the gap in the framing effect for respondents choosing "too much." These estimates ranged between 27% and 51% for different years. Using GCS, we calculated the corresponding effects and found that for the one-question survey the gap was 30% and for the ten-question survey it was 22%. Thus, similar to the analyses focusing on the "too little" answer, these results are roughly similar to previous work on the low side—suggesting that they may be mildly attenuated.

Table 5 Comparison of treatment effect across respondents

	% said "too little"		
	Control	*Treatment*	*Treatment effect*
All respondents (770)	49	25	24
Pass MC (634)	52	24	28
Fail MC (136)	37	25	12
>4 s (683)	49	23	26
<4 s (87)	51	40	11

Note: Number of respondents is in parentheses.

inattention is playing some role in attenuating this treatment effect. Indeed, the result of 28% is approaching the range of results from previous in-person samples (35–40%).[22]

Finally, GCS also provides, at no additional charge, response times for each question. Comparing the answers to our manipulation check to the response times for the treatment question, we found that the average time spent on the treatment question for those passing the manipulation check was almost 9 s, while for those not passing, it was 4 s. This result opens up the possibility of economizing on questions by replacing the manipulation check with the response time for the treatment question. Thus, in the last two rows of Table 5, we also show that one gets a similar (though somewhat smaller) difference in estimated treatment effects using response times as when using the explicit manipulation check. Indeed, if one only included respondents who took longer than 4 s to answer the treatment question, this would include 86% of those that answered the manipulation check correctly and only 14% of those who did not—a degree of noise in the manipulation check that may in some cases be acceptable in exchange for not using up a question.

This result clearly suggests that we should also use response times for treatment questions as pseudo-manipulation checks for the other experiments we replicated but for which we could not include an explicit manipulation check (i.e., our one-question versions of the welfare and Asian Disease experiments, both of which showed some evidence of attenuation). We did this and found that when we focus only on those respondents who spent sufficient time with the treatment to have read it, the treatment effect gets bigger in each case—increasing by 3% in the Asian Disease experiment and 3% in the one-question version of the welfare experiment.[23]

10 Exploring Response Heterogeneity

Exploring treatment effect heterogeneity is an important part of most experimentalists' research because it is one of the main ways experimental researchers try to examine the causal mechanisms underlying their results. To examine the utility of GCS for this kind of work, we replicate Green and Kern's (2012) reanalysis of the welfare framing experiment described above, which focused on exploring treatment effect heterogeneity with respect to a set of attitudinal and demographic variables.[24] This is useful both because it demonstrates that such studies are possible in the ten-question format of the GCS survey, but also because we can explore whether such studies of heterogeneity can usefully be done using the inferred demographics supplied by Google. We examine each of these questions below.

In their replication, Green and Kern find that Democrats, liberals, and those with more favorable attitudes about blacks are much less likely to change their opinions when the frame of the question is changed than are independents, moderates, Republicans, conservatives, and those with less favorable attitudes toward blacks. Table 6 provides our estimates of treatment effects for each

[22]Of course this result already exceeds Huber and Paris's (2013) 15–19% estimates (which they get using other web-based platforms).

[23]In each case, we chose a cutoff by examining the full distribution of times respondents spent with the question.

[24]Note that, unlike the results for the welfare experiment reported above, the dependent variable in their studies is whether respondents choose "too much." Thus, the sizes of the effects discussed here are not directly comparable to those in Table 2.

Table 6 Heterogeneous treatment effect

	Treatment effect	*Standard error*
Democrat	0.08	0.04
Independent	0.24	0.04
Republican	0.27	0.05
Liberal	0.03	0.04
Moderate	0.23	0.04
Conservative	0.33	0.05
Least favorable to Blacks	0.17	0.06
Moderately favorable to Blacks	0.24	0.04
Most favorable to Blacks	0.25	0.04

Note: The treatment effect here refers to the change in probability of respondents choosing "too much".

Table 7 Heterogeneous treatment effect (inferred versus self-reported)

Variables	*Treatment effect (inferred)*	*Standard error*	*Treatment effect (self-reported)*	*Standard error*
Male	0.13	0.05	0.2	0.05
Female	0.38	0.06	0.26	0.05
18–24	0.13	0.06	0.17	0.06
25–34	0.18	0.05	0.19	0.04
35–44	0.22	0.04	0.22	0.04
45–54	0.27	0.05	0.24	0.04
55–64	0.31	0.06	0.26	0.05
65+	0.34	0.07	0.27	0.07
$0–24.9K	0.18	0.07	0.13	0.05
$25–49.9K	0.21	0.04	0.18	0.04
$50–74.9K	0.23	0.04	0.22	0.03
$75–99.9K	0.26	0.07	0.26	0.04
$100K +	0.29	0.11	0.29	0.06

Note: The treatment effect here refers to the change in probability of respondents choosing "too much."

of these groups and we see that our results for partisanship and ideology are quite similar to what Green and Kern found: Democrats and liberals are less "frameable" on welfare spending attitudes than Republicans and conservatives. However, in contrast to their results, we find little heterogeneity in welfare attitudes corresponding to attitudes toward Blacks. This later finding may reflect the fact that in order to economize on questions, we transformed Green and Kern's four-question battery about racial attitudes into one question with a multiple-response format.[25]

In Table 7, we provide a final analysis that is useful in evaluating GCS as a vehicle for examining heterogeneity of treatment effects in surveys experiments. In this analysis, we again explore heterogeneity in the welfare framing experiment, but this time with respect to age, gender, and income. This allows us to compare the results we get for self-reported and inferred versions of these variables.

The results show that the treatment effect varies in the same direction for both the inferred and the self-reported demographics. Indeed, a close examination of the sizes of the various estimated effects shows that they are quite close indeed. This demonstrates that, at least for this experiment, the noise in the inferred demographics described above is not fatal to the use of these variables to study heterogeneity.

[25]While more evidence is needed to make a definitive statement, this result certainly draws attention to the idea that question format may well matter a lot and GCS does impose limits in both the number of questions and their formats. Thus, researchers should carefully consider if the platform is appropriate for measuring concepts, like the one here, for which batteries of questions provide the best measurement.

11 Representativeness of the GCS Samples

The final issue we address is the representativeness of the GCS sample. Like almost every other online survey service (most of whom use opt-in panels), GCS's survey wall methodology does not produce a probability sample of any population. Thus, just like these other services, GCS uses two methods to produce samples intended to be representative.[26] First, they use stratified sampling to try to select respondents to match census targets (on gender, age, and region). Second, they provide post-stratification weights constructed using (inferred) demographics and also provide these demographics to researchers (at no additional cost) who can use them to make post-sampling adjustments in whatever way they prefer (e.g., building their own weights or applying multilevel - regression with post-stratification as recommended by Andrew Gelman and his colleagues [Wang et al. 2015]). This two-pronged strategy to achieve a representative sample is the same one taken by YouGov, SSI, Toluna, and many other online survey houses (see the review by Bremer [2013]).[27]

With respect to the strategy of dynamically targeting specific demographic categories, the success of the effort (assuming one is targeting the right demographics) depends a great deal on the level of non-response (if selected individuals do not actually take the survey, it's hard to target them successfully). Further, GCS seems to suffer from higher levels of non-response than opt-in panels (see Table A7.1 in Section 7 of the Supplementary Materials) especially for longer surveys (which, as we explained above, is unsurprising given the mechanics of the survey wall). Indeed, for the five surveys we conducted for this paper, we had overall response rates (for the four one-question surveys) of 29.5%, 25.5%, 26%, 24.5%, and (for our ten-question survey) 6.4%. Correspondingly, the average root mean-squared error (and maximum bias) of GCS's inferred demographics from census targets on age, region, and gender were 2.7% (12.2 for gender), 2.9% (12.9 for gender), 2.9% (10.6 for gender), 3.6% (14.5 for gender), and 4.3% (19.2 for gender), respectively. Since, in our opinion, most social scientists would likely use the ten-question version of GCS, this suggests that non-response is likely to be very high and so targeting is unlikely to be successful on its own. Thus, the second strategy (post-stratification adjustment via weighting or MRP) will likely be necessary for researchers aiming at representative samples.

Given that most researchers using GCS will have to make post-sample adjustments, it is important to ask whether the post-stratification weights provided by GCS (or the demographic information provided by GCS to construct one's own weights) are adequate to produce a representative sample. The worry, beyond the usual one that the demographic categories adjusted for may be insufficient, is that the noise introduced into a weighted analysis by use of inferred demographics (in constructing the weights) may undermine their usefulness. Ultimately, of course, this is an empirical question, and a number of studies (mostly unpublished) have explored whether weighted data from GCS can be used to reproduce various benchmark inferences. Specifically, McDonald et al. (2012) compared responses from a probability-based Internet panel, a non-probability-based Internet panel, and a GCS sample against a set of media consumption benchmarks.[28] Pew Research Center (see Keeter and Christian 2012) also conducted a (pseudo) mode study in which they compared the results for more than 40 questions asked of a sample of telephone respondents with those asked (over the same time period) of GCS respondents.[29] Finally, Tanenbaum et al. (2013) compared GCS's estimates of home cell-phone usage with three national probability samples. On many of the variables examined in these studies, GCS was quite close to

[26]Here, we will focus on representativeness with respect to the U.S. adult Internet using population (84% of U.S. adults).

[27]Debates about the representativeness of online convenience samples versus probability samples (usually executed via random digit dialing [RDD]) are ongoing and we have little to add, except to point out that, given high levels of non-response in RDD samples (less than 9% in 2012—see Kohut et al. [2012]; Holbrook, Krosnick, and Pfent [2007]; and Steeh et al. [2001]), even probability samples usually require just the kinds of post-survey adjustment that online panels use to try to produce representative samples (see Wang et al. 2015).

[28]The benchmarks are video on demand, digital video recorder, and satellite dish usage in American households as measured by a semi-annual RDD telephone survey of 200,000 respondents.

[29]We say "pseudo mode study" because the respondents were not randomly assigned to a mode, but were sampled separately in what was actually a dual frame design. Thus, some of the differences reported may well come from the different sampling frames rather than mode differences.

the probability-sample benchmarks. Further, in a direct comparison of GCS to alternatives, Silver (2012) examined the predictive accuracy of 23 organizations polling the 2012 U.S. presidential race—concluding that GCS ranked second in average error. That said, none of these results are published in peer-reviewed journals and, while the overall pattern of results clearly supports the accuracy of inferences using GCS, there are certainly some variables examined in these studies that show differences with benchmarks.

With respect to the use of self-reported versus inferred demographics in weighting, we built post-stratification weights using both inferred and self-reported demographics for the ten-question survey described above (keeping the sample constant—so dropping any observations for which either weight could not be calculated). These weights used gender and age, relied on the same population totals used by GCS, and used iterative proportional fitting (raking). The correlation between the weights we produced using inferred demographics and GCS's supplied weight was 0.94. The correlation between our weight using inferred demographics and self-reported demographics was 0.63. The weighted treatment effect for the survey experiment embedded in that survey was −0.306 when using inferred demographics, while for self-reported demographics it was −0.266. Using Google's provided weight on the same sample gave an estimate of −0.317. The unweighted estimate on the same sample is −0.259. Thus, we get quite similar answers when we weight using self-reported and inferred demographics.

Taking all of this together, we think the safest conclusion to draw from those studies that have evaluated GCS's predictive accuracy, as well as our analysis of weighting with inferred demographics—and one in keeping with the simple fact that GCS uses the same adjustment and targeting strategies as most other online services—is simply that there is no reason to think that GCS, after adjustment, is either more or less representative than other non-probability samples.

12 Other Potential Uses of GCS

In this article, we have argued that that GCS may be a useful platform for exploring causal relationships when those relationships can be identified through random assignment of treatments in a survey. That said, scholars often have goals (or proximate goals) other than establishing causal relationships and it may be that GCS is useful for some of these as well. Indeed, in this section we point out two areas in which GCS could help facilitate the adoption of innovative new technologies designed to build better and more efficient surveys.

Our first example is from Montgomery and Cutler (2013), who have recently proposed an algorithm (computer-adaptive testing, or CAT) designed to measure a latent trait based on a small number of items drawn from a much larger battery of items. The algorithm sequentially chooses (in real time) the next item to ask a respondent so that the item is maximally discriminatory (on the unobserved scale) given the particular respondent's background and answers to previous items (e.g., on a political knowledge scale, a low-knowledge individual would be asked a series of increasingly easy questions). In this way, each respondent does not get the whole large battery of items but only a much smaller set that more efficiently places her on the scale. A challenge to practically implementing this kind of innovation, however, is that it requires *a priori* estimates of difficulty and discrimination parameters for each item in the larger battery—a requirement that usually necessitates a separate sample of respondents who respond to all items in the battery. Given that the item batteries are usually large and expensive to field (which is, of course, the whole problem that the method is trying to overcome), this is a substantial practical barrier to the adoption of this innovation.[30] Consider, for example, the empirical illustration Montgomery and Cutler present in their article: They used the CAT algorithm to optimally select (for each respondent) five questions from a 64-item battery of political knowledge items based on previously estimated difficulty and discrimination parameters for each. The "calibration sample" used to produce these estimates included 810 respondents, each of whom was asked about all 64 items. Using the cost estimates provided in footnote 3 and Fig. 1, this calibration survey would have cost

[30]This may explain why the method has not yet seen greater use—we could find no published or unpublished examples.

approximately $18,144 using Toluna's U.S. panel, $32,300 using YouGov's omnibus, and more than $45,000 using Knowledge Networks (all in addition to the cost of the main survey). Of course, such costs undermine the very reason to adopt CAT in the first place and it is not at all clear that the savings gained by asking each respondent fewer questions in the main survey would outweigh the extra costs of conducting the calibration survey. Indeed, Montgomery and Cutler confronted this exact problem in their empirical illustration—opting to use an MTurk calibration sample while warning readers that the resulting estimates were intended only to "illustrate... the usefulness of the CAT method." In contrast, in a substantive application of the method they advise that "researchers should avoid using convenience samples to calibrate the CAT algorithm unless they are comfortable that the distribution of the calibration sample is 'representative' of the overall population of interest on the given latent trait."

Thus, if one were comfortable that GCS's sample is reasonably representative of one's target population, it could be used in the calibration stage to allow for the practical implementation of the CAT method at a price that does not undermine its original purpose. For example, in Montgomery and Cutler's application to political knowledge, the calibration sample (using 810 respondents) would have cost only $5670 (for up to 70 items).[31]

Our second example of how GCS may provide a low-cost means of exploiting new and innovative survey methodologies is Salganik and Levy's (2015) *wiki survey*. The main idea of a wiki survey is that it can solicit respondent input into the design of the survey, adapting questions and answer choices in real time as it is fielded. Two features of the GCS platform facilitate these possibilities. First, like other survey construction platforms, GCS makes it easy to solicit input from respondents using either open-ended questions or closed-list questions with an option to contribute a respondent-generated response. Further, unlike almost every other web-survey platform, GCS allows one to field surveys with very small number of respondents (including only one respondent). Together, these features mean one can dynamically solicit, for example, new response options from respondents and immediately incorporate them into a set of active surveys. For example, Salganik and Levy discuss an effort by the New York City Mayor's Office to solicit (and gauge public support for) ideas for enhancing sustainability. The effort started with 25 ideas, all of which might be used in a forced-choice question (asking which idea the respondent likes best) with the option of supplying a new idea (GCS allows one to include all 25 options in the response set, but randomly selects only seven be seen by a given respondent). This initial survey could be released to a small number of respondents, whose responses could be subsequently examined for new ideas. Any useful new ideas are then easily added to the existing response set and released to more respondents.[32]

13 Conclusion

In this article, we asked whether GCSs is likely to be a useful platform for survey experimenters doing rigorous social scientific work. Overall, our answer is yes. This is based on results from our reanalysis of four canonical experiments and one study of treatment heterogeneity. First, in each experiment, we successfully replicated the directional findings from the previous literature. Second,

[31]This assumes it is implemented in seven 10-question surveys. Further cost savings could also be realized by conducting the GCS surveys in batches and then estimating discrimination and difficulty parameters until one found questions that adequately covered the desired parameter space. Since the CAT algorithm would see two questions with the same parameters as identical, only one such question need be identified. With a judicious selection of questions, this might be considerably fewer than the whole battery.

[32]Salganik and Levy explored this example using a "pairwise wiki survey," which is a particular implementation of the general wiki survey idea that pairs two ideas against each other (with the option of a new idea). Like the design described above, it requires an active researcher to curate contributed ideas and add them to the active response set. One substantive difference between the GCS design described above and the one implemented on their Web site (which leaves open the question of how one drives sample to the Web site—and its representativeness) is that they could ask the same respondent about many different pairings (as many as she was willing to answer). That is not possible in GCS, though one could simply ask many different respondents one question each. To get a sense of the practicality of doing this: in their implementation of the mayoral office survey, the survey was active for a month and ultimately 1436 respondents who contributed 464 new ideas (about half of which were incorporated into the survey) and gave 31,893 responses (pairwise comparisons). Alternatively, a similar design using GCS (on a national sample implemented in the way described above) that obtained 31,839 unique respondents would have cost $31,839.

our estimates of treatment effects were in all but one case within the range of previous results, though usually on the low end. Third, we demonstrated unequivocally that the platform can be used effectively to achieve balance across treatment groups on measured demographics and attitudinal variables. Fourth, our replication of Green and Kern and our examination of the scope of many recent survey experiments suggests that the GCS's ten-question survey, along with additional information provided by Google (inferred demographics and question response times), provides sufficient space to be useful to many survey experimenters interested in exploring treatment effect heterogeneity. Fifth, our analysis of Google's inferred demographics revealed that, while they are certainly noisy, they may be sufficiently accurate to be useful in post-stratification weighting and explorations of heterogeneity (at least when compared with the alternative of using self-reported data). Finally, we conjectured that the mild attenuation of treatment effects that we observed in each experiment could result from the peculiar environment of the survey wall, which is likely to encourage non-cooperation. Our tests using manipulation checks and response frequencies supported this conjecture, though these effects were modest in size. Still, the result suggests that researchers can effectively use such checks to ameliorate some of the negative consequences of the survey wall.

Conflict of interest statement. None declared.

References

Berinsky, Adam J., Michelle 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* 58:739–53.

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* 20:351–68.

Bremer, John. 2013. The interaction of sampling and weighting in producing a representative sample online: An excerpt from the ARF's "foundations of quality 2" initiative. *Journal of Advertising Research* 53:363–71.

DeRouvray, Cristel, and Mick P. Couper. 2002. Designing a strategy for reducing "no opinion" responses in web-based surveys. *Social Science Computer Review* 20:3–9.

Druckman, James N. 2001. Evaluating framing effects. *Journal of Economic Psychology* 22:91–101.

Green, Donald P., and Holger L. Kern. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* 76:491–511.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. Social desirability bias in voter turnout reports. *Public Opinion Quarterly* 74:37–67.

Holbrook, Allyson, Jon A. Krosnick, and Alison Pfent. 2007. The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In *Advances in Telephone Survey Methodology*, eds. James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster. New York: Wiley-Interscience 499–528.

Huber, Gregory A., and Celia Paris. 2013. Assessing the programmatic equivalence assumption in question wording experiments: Understanding why Americans like assistance to the poor more than welfare. *Public Opinion Quarterly* 77:385–97.

Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171(2):481–502.

Janus, Alexander L. 2010. The influence of social desirability pressures on expressed immigration attitudes. *Social Science Quarterly* 91:928–46.

Jou, Jerwen, James Shanteau, and Richard Harris. 1996. An information processing view of framing effects: The role of causal schemas in decision making. *Memory & Cognition* 24:1–15.

Kane, James G., Stephan C. Craig, and Kenneth D. Wald. 2004. Religion and presidential politics in Florida: A list experiment. *Social Science Quarterly* 85:281–93.

Keeter, Scott, and Leah Christian. 2012. *A comparison of results from surveys by the Pew Research Center and Google Consumer Surveys*. Pew Research Center, Washington, DC.

Kohut, Andrew, Scott Keeter, Carroll Doherty, Michael Dimock, and Leah Christian. 2012. *Assessing the representativeness of public opinion surveys*. Pew Research Center, Washington, DC.

Kuhberger, Anton. 1995. The framing of decisions: A new look at old problems. *Organizational Behavior & Human Decision Processes* 62:230–40.

McDonald, Paul, Matt Mohebbi, and Brett Slatkin. 2012. Comparing Google Consumer Surveys to existing probability and non-probability-based Internet surveys. Google White Paper.

Montgomery, Jacob, and Joshua Cutler. 2013. Computerized adaptive testing for public opinion surveys. *Political Analysis* 21(2):141–71.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: methods and principles for social research*. Cambridge: Cambridge University Press.

Mutz, D. C., and R. Pemantle. 2013. The perils of randomization checks in the analysis of experiments. Typescript. University of Pennsylvania.

Pearl, Judea. 2011. The structural theory of causation. In *Causality in the sciences*, eds. P. McKay Illari, F. Russo, and J. Williamson. Oxford: Oxford University Press 697–727.

Rasinski, Kenneth A. 1989. The effect of question wording on public support for government spending. *Public Opinion Quarterly* 53:388–94.

Salganik, Matthew J., and Karen Levy. 2015. Wiki surveys: Open and quantifiable social data collection. *PLoS One* 10(5):e0123483. doi:10.1371/journal.pone.0123483

Santoso, Philip, Robert Stein, and Randy Stevenson. 2016. *Replication data for: Survey experiments with Google Consumer Surveys: Promise and pitfalls for academic research in social science*. http://dx.doi.org/10.7910/DVN/FMH2IR, Harvard Dataverse, Draft version [UNF:6:lh9o42tCLawnMVwewlaPhA==].

Senn, Stephen. 1994. Testing for baseline balance in clinical trials. *Statistics in Medicine* 13:1715–26.

Shih, Tse-Hua, and Xitao Fan. 2008. Comparing response rates from web and mail surveys: a meta-analysis. *Field Methods* 20:249–71.

Silver, Nate. 2012. Which polls fared best (and worst) in the 2012 presidential race? FiveThirtyEight (blog), *New York Times,* November 10.

Steeh, Charlotte G., Nicole Kirgis, Brian Cannon, and Jeff DeWitt. 2001. Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *Journal of Official Statistics* 17:227–47.

Streb, Matthew J., Barbara Burrell, Brian Frederick, and Michael A. Genovese. 2008. Social desirability effects and support for a female American president. *Public Opinion Quarterly* 72:76–89.

Takemura, Kazuhisa. 1994. Influence of elaboration on the framing of decision. *Journal of Psychology* 128:33–39.

Tanenbaum, Erin R., Parvati Krishnamurty, and Michael Stern. 2013. How representative are Google Consumer Surveys? Results from an analysis of a Google Consumer Survey question relative national level benchmarks with different survey modes and sample characteristics. Paper presented at the American Association for Public Opinion Research (AAPOR) 68th Annual Conference, Boston, MA.

Tversky, Amos, and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211:453–58.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. Forecasting elections with non-representative polls. *International Journal of Forecasting* 31:980–91.