methods, data, analyses | Vol. 10(2), 2016, pp. 195-220 DOI: 10.12758/mda.2016.008

Prior Exposure to Instructional Manipulation Checks does not Attenuate Survey Context Effects Driven by Satisficing or Gricean Norms

David J. Hauser [1], *Aashna Sunderrajan* [2], *Madhuri Natarajan* [1] & *Norbert Schwarz* [3]

1 *University of Michigan*
2 *University of Illinois Urbana-Champaign*
3 *University of Southern California*

Abstract

Instructional manipulation checks (IMCs) are frequently included in unsupervised online surveys and experiments to assess whether participants pay close attention to the questions. However, IMCs are more than mere measures of attention – they also change how participants approach subsequent tasks, increasing attention and systematic reasoning. We test whether these previously documented changes in information processing moderate the emergence of response effects in surveys by presenting an IMC either before or after questions known to produce classic survey context effects. When the items precede an IMC, familiar satisficing as well as conversational effects replicate. More important, their pattern and size does not change when the items follow an IMC, in contrast to experiments with reasoning tasks. Given a power of 82% to 98% to detect an effect of $d = .3$, we conclude that prior exposure to an IMC is unlikely to increase or attenuate these types of context effects in surveys.

Keywords: instructional manipulation checks; survey context effects; satisficing; Gricean conversational norms; survey methods

1 Introduction

With the surge in cheap, fast research via online labor markets (Buhrmester, Kwang, & Gosling, 2011), the issue of participant attentiveness has received considerable attention from behavioral researchers (Goodman, Cryder, & Cheema, 2013; Berinksy, Margolis, & Sances, 2013; Paolacci, Chandler, & Ipeirotis, 2010). Some have expressed concern over participant attentiveness in online tasks (see "Quality Assurance" section in Mason & Suri, 2012). Furthermore, many researchers see it as a major issue for research conducted on online labor markets (see informal poll in Chandler, Mueller, & Paolacci, 2014).

One popular method of ensuring attention is the Instructional Manipulation Check (IMC; Oppenheimer, Meyvis, & Davidenko, 2009). The typical IMC is a question that requires close attention to the instructions in order to answer the question correctly; hence, not answering the question correctly is treated as an indication of not paying close attention to the instructions. The standard IMC on the surface looks like a humdrum survey question but contains less noticeable text in the instructions that informs participants to provide an unconventional response in place of an intuitively correct response (Oppenheimer et al., 2009). As an example, a bolded lure question might inquire about which sports you play, but hidden in the instructions may be a command to click the title of the question in order to demonstrate attention. Other methods of checking on participant attention involve asking questions with factually correct, obvious answers, such as, "While watching television, have you ever had a fatal heart attack?" Participants selecting any response other than "never" are presumed to have not been paying attention while responding (Paolacci et al., 2010). These inattentive participants often contribute substantial error to datasets by failing to read the entirety of instructions or by not giving enough thought to questions, which can justify excluding them from analyses (Oppenheimer et al., 2009). Hence, the routine use of IMCs is frequently recommended by online research methodologists as a way to validate online participant pool platforms (e.g., Paolacci et al., 2010; Goodman et al., 2013; Berinsky, Margolis, & Sances, 2013), and they have been become prevalent research tools.

Despite their prevalence as measures of attention, little research has explored how the administration of an IMC itself may affect participants' inferences about the study and their responses to a questionnaire. As research into context effects in self-report highlights, every question is also a treatment that may affect responses

Acknowledgments

We thank Allison Earl for her advice with the research.

Direct correspondence to

David J. Hauser, Department of Psychology, 3233 East Hall
530 Church St, Ann Arbor, MI 48109-1043
E-mail: djhauser@umich.edu

to subsequent questions (for reviews, see Schwarz, 1999; Sudman, Bradburn, & Schwarz, 1996). This may be particularly likely for IMCs, which stand out as unique, salient questions in the context of a standard survey. These questions usually convey the message that researchers want to know if participants are paying attention. This highlights that paying close attention and reading all instructions is important and highly valued in this survey. Furthermore, these questions often attempt to lure participants into responding incorrectly. Thus, IMCs also inform participants that questions may not be what they seem and that the survey may involve "trick" questions that should not be taken at face value. These lessons may increase attention to detail and may prompt a more systematic reasoning strategy than respondents might otherwise adopt.

Initial research suggests that this may be the case. Hauser and Schwarz (2015a, Experiment 1) had participants answer a standard IMC and complete the Cognitive Reflection Test, a series of math questions designed to measure a person's propensity to engage in reflective thinking (Frederick, 2005). For example, a question would read, "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?" (taken from Frederick, 2005). People tend to intuitively respond "100," but the actual answer is "5," which requires more careful, reflective thinking to reach. Hauser and Schwarz varied the order of the tasks, such that the CRT questions either preceded or followed a single IMC question. As expected, participants performed better on the CRT when they had first answered an IMC question. A follow-up study further showed that answering an IMC improves performance on subsequent probabilistic reasoning tasks (Hauser & Schwarz, 2015a, Experiment 2). These findings converge on the conclusion that IMCs do more than "assess" participants' attention: they teach participants that there may be more to a question than meets the eye, which influences how they approach later questions in the survey. As a result, participants who were exposed to an IMC engage in more careful reasoning on subsequent questions, compared to participants who were not exposed to an IMC.

Whether this is a desirable or undesirable effect of using IMCs depends on the researcher's goals. If one wants the most careful answers possible, IMCs may be helpful in achieving the goal. But if one wants to capture how and what people think spontaneously, IMCs may systematically bias one's results. Using the above reasoning tasks as an example, a preceding IMC may be desirable when one wants to test how well people can do when highly motivated. Yet the sample's enhanced performance when an IMC is administered is likely to differ from the performance one would observe under many natural conditions, resulting in erroneous population estimates.

At this point, it is unknown how general the influence of IMCs is. On the one hand, IMCs may only affect performance on tasks that look "tricky" to begin with, such as complex reasoning tasks where correct responses are nonobvious and

require overriding intuitive responses. The tasks affected by IMC administration to this point have fallen into this category, so it is currently unknown whether IMCs may affect other subsequent tasks. On the other hand, participants' motivation and their assumptions about the cooperative nature of the research conversation have been shown to play a key role in all self-report tasks. For instance, minute aspects of surveys such as the survey's letterhead (Norenzayan & Schwarz, 1999), question order (Schwarz, Strack, & Mai, 1991), and administration mode (Schwarz, Strack, Hippler, & Bishop, 1991) all affect survey behavior. Thus, it seems possible that an IMC may influence many common survey tasks because of the unique information that it conveys. Next, we review survey tasks that may be particularly likely to be influenced by IMC placement, namely tasks that give rise to satisficing and Gricean conversational norm effects.

2 Satisficing

Participants often exert less than optimal effort in answering questions. Termed satisficing (Krosnick, 1991, 1999), the phenomenon refers to the practice of taking mental shortcuts rather than considering the full range of inputs in responding to survey questions. Satisficing manifests in specific patterns of survey behavior. *Response order effects* emerge when satisficing participants select the first most reasonable response, resulting in different responses when response option order is manipulated (Schuman & Presser, 1981). Satisficing participants also display *non-differentiation* (Krosnick, 1991, 1999; Krosnick & Alwin, 1988), assigning similar ratings to items using the same scale. *Acquiescence bias* describes the tendency for satisficing participants to simply agree or disagree with statements regardless of their content (Moum, 1988; Winkler, Kanouse, & Ware, 1982). Satisficers also tend to respond more often with "don't know" (DKing) when such a response is offered (Schuman & Presser, 1981), and satisficers show *mark all effects*, selecting less items when questions ask respondents to "mark all items that apply" vs inquire about the relevance of every item individually (Smyth, Dillman, Christian, & Stern, 2006).

The extent to which participants satisfice varies with aspects of survey design. For example, longer surveys, which fatigue respondents, are more prone to satisficing behaviors (Krosnick & Alwin, 1988), and surveys on trivial or non-personally relevant topics, which participants spend less time thinking about, are also prone to satisficing (Krosnick, 1991; Holbrook, Green, & Krosnick, 2003; Holbrook, Krosnick, Moore, & Tourangeau, 2007). Satisficing also increases when questions are difficult to answer (Gage, Leavitt, & Stone, 1957). In addition, satisficing varies with individual difference variables, and satisficers have been found to be less intelligent and less politically informed (Holbrook et al., 2007; Krosnick & Alvin, 1988;

Narayan & Krosnick, 1996). Finally, the IMC development literature also suggests that satisficers are more likely to fail an IMC (Oppenheimer, et al., 2009).

Satisficing is conceptualized as existing on a continuum rather than being a dichotomous measure of present vs absent (Krosnick, 1991). Thus, participants may pass an IMC while still displaying some level of satisficing (Berinsky et al., 2013). Whereas previous research used IMCs as measures of attention, the present research asks whether exposure to an IMC is itself a treatment that influences how much attention respondents pay to subsequent questions. Do respondents show less satisficing after (than before) encountering an IMC question?

3 Conversational Effects

In everyday life, conversations follow a cooperation principle (Grice, 1975) that allows listeners to assume that speakers attempt to be informative, relevant, and clear. When speakers fail to live up to these expectations, listeners draw on the context of the utterance to infer its likely meaning (for reviews see Clark & Clark, 1977; Schwarz, 1994, 1996). Research participants bring these expectations to the research situation and consider all contributions of the researcher to be relevant to their task. These contributions include formal features of questionnaire design, from scale format to graphics and question wording. As a result, many "technical" aspects of questionnaires become a source of information that respondents systematically use to determine what is asked of them (for reviews, see Conrad, Schober, & Schwarz, 2014; Schwarz, 1994, 1996).

For instance, respondents draw on the numeric values of rating scales to interpret the intended meaning of verbal labels (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991; Schwarz, Grayson, & Knauper, 1998), resulting in *scale value effects***.** They also assume that values in the middle range of a frequency scale reflect the population average (Schwarz, Hippler, Deutsch, & Strack, 1985), resulting in *scale range effects.* When encountering an ambiguous question, they draw on the content of prior questions to interpret its meaning, resulting in *question context effects* (Strack, Schwarz, & Wänke, Study 1, 1991). Throughout, respondents assume that the researcher is a cooperative communicator whose contributions are relevant to their task, consistent with the tacit assumptions underlying conversational conduct in everyday life (Schwarz, 1996). Accordingly, they pay close attention to subtle contextual features, in particular when they encounter ambiguous questions. The experience that the researcher presents a "trick" question may influence the emergence of Gricean conversational effects in different ways. On the one hand, learning that attention is called for may increase attention and hence the impact of subtle contextual cues; on the other hand, realizing that the researcher is not always a cooperative communicator may undermine reliance on conversational

norms and hence attenuate the influence of conversational inferences. Next, we turn to these potential influences.

4 Implications of IMCs for Survey Research

If IMCs alert participants that a question may not be what it seems at first glance (as shown by Hauser & Schwarz, 2015a), they may influence responses in a variety of ways. First, they may increase attention to ensure that one isn't "tricked" in subsequent questions. Second, they may teach respondents that the researcher is not a fully cooperative communicator, which may undermine respondents' reliance on conversational norms in making sense of the questions asked. These two possibilities result in differential predictions.

Increased attention

In survey questionnaires, increased attention to the details at hand should attenuate satisficing effects (Krosnick, 1991, 1999), that is, response effects that are commonly attributed to low attention and mental short cuts. The more attention respondents pay to the questions, the less they should resort to "top-of-the-head" answers. In contrast, increased attention to the details at hand should increase conversational inference effects, that is, response effects that are commonly attributed to the operation of conversational norms (Schwarz, 1994, 1995, 1996). These effects require close attention to minor question details (such as numerical values or scale range) in drawing inferences about a question's intended meaning; they should therefore benefit from increased attention. Note that these considerations entail that increased attention and effort have opposite effects on the emergence of satisficing and Gricean norm effects.

Cooperativeness

Complicating predictions, answering an IMC may also teach respondents that the researcher is not a fully cooperative communicator. Asking a question that seems to inquire about X, while noting along the way that X should be ignored in favor of a substantively unrelated response, violates the norms of cooperative conversational conduct (Grice, 1975). The impression that the researcher is not a cooperative communicator, in turn, may reduce the likelihood that participants draw on other features of the questionnaire to infer what the researcher may have had in mind (Schwarz, 1996). If so, response effects based on Gricean conversational processes should be attenuated (rather than increased) when the respective question is preceded by an IMC.

Motivation

Finally, being asked an IMC may also undermine respondents' motivation and willingness to live up to their role – they didn't agree to being "tricked", after all. If so, it may result in more missing data, early termination of online surveys, and so on. Our studies are not suited to assess this possibility because they draw on Amazon Mechanical Turk (MTurk) workers as participants. These online participants are paid for good performance and rely on positive ratings from their employers, which are the basis of reputation scores that drive their future employment. Accordingly, a transparent lack of cooperation is unlikely to be observed in samples of MTurk workers (see Hauser & Schwarz, 2015b).

Manipulation versus measure

Note that our analysis of IMCs treats IMCs as a manipulation of attention, not merely a measure of attention. Our predictions therefore deviate from the more familiar prediction that those who pass an IMC will show less satisficing than those who fail an IMC. The latter prediction pertains to an individual difference in attention and/or motivation and uses IMCs as a measure. In MTurk samples, more than 90 percent of participants routinely pass IMCs (see Hauser & Schwarz, 2015b), indicating that the situational incentives provided by performance-dependent payment and reputation ratings trump variations at the individual difference level. Thus, in the studies that follow, we restrict our analyses to only the participants who pass the IMC in order to assess its potential as a manipulation of attention.

5 Replication, Logic of Analysis, and Data Collection

We test whether previously documented changes in information processing moderate the emergence of context effects in surveys by presenting an IMC either before or after classic survey context effects. This design incorporates replications of classic effects into our investigation. We expect effects driven by satisficing and Gricean norms to replicate when such items precede an IMC, and we test predictions about how an IMC may affect their emergence and size when these items follow an IMC. Note that testing the effect of IMC order is mute if a classic effect does not replicate when administered before an IMC to begin with.

In two online surveys, we presented an IMC either before or after questions expected to elicit classic survey context effects. For ease of presentation, we discuss the satisficing and conversational experiments separately and note in which of the two surveys they appeared. The Method section that follows provides details regarding the online surveys.

6 Method

6.1 Survey 1

Participants

Seven hundred and ninety-eight American Amazon Mechanical Turk (MTurk) workers (456 male, age range 18 - 81) completed a survey in exchange for 40 cents. An a priori power analysis suggested this sample size yields an estimated 98% power for finding an effect of IMC order on satisficing measures when $d = .30$ for the effect of IMC order (Faul, Erdfelder, Lang, & Buchner, 2007).

Materials and procedure

Participants were directed to an online Qualtrics survey ostensibly on current issues. After consenting to the research, participants completed a battery of eight tasks and an IMC. Crucially, random assignment determined the order in which the task battery and IMC were administered. In one condition (IMC first), participants completed the IMC first, followed by the task battery. In the other condition (IMC last), participants completed the task battery, then the IMC. See Appendix A for wording of all questions.

Instructional manipulation check

The IMC was a standard attention check (adapted from Oppenheimer et al., 2009) which has been shown to affect systematic thinking in prior research (Hauser & Schwarz, 2015a) and which has been used extensively in unsupervised online research. In this question, a lure prompt asks participants to choose which of a long list of sports activities they regularly engage in, asking them to check all sports that apply. However, an instruction block informs participants that researchers are interested in their attention levels and, in order to demonstrate attention to the instructions, participants should only select the "other" option below and type in to the accompanying textbox "I read the instructions." Participants who followed these instructions were scored as "passing" the trap question.

Task battery

A battery of eight tasks assessed the degree to which participants exhibited survey context effects. Participants were randomly assigned to receive the tasks in different orders.

Question context and a fictitious issue

In an effort to cooperatively answer questions, participants often assume adjacent questions are related and use prior questions to draw inferences about ambiguous concepts. Modeled on Strack, Schwarz, and Wänke (1991), participants reported whether they favored or opposed (forced choice) a fictitious "Data Sharing Act".

This question was preceded by a question that either referred to Google's decision to grant users control over their personal data or to the U.S. governments' mass collection of private emails and browsing histories; these questions are predicted to provide a positive vs. negative context for interpreting what the fictitious Data Sharing Act refers to, resulting in differential support. These questions constitute a novel conceptual replication of previous experiments on fictitious issues.

Response order

Taken from Schuman & Presser (1981), two tasks assessed satisficing-driven response order effects**.** People taking mental shortcuts don't give full consideration to all response options and tend to select the first reasonable response they consider. When response options are presented visually, the first option is the first considered and is more often selected (Krosnick & Alwin, 1987; Schwarz, Strack, Hippler, & Bishop, 1991). Participants reported which of two statements they agreed with regarding the world's oil supply ("we will still have plenty of oil 25 years from now" or "it will all be used up in about 15 years") and the government's role in supplying adequate housing ("the federal government should see to it that all people have adequate housing" or "each person should provide for his own housing"); the order of responses options was manipulated.

Nondifferentiation

When faced with rating many items on the same scale, satisficers tend to assign many items the same rating. Modeled on Krosnick and Alwin (1988), in a single question matrix, participants rated their interest in thirteen reality television shows on a five point scale (1 = extremely interesting, 2 = very interesting, 3 = fairly interesting, 4 = not too interesting, 5 = not interesting at all). To compute nondifferentiation scores, we counted the number of shows to which participants assigned the same rating.

Don't know

Satisficers are more likely to give "don't know" (DK) responses when these options are offered as it is an easy response. Questions taken from Schuman and Presser (1981) asked about the severity of local courts and about federal government power, and participants were either offered a DK response option or not. All participants typed their response into textboxes, which we coded as falling into the various response options or as expressing a DK response.

Mark all effects

When asked to "mark all items that apply," satisficers tend to consider and mark only a few of the items. This results in less items selected compared to a question that forces respondents to consider each option individually. Modeled on Smyth, Dillman, Christian, and Stern (2006), participants indicated from which of 16 Amazon.com departments they had purchased items in the last 18 months. Partici-

pants were randomly assigned to either "mark all departments that apply" or were asked about each department individually.

Acquiescence
Satisficers often agree or disagree with a majority of statements and contradict themselves in their answers. Modified from Winkler, Kanouse, and Ware (1982), participants selected whether they agreed or disagreed with twenty statements concerning doctors and healthcare. Five pairs of statements (ten statements in total) were logical opposites, which assessed acquiescence bias. The remaining ten statements were filler items.

Task order
We varied the order in which the eight tasks were presented in order to a) assess whether the effects of the IMC on subsequent tasks vary as a function of distance from the IMC and b) assess the sensitivity of our measures to satisficing. We were interested in whether the effects of the IMC "wore off" and became less strong as an item was moved further away from the IMC. Half of the participants received the tasks in the following order: data sharing act, oil supply, reality TV shows, court punishment, adequate housing, Amazon purchasing, government power, and healthcare attitudes. The other half received the tasks in this order: data sharing act, adequate housing, Amazon purchasing, government power, oil supply, reality TV shows, court punishment, healthcare attitudes.

6.2 Survey 2

Participants

Three hundred and ninety seven participants from MTurk participated in the study (254 male, 143 female) in exchange for 40 cents. An a priori power analysis showed that when $d = .30$ (a conservative estimate of the effect size of IMC order) this sample size has 82% power for finding an effect of IMC order (Faul, et al., 2007). The median time to complete the survey was two minutes. We excluded the data of one participant who took twenty-seven minutes (nearly twelve standard deviations beyond the mean survey completion time) to complete the survey, bringing our total number of participants down to 396.

Materials and procedure

Participants were directed to a survey ostensibly addressing current issues. Participants completed an IMC and a series of Gricean conversational norm tasks. They were randomly assigned to receive the IMC as either the first or last question in the survey.

Instructional manipulation check

The IMC (adopted from Oppenheimer et al., 2009) followed the same format as in Study 1. However, unlike Study 1, participants were also randomly assigned to receive feedback on their response. Feedback informed participants of incorrect answers on the trap question and returned them back to the IMC with the instructions "Please try again" in the event of an incorrect response. Participants assigned to receive no feedback were not informed of incorrect answers, and thus simply progressed to the next page of the survey in the event of an incorrect response. However, because we restricted our analyses to only the participants who answered the IMC correctly (as detailed in the upcoming results section), none of our participants whose data was analyzed actually received feedback. Therefore, this manipulation was not included in our analyses and won't be discussed further.

Task battery

Participants completed three tasks designed to measure context effects due to inferences from conversational norms. The wording of all tasks is shown in Appendix A.

Scale range effects

Participants view scale ranges presented by researchers as being informative inputs for their judgments, assuming that middle values in the range reflect population averages. When asked how many hours of television they watch per day, participants given scales that contain more values below the population average (low-skewed scales) report watching less hours of television than participants given scales that contain more values above the population average (high-skewed scales). Additionally, when asked how important a role TV plays in their leisure time, participants given low-skewed scales report a more important role of TV than participants given high-skewed scales. Because participants given low-skewed frequency scales often rate their TV watching frequency above the scale's midpoint, this prompts them to infer that they watch more TV than average and think that TV plays a rather important role in their leisure time (and vice versa for high-skewed frequency scales; Schwarz, Hippler, Deutsch, & Strack, 1985).

Adapted from Schwarz et al. (1985), participants rated how many hours of TV they watch daily. Participants were randomly assigned to either a low frequency scale (ranging from "up to .5 hour" to "more than 4.5 hours") or a high frequency scale (ranging from "up to 4.5 hours" to "more than 8.5 hours"). The scale was created around the actual mean hours of TV viewed per day in America (4.5 hours; Nielsen, 2011), and both scale range conditions contained that mean. Following this question, participants were then asked, "How important is the role of TV in your leisure time?" with an 11-point scale (0 = "not at all important" to 10 = "very important").

Scale label effects

Participants draw on the numeric values of rating scales to infer question meaning. When asked how successful they have been in life, respondents report higher success when the scale runs from -5 ("not at all successful") to +5 ("extremely successful") rather than from 0 ("not at all successful") to 10 ("extremely successful"). This reflects that the bipolar -5 to +5 format suggests an interpretation that spans the whole range from failure (-5) to success (+5), whereas the unipolar 0 to 10 format covers only differential degrees of success (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). We replicated this experiment.

Similarly, participants provide higher ratings of the frequency with which they engage in rare behaviors when the rating scale runs from 0 ("rarely") to 10 ("often") rather than 1 ("rarely") to 11 ("often"). This is the case because "rarely" is interpreted as "never" when combined with 0 and interpreted as a small frequency when combined with 1, resulting in corresponding shifts on the scale (Schwarz, Grayson, & Knauper, 1998). We replicated this experiment with questions about the frequency of getting a haircut, visiting a museum, and attending a poetry reading.

7 Results

IMC performance

In survey 1, 747 participants (93.5%) answered the IMC correctly, while only 52 participants (6.5%) answered it incorrectly. This high IMC pass rate is consistent those of recent research on MTurk (Hauser & Schwarz, 2015b; Nauts, Langner, Huijsmans, Vonk, & Wigboldus, 2014; Wolf, Levordashka, Ruff, Kraaijeveld, Lueckmann, & Williams, 2014). Following convention (Oppenheimer et al., 2009) we restricted our survey 1 data to the sample of participants who answered the IMC correctly because this is the primary sample of interest. Moreover, the small number of participants who failed the IMC does not allow for meaningful comparisons of the question effects of interest.

In survey 2, 369 participants (93%) answered the IMC correctly on their first try. As with survey 1, we restricted our survey 2 sample to the 369 (93%) participants who responded correctly to the IMC because the sample of participants who responded incorrectly was not large enough for drawing firm conclusions.

Satisficing effects

For each question experiment, we first present replication analyses that assess whether the standard satisficing effect emerges when the IMC is the last task in the sequence, that is, under normal survey conditions without a potential IMC intervention. Subsequently, we test whether an observed effect is attenuated when the

Table 1 Summary of satisficing effect results

Satisficing-driven survey context effect	Replicates?	Moderated by IMC order?
response order effects (Schuman & Presser, 1981)		
oil supply	no	no
adequate housing	yes	no
nondifferentiation (Krosnick & Alwin, 1988)		
reality TV shows	yes	no
DKing (Schuman & Presser, 1981)		
court punishment	yes	no
government power	yes	no
mark all effects (Rasinksi et al., 1994)		
Amazon purchasing	yes	no
acquiescence (Winkler et al., 1982)		
healthcare attitudes	–	no

IMC precedes rather than follows the items of interest. Table 1 summarizes the conclusions.

Response order effects

One of the two response order questions in survey 1 asked whether the government should provide adequate housing (taken from Schuman & Presser, 1981). When the IMC was presented last, participants were more likely to choose "the government" as their response when it was the first response option listed (49%) than when it was the last option listed (36%); $\chi^2(N = 372, 1) = 5.74$, $p = .017$, $\varphi = .12$. This replicates to the standard response order effect.

To assess if prior IMC administration attenuated this effect, we conducted a logistic regression with IMC order (IMC first, IMC last), response option order (government first, government last), task order (2nd task, 5th task), and their interactions entered as mean-centered categorical predictors of response to the adequate housing question (1 = government, 2 = each person). Importantly, this response order effect was unaffected by prior answering an IMC, $\beta = -0.04$, $Wald = 0.35$, $p = .56$ for the 2 way interaction of IMC order and response order. As suggested by the replication analysis, the main effect of response option order was significant, $\beta = .30$, $Wald = 16.12$, $p > .001$, $OR = 1.35$. All other main effects and interactions failed to reach significance, ps $> .12$. In sum, prior exposure to an IMC did not attenuate the classic response order effect on this task.

A second response order question in survey 1 pertained to the *oil supply* (Schuman & Presser, 1981). Under standard conditions (IMC last) the familiar response order effect did not replicate, χ^2 (N = 372, 1) = 1.11, p = .29. Hence, this item cannot serve as an index of satisficing in our sample. (For additional analyses of this item see Appendix B.)

Nondifferentiation

One question in survey 1 concerning interest in reality TV shows assessed non-differentiation behavior. Survey fatigue effects suggest nondifferentiation should increase when the task is administered later in the survey (Krosnick, 1991). We replicated this effect when the IMC was presented last; the mean number of identically-rated shows was higher when the reality TV show question was presented sixth in the task battery (M = 9.38, SD = 2.80) compared to when it was presented third in the battery (M = 8.53, SD = 2.63); F(1, 369) = 9.01, p = .003, η_p^2 = .024, 95% CI [-1.40, -0.29] for the effect of task order.

To test for a potential effect of IMC placement, we conducted a 2 (IMC order: IMC first, IMC last) x 2 (task order: 3rd task, 6th task) between subjects analysis of variance on the number of shows given an identical rating. First answering an IMC did not affect nondifferentiation; F < 1 for the main effect of IMC order. The interaction of IMC order and task order also did not reach significance; F(1, 741) = 1.91, p = .168. As shown in the replication analysis, the main effect of task order was significant, F(1, 741) = 8.15, p = .004, η_p^2 = .011, 95% CI [-0.48, -0.09]. Thus, prior exposure to an IMC did not lessen participants' nondifferentiation behavior.

DK effects

Two questions in survey 1 assessed the influence of offering a DK option. When the IMC was presented last, the standard effect replicated for both questions. On the question regarding court punishment (Schuman & Presser, 1981), participants were much more likely to indicate a "don't know" response when a DK option was explicitly offered (58.1%) than when it was not explicitly offered (0%); χ^2(N = 373, 1) = 152.83, p < .001, φ = .64 for the effect of DK option.

In order to assess whether the experimental treatments significantly affected DK responses to this question, we limited our sample to the participants who were offered a DK option and conducted a logistic regression with IMC order, task order (4th task, 7th task), and their interaction entered as mean centered categorical predictors of giving a DK response (0 = non-DK, 1 = DK). Task order did not affect DK responses, β = -.17, *Wald* = 2.69, p = .101, *OR* = 0.84 for the main effect of task order. Prior answering an IMC also did not affect DK responses, β = .02, *Wald* = 0.07, p = .813 for the main effect of task order. The interaction of task order by IMC order was also not significant, β = .09, *Wald* = .78, p = .377. Thus, while standard DK effects replicated, prior exposure to an IMC did not significantly lessen the extent to which participants selected a DK response.

On the question regarding government power (Schuman & Presser, 1981), participants were again more likely to indicate a "don't know" response when a DK option was offered (14.0%) than when it was not (0%); $\chi^2(N = 372, 1) = 27.95$, $p < .001$, $\varphi = .27$.

In order to assess whether the experimental treatments significantly affected DK responses to this question, we limited our sample to the participants who were offered a DK option and conducted a logistic regression with IMC order, task order (4th task, 7th task), and their interaction entered as mean centered categorical predictors of giving a DK response (0 = non-DK, 1 = DK). DK responses were no more likely in either task order; $\beta = -.05$, *Wald* = 0.10, $p = .748$, for the main effect of task order. DK responses were also not affected by prior seeing an IMC; $\beta = -.01$, *Wald* = 0.00, $p = .968$ for the main effect of IMC order. Finally, the interaction of IMC order and task order was not significant, $\beta = -.23$, *Wald* = 2.27, $p = .131$. Thus, while standard DK effects replicated, prior exposure to an IMC did not significantly alter the extent to which participants selected a DK response for either question.

Mark all effects

One question in survey 1 regarding Amazon.com department purchases assessed mark all effects. Participants tend to select fewer options when given a mark all question type than when asked about each option individually (Smyth et al., 2006). We replicated this effect when the IMC was presented last; participants selected less departments when asked to *mark all* ($M = 3.7$, $SD = 2.5$) than when asked about each department separately ($M = 4.7$, $SD = 3.3$); $F(1, 370) = 11.69$, $p = .001$, $\eta_p^2 = .03$, 95% CI [-1.64, -0.44].

In order to assess if prior answering an IMC attenuates this effect, we conducted a 2 (IMC order: IMC first, IMC last) x 2 (task order: 3rd task, 6th task) x 2 (question type: mark all, individual questions) between subjects analysis of variance on the number of Amazon.com departments selected. As shown in the replication analysis, the main effect of question type was significant, $F(1, 738) = 22.60$, $p < .001$, $\eta_p^2 = .03$, 95% CI [-0.70, -0.29]. The effect of question type was also marginally moderated by task order: interaction of task order x question type, $F(1, 738) = 3.40$, $p = .066$, $\eta_p^2 = .01$, 95% CI [-0.01, 0.39]. Simple effects tests showed that when the task appeared as the 3rd task in the battery, participants selected less departments when given a *mark all* item type ($M = 3.5$, $SD = 2.3$) than when given an *individual questions* item type ($M = 4.9$, $SD = 3.3$); $F(1, 738) = 21.94$, $p < .001$, $r = .17$ for the simple main effect. When the task appeared as the 6th task in the battery, the effect of question type was in the same direction but less strong. In these conditions, participants selected less departments when given a *mark all* item type ($M = 3.6$, $SD = 2.4$) than when given an *individual questions* item type ($M = 4.2$, $SD = 3.2$); $F(1, 738) = 4.2$, $p = .041$, $r = .07$ for the simple main effect.

Importantly, the effect of question type was not attenuated by prior answering an IMC: $F < 1$ for the interaction of question type and IMC order. All other interac-

tions and main effects failed to reach significance, *p*s > .20. Thus, prior exposure to an IMC did not lessen classic "mark all" effects.

Acquiescence

Survey 1 also included an empirically-validated acquiescence scale that assesses how many contradictory statements regarding healthcare that a respondent endorses (Winkler et al., 1982). Prior exposure to an IMC did not lessen acquiescence on this scale, $F < 1$ for the effect of IMC order on the number of contradictory statement pairs each participant selected.

Gricean conversational norm effects

Next, we turn to Gricean conversational norm effects. For each experiment, we again report whether the original effect replicated and then assess whether its emergence and size is moderated by the placement of an IMC. Table 2 summarizes the analyses.

Question context and a fictitious issue

One question in survey 1 assessed whether participants used a preceding context question to disambiguate the meaning of a fictitious Data Sharing Act. Replicating the findings of Strack, Schwarz, and Wänke (1991) with a novel question set, when the IMC was presented last, a favorable context prompted more "favor" responses to the fictitious issue (46.5% favor) than an unfavorable context (9.1% favor) χ^2 (1, N = 372) = 34.95, $p < .001$, $\varphi = .42$.

In order to assess if this effect was moderated by IMC order, we conducted a logistic regression with IMC order (IMC first, IMC last), prior question context (favorable, unfavorable), and their interaction entered as mean-centered categorical predictors of approval of the fictitious issue (1 = favor, 2 = oppose). Consistent with the replication analysis, the main effect of prior question context was significant, β = .93, *Wald* = 89.43, $p < .001$, *odds ratio [OR]* = 2.53. All other effects failed to reach significance; β = .07, *Wald* = .45, $p = .50$, for the main effect of IMC order and β = .15, *Wald* = 2.38, $p = .12$, for the interaction of IMC order and context. Thus, placement of the IMC did not influence the extent to which participants drew on question context in interpreting an ambiguous issue.

Scale range effects – behavioral report

One question in survey 2 assessed whether reports of TV consumption were affected by the range of the frequency scale. Today, the average TV consumption in the United States is about 4.5 hours (Nielsen, 2011). When the IMC was presented last, 19.6% of the participants reported watching more than 4.5 hours when given the high frequency scale, whereas only 3.4% did so when given the low frequency scale; χ^2 (1, N = 369) = 11.47, $p = .001$, $\varphi = .25$. This replicates the original pattern reported by Schwarz et al. (1985) with values that have been adjusted to reflect current TV consumption.

Table 2 Summary of Gricean norm effect results

Gricean-driven context effect	Replicates?	Moderated by IMC order?
Question context and a fictitious issue (Strack et al., 1991)		
data sharing act	yes	no
Scale labels (Schwarz et al., 1991; Schwarz et al., 1998)		
life success	yes	no
rare behavior frequency	no	no
Scale range (Schwarz et al., 1985)		
TV consumption – behavioral report	yes	no
TV consumption – comparative judgment	yes	yes

To test if scale range effects are moderated by prior exposure to an IMC, we conducted a logistic regression with IMC order (first, last), scale range (low, high), and their interaction entered simultaneously as mean-centered categorical predictors of the likelihood of participants saying they watch more than the mean amount of TV per day (0 = no, 1 = yes). Importantly, IMC order did not moderate scale range effects, $\beta = 0.53$, $Wald = 0.40$, $p = .527$ for the two way interaction of IMC order and scale range. Consistent with the replication analysis above, the effect of the scale range was significant, $\beta = 1.66$, $Wald = 15.95$, $p < .001$, $OR = 5.28$ for the main effect. The main effect of IMC order was not significant, $\beta = -0.25$, $Wald = 0.38$, $p = .540$. Thus, IMC order did not affect this Gricean norm effect.

Scale range effects – comparative judgment

A follow-up question in survey 2 assessed whether judgments of TV's importance in participant's leisure activities were affected by the frequency scale presented with the behavioral question. Participants who report their behavioral frequency along a low (high) frequency scale endorse values in the higher (lower) range of the respective scale. As observed in previous research (Schwarz et al., 1985), participants infer their likely placement in the distribution from their placement on the scale. Hence, a low frequency scale suggests that their own TV consumption is above average, whereas a high frequency scale suggests that it is below average. This, in turn, affects judgments of how important TV is in their own lives. Replicating this effect, participants given a low frequency scale range rated TV as being more important to their leisure time ($M = 5.38$, $SD = 2.41$) than participants given a high scale range ($M = 4.62$, $SD = 2.63$); $F(1, 187) = 4.26$, $p = .040$, $\eta_p^2 = .02$, 95% CI [0.03, 1.49] for the effect of scale range when the IMC is presented last.

In order to investigate if IMC order moderates this effect, we conducted a 2 (IMC order: first, last) x 2 (scale range: low, high) between subjects analysis of variance on the importance of TV in participants' lives. There were no main effects, *p*s > .10. However, IMC order did marginally moderate the effect of scale range: $F(1, 361) = 3.18$, $p = .075$, $\eta_p^2 = .01$, 95% CI [-0.49, 0.02] for the interaction of IMC order and scale range. As shown before, when participants received the IMC last, there was the typical effect of scale range; those participants presented with a low scale range reported TV as being more important in their lives compared to those participants who received the high scale range: $F(1, 365) = 4.26$, $p = .040$, $r = .11$, 95% CI [0.02, 0.74] for the simple effect of scale range. However, this effect was eliminated when participants answered the IMC first. In this case, TV importance ratings did not differ ($M = 4.63$ and 4.80, $SD = 2.57$ and 2.49 for the low and high frequency conditions, respectively), $F < 1$ for the simple effect of scale range. Thus, IMC order moderated this effect. We discuss the implications of this observation in the General Discussion.

Scale label effects

Two tasks in survey 2 assessed scale label effects. When asked about their success in life, participants provide more modest ratings when the numeric values of the rating scale suggest that the low anchor of the scale refers to the absence of outstanding achievements (0 = not at all successful to 10 = very successful) rather than the presence of explicit failure (-5 = not at all successful to +5 = very successful; Schwarz et al., 1991). Replicating this effect, 44.7% of the participants endorsed a value in the lower half of the 0-to-10 scale, whereas only 30.5% of the participants did so on the -5 to +5 scale; $\chi^2 (1, N = 189) = 4.04$, $p = .045$, $\varphi = .15$ for the effect of scale values when the IMC was asked last.

To assess if IMC order moderates this effect, we conducted a logistic regression, where IMC order (first, last), scale label numeric values (-5 to +5, 0 to 10), and their interaction were entered simultaneously as mean-centered categorical predictors of participants' placing themselves in the lower half of the respective life success scale (0 = no, 1 = yes). IMC order did not moderate the impact of the numeric scale values, $\beta = 0.42$, *Wald* = 0.94, $p = .332$ for the two way interaction of scale label and IMC order. Consistent with the replication analysis, the main effect of scale labels was marginally significant, $\beta = 0.41$, *Wald* = 3.56, $p = .059$, *OR* = 1.50. The main effect of IMC order was also not significant, $\beta = -0.23$, *Wald* = 1.15, $p = .283$. Thus, IMC order does not increase this Gricean norm effect.

For the second scale label task, participants reported their frequency of engaging in rare behaviors. In previous research, participants interpreted the verbal end anchor "rarely" as "never" when it was paired with the numeric value 0, but not when paired with the numeric value 1. As a result of this shift in scale interpretation, they provided higher ratings along a 0 to 10 scale than along a 1 to 11 scale (Schwarz, Grayson, & Knäuper, 1998). This influence of numeric scale values was

not observed in our sample of participants receiving the IMC last, $F < 1$. This nonreplication renders the task unsuitable for exploring the potential influence of IMC order on Gricean task interpretations.

8 General Discussion

Instructional manipulation checks (IMCs) aim to identify research participants who pay little attention. These participants may introduce noise. Hence, identifying and excluding them has been found to increase data quality (Oppenheimer et al., 2009). However, cognitive research into the question-answering process highlights that every measurement is also a treatment (e.g., Nebel, Strack, & Schwarz, 1989; for a discussion, see Sudman, Bradburn, & Schwarz, 1996). If so, answering an IMC may influence participants' performance on subsequent tasks. Supporting this possibility, Hauser and Schwarz (2015a) found that participants performed better on reasoning tasks that required careful analytic reasoning when an IMC preceded rather than followed the task. This observation is potentially worrisome for survey researchers – although attention to survey tasks is generally desirable, inducing the sample to pay more attention to a task than the population ever may under natural conditions can result in erroneous population estimates.

As far as standard survey questions are concerned, the present findings indicate that there is less reason to worry than the Hauser and Schwarz (2015a) results suggested. In two online surveys with MTurk workers we administered twelve question experiments, seven pertaining to satisficing effects and five pertaining to Gricean norm effects. Two conclusions stand out. First, as shown in Tables 1 and 2, the classic response effects were highly robust and replicated well. The two exceptions were a nonreplication of a response order effect on Schuman and Presser's (1981) oil supply item and an influence of the numeric values of a rating scale on behavioral reports (Schwarz et al., 1998). There are no obvious reasons for these nonreplications and their cause is of limited interest for the present research, which requires the replication of response effects to assess their potential moderation through the placement of IMCs.

Second, and more important, the placement of IMCs did not affect the emergence, direction, or size of response effects (see Tables 1 and 2). The single exception is the observation that the range of a behavioral frequency scale influenced subsequent comparative judgments under standard conditions (replicating Schwarz et al., 1985), but not when an IMC preceded the question. Considered in isolation, this observation would be consistent with the assumption that IMCs undermine participants' belief that the researcher is a cooperative communicator. However, this interpretation is thwarted by the fact that a preceding IMC did not attenuate the influence of the scale manipulation on the behavioral report itself; nor did IMCs attenuate any of the other Gricean effects.

In combination, our findings are good news for survey methodologists. Although IMCs can influence how participants approach complex reasoning tasks (Hauser & Schwarz, 2015a), they seem unlikely to affect how they approach standard survey questions. We assume that the crucial difference is in the apparent nature of the task. Reasoning tasks of the type used by Hauser and Schwarz (2015a; taken from Frederick, 2005, and Toplak, West, & Stanovich, 2011) invite erroneous answers because the first answer that leaps to mind is objectively wrong, which more effortful systematic thinking elucidates. These tasks assess intuitive versus reflective thinking and were designed in such a way that a person must reflect in order to recognize that the initial intuitive answer is wrong. Thus, these questions require an element of error detection for correct answers and many people experience the questions as "tricky".

This is not the case for questions that give rise to satisficing effects and Gricean effects in survey research. These questions often ask people's opinions about issues or estimations of their own behaviors and are hardly perceived as "tricky." Further, these questions often lack a clearly right or wrong answer, and are thus unlikely to initiate error detection processes. Accordingly, questions relating to satisficing may not invite the same suspicion as complex reasoning tasks. If so, prior exposure to an IMC may only initiate systematic thinking on later *tricky-seeming* tasks that have objectively correct answers (which participants can check via systematic reasoning) while having no effects on other tasks. These conjectures await systematic testing.

References

Berinsky, A.J., Margolis, M.F., & Sances, M.W. (2013). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 1-15.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112-130.

Conrad, F.G., Schober, M.F., Schwarz, N. (2014). Pragmatic processes in survey interviewing. In T. Holtgraves (Ed.), *Handbook of language and social psychology* (pp. 420-437). Oxford, UK: Oxford University Press.

Clark, H.H., & Clark, E.V. (1977). *Psychology and language.* New York: Harcourt, Brace, Jovanovich.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 25-42.

Gage, N.L., Leavitt, G.S., & Stone, G.C. (1957). The psychological meaning of acquiescence set for authoritarianism. *The Journal of Abnormal and Social Psychology*, *55*, 98-103.

Goodman, J.K., Cryder, C.E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213-224.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41-58). New York: Academic Press.

Hauser, D.J. & Schwarz, N. (2015a). It's a trap! Instructional manipulation checks prompt increased effort on "tricky" tasks. *SAGE Open*, *5*, 1-6. doi:10.1177/2158244015584617

Hauser, D.J. & Schwarz, N. (2015b). Attentive Turkers: MTurk participants perform better on attention checks than do subject pool participants. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-015-0578-z

Holbrook, A.L., Green, M.C., & Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*, 79-125.

Holbrook, A.L., Krosnick, J.A., Moore, D., & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly*, *71*, 325-348.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.

Krosnick, J.A. & Alwin, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201-19.

Krosnick, J.A., & Alwin, D.F. (1988). A test of the Form-Resistant Correlation Hypothesis Ratings, Rankings, and the Measurement of Values. *Public Opinion Quarterly*, *52*, 526-538.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1-23.

Moum, T. (1988). Yea-saying and mood-of-the-day effects in self-reported quality of life. *Social Indicators Research*, *20*, 117-139.

Nauts, S., Langner, O., Huijsmans, I., Vonk, R., & Wigboldus, D.H. (2014). Forming impressions of personality. *Social Psychology*, *45*, 153-163

Nielsen. (2011). *State of the media: The cross-platform report.* New York: The Nielsen Company.

Narayan, S., & Krosnick, J.A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58-88.

Nebel, A., Strack, F., & Schwarz, N. (1989). Tests als Treatment: Wie die psychologische Messung ihren Gegenstand verändert. [Tests as treatments.] *Diagnostica, 35*, 191-200.

Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*, *29*, 1011-1020.

Oppenheimer, D.M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867-872.

Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*, 411-419.

Schuman, H., & Presser, S. (1981). *Questions and answers: Experiments on question form, wording, and context in attitude surveys*. New York, NY*:* Academic.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology, 26,* 123-162.

Schwarz, N. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. (The 1993 Morris Hansen Lecture) *International Statistical Review, 63*, 153-177.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods and the logic of conversation*. Hillsdale, NJ: Erlbaum.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, *54*, 93-105.

Schwarz, N., Grayson, C.E., & Knäuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10*, 177-183.

Schwarz, N., Hippler, H.J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*, 388-395.

Schwarz, N., Knäuper, B., Hippler, H.J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570-582.

Schwarz, N., Strack, F., Hippler, H.J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, *5*, 193-212.

Schwarz, N., Strack, F., & Mai, H.P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, *55*, 3-23.

Smyth, J.D., Dillman, D.A., Christian, L.M., & Stern, M.J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, *70*, 66-77.

Strack, F., Schwarz, N., & Wänke, M. (1991). Semantic and pragmatic aspects of context effects in social and psychological research. *Social Cognition*, *9*, 111-125.

Sudman, S., Bradburn, N.M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass.

Toplak, M.E., West, R.F., & Stanovich, K.E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275-1289.

Winkler, J.D., Kanouse, D.E., & Ware, J.E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, *67*, 555-561.

Wolf, W., Levordashka, A., Ruff, J.R., Kraaijeveld, S., Lueckmann, J.M., & Williams, K.D. (2014). Ostracism Online: A social media ostracism paradigm. *Behavior Research Methods*, 1-13.

Appendix A

Survey 1 materials

Data sharing act

(favorable context) How do you feel about Google's decision to allow users complete control over the data they share (or choose not to) with advertisers?

(unfavorable context) How do you feel about the government's decision to allow government agencies to collect privately-shared data from internet users' email accounts and browsing histories?

Congress has been considering the Data Sharing Act of 2013. Do you favor or oppose the passage of this act?

Oil supply

Some people say that we will still have plenty of oil 25 years from now. Others say that at the rate we are using our oil, it will all be used up in about 15 years. Which of these ideas would you guess is most nearly right?

Adequate housing

Some people feel the federal government should see to it that all people have adequate housing, while others feel each person should provide for his own housing. Which comes closest to how you feel about this?

Reality TV shows

Please look at the reality television shows listed below. Could you please tell me whether you find the reality television show to be extremely interesting, very interesting, fairly interesting, not too interesting, or not interesting at all?

- The Real Teenagers of Beverly Hills
- Survivor
- Fish Tank Kings
- The Biggest Loser
- Hell's Kitchen
- So You Think You Can Dance?
- Shahs of Sunset
- Geeks vs. Greeks
- Married to a Vampire
- America's Next Top Model
- Millionaire Matchmaker
- The Bachelor
- The Apprentice

Court punishment

In general, do you think that the local courts in your area deal too harshly or not harshly enough with criminals (or do you not have enough information to say)? Enter "too harshly" or "not harshly enough" (or "not enough info") in the text box below.

Government power

Some people are afraid the government in Washington is getting too powerful for the good of the country and the individual person. Others feel that the government in Washington is not getting too strong. (Have you been interested enough in this to favor one side over the other? If so,) What is your feeling, do you think the government is getting too powerful or do you think the government is not getting too strong? Enter ("not interested enough,") "too powerful" or "not too strong" in the text box below.

Amazon purchasing

(mark all) From which of the following departments on Amazon.com have you made a purchase in the last eighteen months? (Check all that apply)

(individual questions) Have you or have you not purchased from the following departments on Amazon.com in the last eighteen months? (Select Yes or No)

- Unlimited Instant Videos
- MP3s and Cloud Player
- Amazon Cloud Drive
- Kindle
- Appstore for Android
- Digital Games and Software
- Audible Audiobooks
- Books
- Movies, Music & Games
- Electronics and Computers
- Home, Garden & Tools
- Grocery, Health & Beauty
- Toys, Kids & Baby
- Clothing, Shoes & Jewelry
- Sports & Outdoors
- Automotive & Industrial

Healthcare attitudes

Please look at the statements below and indicate whether you agree or disagree with each statement.

Doctors don't always explain to their patients the risks involved in certain treatments
(a) There is little a person can do to prevent illness
I'd rather my doctor just told me what to do
(b) Doctors do not always check everything they should check when examining their patients
Good doctors nearly always agree on how to treat a specific illness
(c) Prescription drugs frequently do more harm than good
Good health is largely a matter of luck
(d) Most doctors carefully explain what will happen to their patients
It mainly takes good medical care to get over an illness
Going to the doctor's office for check-ups is necessary
In the long run, people who take good care of themselves stay healthier and get well more quickly
(a) Anyone can learn a few basic health rules, which will go a long way in preventing illness
(e) A person should take medicine only as a last resort
It is important to seek immediate medical advice when you notice something wrong or unusual
(d) Doctors don't usually explain your medical problems to you
Sometimes doctors prescribe treatments that involve unnecessary risks
Your health is based more on genetics than the environment
(b) Doctors are very careful to check everything when examining their patients
(e) It's always silly to suffer if medicine will make you feel better
(c) Prescription drugs are almost always helpful

Survey 2 materials

TV consumption

On average, how many hours of TV do you watch daily?
(low frequency scale) Up to .5 hour, .5 hour to 1.5 hours, 1.5 hours to 2.5 hours, 2.5 hours to 3.5 hours, 3.5 hours to 4.5 hours, More than 4.5 hours
(high frequency scale) Up to 4.5 hours, 4.5 hours to 5.5 hours, 5.5 hours to 6.5 hours, 6.5 hours to 7.5 hours, 7.5 hours to 8.5 hours, More than 8.5 hours
How important is the role of TV in your leisure time?
1 = not at all important to 10 = very important

Life success

How successful have you been in life so far? Please use the following rating scale from -5 (not at all successful) to +5 (extremely successful) [from 0 (not at all successful) to 11 (extremely successful)].

Rare behavior frequency

How often do you get a haircut?
0 (1) = rarely to 10 (11) = often
How often do you visit a museum?
0 (1) = rarely to 10 (11) = often
How often do you attend a poetry reading?
0 (1) = rarely to 10 (11) = often

Appendix B

Table A1 Task order by IMC order by response order on *oil supply* response selection

	oil supply is 2nd question in battery				*oil supply* is 5th question in battery			
	IMC first		IMC last		IMC first		IMC last	
	plenty 1st	plenty 2nd	plenty 1st	plenty 2nd	plenty 1st	plenty 2nd	plenty 1st	plenty 2nd
plenty	58%	48%	58%	58%	47%	62%	59%	48%
used up	42%	52%	42%	42%	53%	38%	41%	52%

We conducted a logistic regression with IMC order (IMC first, IMC last), response option order (plenty first, plenty last), task order (2nd task, 5th task), and their interactions entered as mean-centered predictors of responses to the *oil supply* question (1 = plenty, 2 = used up). While the three way interaction of task order by IMC order by response option order was significant, $\beta = .18$, $Wald = 6.04$, $p = .014$, $OR = 1.20$, the patterns did not replicate the usual response order effect in any of the conditions (see Table A1) and is thus uninformative.