Surveys as a Social Experience: The Lingering Effects of Survey Design Choices on Respondents' Survey Experience and Subsequent Optimizing Behavior

Delia Dumitrescu, PhD

Postdoctoral fellow, Multidisciplinary Opinion and Democracy Group

Department of Political Science, University of Gothenburg

delia.dumitrescu@gu.se

and

Johan Martinsson, PhD

Research director, Laboratory of Opinion Research (LORE)

Department of Political Science, University of Gothenburg

johan.martinsson@pol.gu.se

Abstract

We propose that the design of earlier survey sections affects respondents' response strategy to later unrelated questions. We hypothesize that the structure of the survey is socially construed, and when earlier survey design features are respectful of the rules of social conversation, individuals are more likely to optimize their responses later on and express more satisfaction in end-of-survey evaluations. We find evidence supporting these expectations from two experiments, but more research is needed to sort out the causal mechanism responsible for these effects.

Keywords: survey methodology, context effects, questionnaires, survey optimizing, maxims of conversation

Survey practice has a lot to gain from applying rules of conventional wisdom: Using simple words, keeping the questions simple, not asking about multiple objects at the same time, and so on, yields better quality data (Krosnick & Presser, 2010). An impressive body of literature on the consequences of breaking any one of them, as well as significant advancements in the psychology of survey response (e.g., Tourangeau, Rips & Rasinski, 2000), have provided valuable information about the data biases one expects to find from, for example, poor question formulation, poor choice of scales, poor question order, and poorly formulated questions about sensitive objects (Berinsky, 1999). Much of this literature has been concerned with the effects that a particular question design has on responses to that question or battery or with the cognitive effects that question order has on responses to subsequent questions (in terms of the type and amount of information retrieved, e.g., priming). In this paper we examine how the design of earlier survey sections affects respondents' optimizing strategy (cf. Krosnick & Presser 2010, p. 265) to subsequent, unrelated questions.

Previous studies found evidence suggesting that surveys are like a social conversation between the respondent and the researchers (except that they are a structured, directed conversation for the purposes of getting some specific and, to the extent possible, unbiased information; e.g., Hippler & Schwarz, 1989; Schwarz, 1994; 1996; Schober, 1999). If respondents perceive surveys as conversations, then breaking any social conversation rule during the interview should affect the respondent's behavior down the line. That is, asking a question that a respondent knows she cannot answer correctly, or she feels pushed on, should not just affect the answer to that particular question, but should also send a negative social signal about the survey designers (e.g., of sloppiness or bias) which in turn should affect response strategies to further questions. On the positive side, actively making known the researchers' intention to

collect high quality data (for example, by asking people to weight carefully the various response options before giving their response) should not just affect the answers to the question following this acknowledgment, but should also affect behavior to further questions, as the most likely interpretation of this signal is that it applies to all subsequent survey questions. In other words, just like in social conversations, the details of the survey conversation should have lasting effects on respondents' behavior during the survey interview. If one were to observe a conversation partner yawning, it is fair to say that this detail would not only affect one's immediate reaction, but also the remainder of the interaction. On the positive side, individuals who say something smart at one time may be given the benefit of the doubt in further interactions, even if they may seem to say something unintelligent down the line.

Previous research has shown how the rules of conversation guide respondents' responses to unclear questions (e.g., Schwarz, 1996). However, if a particular rule has either been broken or actively enforced in one specific instance in the survey, there is no reason to believe that this incident is encoded as relevant information only for some selected items. There is an abundance of results showing that individuals rely on minimal cues (such as facial traits) in judging others (in occurrence political candidates, e.g., Hall, Goren, Chaiken, & Todorov, 2009) in the span of milliseconds. Automatic overarching evaluations are part of social interactions, and small details can play an important role.

In this paper we propose that respondents take cues from how researchers behave with respect to the rules of conversation in a survey and adapt their subsequent response behavior and overall survey evaluations accordingly. To test this expectation, we conduct two survey experiments in which we manipulate response options and instructions to actively enforce the maxims of conversation in one survey section. We then observe respondents' response

optimizing behavior to questions in subsequent sections. We examine four empirical indicators of data quality: (1) correlations between sets of items on a new topic, (2) time spent per item following the manipulation, (3) end-of-survey evaluations measuring respondents' satisfaction with the response options provided during the entire survey, and (4) attention paid to questions throughout the survey.

The Rules of Social Conversation and Survey Data Quality

Compliance with the rules of social conversation is part of the standard advice when it comes to survey design (e.g., Tourageau et al., 2000, Bradburn, Sudman, & Wansink, 2004). In line with the initial formulation of the conversation logic by Grice (1975), the structure of the interaction between respondents and surveyors has been shown to be influenced by at least four conversational maxims: manner, quantity, quality, and relation (Clark & Schober, 1992; Schwarz, 1994; 1996; Schober, 1999). In the survey setting, the maxim of manner implies that all information provided by researchers should be comprehensible by respondents. The maxim of quantity requires researchers to provide neither more nor less information than is needed to accurately answer the questions. The maxim of quality requires that all the information provided to the respondents be accurate and truthful. Finally, the maxim of relation implies that all the information provided and asked for in the survey be relevant to the purpose of the survey. The amount of information provided by survey designers encompasses introductions, instructions, question wording, scales, question order, and visual design.

Work by Schwarz and colleagues (e.g., Hippler & Schwarz, 1989; Norenzayan & Schwarz, 1999; Schwarz 1994; 1996; 2007; Schwarz, Grayson & Knäuper, 1998) illustrates the applicability of the conversational maxims in several survey situations. For example, people's willingness to offer opinions about fictitious attitude objects is predicated on their assumptions

of the survey designers' truthfulness and earlier provision of information (i.e., the maxim of quality and quantity; Schwarz, 1995). Scale effects in frequency questions can be explained by a similar reliance on the maxims of quality and quantity (Schwarz et al., 1998): Respondents assume that scales reflect underlying distributions of responses known to the designers and consequently adapt their frequency reports in light of this inference. In line with the maxim of quantity, respondents can use even seemingly irrelevant information, provided at an introduction stage of the survey, to guide their responses to particular questions; Norenzayan and Schwarz (1999) and Galesic and Tourangeau (2007) show that people adapt their responses as a function of who is responsible for the survey. Importantly, when respondents can observe that the conversational maxims do not apply in the survey context, such response biases are lessened (Igou & Bless, 2007; Schwarz, 1995; Zhang & Schwarz, 2012).

While conversational rules are difficult to always uphold, evidence obtained in lab settings suggests that, by and large, people give researchers the benefit of the doubt in this respect when encountering an item they don't know how to answer (Zhang & Schwarz, 2012). However, this may not hold in mass surveys, particularly when respondents feel they do not know how to answer the question because the response options are too restrictive, the question is poorly formulated, or it is unreasonably difficult. Moreover, previous research has considered only some aspects of respondents' response behavior. For example, in the case of the fictitious attitude object question, the focus has been on understanding why people give an opinion but not on what information they derive from having been asked this difficult question in the first place or how they use this for the rest of the survey.

In short, survey researchers, by-and-large, agree that respondents interpret the meaning of questions in context (e.g., Hippler & Schwarz, 1989; Schwarz et al., 1998; Schober, 1999;

Smyth, Dillman, & Christian; 2007; Tourangeau, Couper, & Conrad, 2004). However, in previous studies, the focus has been on how conversational rule breaking affects the interpretation of specific items. Nevertheless, it is likely that people encode survey design features on a more general level too. They may use them to pass a judgment of (good/poor quality) the survey designers, or they may encode them on an affective level (i.e., they may remember the frustration or, alternatively, the pleasure of answering certain items). The durability of the cues people derive from the survey's design is still an extant question. We provide a first indirect test of the expectation that survey features are interpreted within a more general framework and not just related to particular items. While we do not directly measure affective states or social judgments, we examine how response behavior changes after conversational rule breaking.

Offering a DK Option: a Conversational Rule Perspective

Whether people have opinions on all political items in surveys has been a matter of debate for decades. Converse's seminal article (1964) provided evidence that many individuals may not hold a stable opinion on public policy issues. Zaller and Feldman (1992) showed that individuals are sensitive to the question format. In addition, as mentioned earlier, survey respondents have even been shown to express opinions on bogus items on which they could not reasonably have formed opinions. One explanation is that they assume that certain conversational rules hold (Schwarz, 1995), and, if given the option to respond "don't know" (DK) to an item, a sizable share of them do so (Bishop, 2005). In line with the conversational logic, previous studies found that respondents are more reluctant to give an opinion if the response instructions imply that such an opinion requires a lot of prior thought (Hippler & Schwarz, 1989). Therefore, there is evidence to argue that people do not have set opinions on all

issues. If given the chance to say they don't know, many will do so. And because they do not have a set opinion, they will be more sensitive to the question format and other context features. The basic logic guiding the use of DK options in surveys states that they should be used when researchers expect non-attitudes. In our view, this is largely unknown for many questions included in social science surveys. Thus, using DK options in many cases remains a choice to be made, and practitioners have been encouraged using them for improved data quality (e.g., Converse & Presser, 1986).

Recent reviews (Krosnick, 1999; Krosnick & Presser, 2010) summarize the evidence on opinionation and arrive at a different conclusion. People are more likely to give an opinion when they feel strongly about the object of the question, when they are motivated, and when they have the cognitive abilities. Moreover, people who say DK once may express an opinion later on if pressed to do so, and that opinion can be a strong predictor of other behavior or attitudes (Gilljam & Granberg, 1993; Krosnick et al., 2002). The underlying mechanism for why DK responses may mask reasonable attitudes is related to the psychological demands of the survey response. Answering questions reliably can be cognitively burdensome, as individuals must understand the question, retrieve relevant information, make a summary judgment, and map that judgment on a scale. Faced with an effortful task, someone may indicate that they have no opinion even if, in reality, they could articulate one from the available questionnaire options with a bit more effort. Thus, the availability of a DK option may also offer an opportunity to not reveal one's true preferences.

The choice of omitting a DK option is also based on some implicit assumptions about the maxims of manner and quantity, and it is on this underlying social conversation logic that we focus here. Not offering a DK option assumes that both the question object and the scale are

formulated in such terms to be comprehensible to all respondents. It also assumes that in the true state of the world, if people search enough, they will be able to map their attitude on the scale provided. If either one of these assumptions is faulty, but respondents don't realize it, we know, based on prior studies, that the specific items lacking a DK will be prone to measurement error. But what if they believe the maxims of conversation have not been observed?

There are several reasons for why respondents may perceive the lack of DK as a violation of conversational rules. First, the array of opinions that is often asked in surveys is sufficiently large to make it very likely that individuals will not have an opinion on all the items. Assuming even that a highly informed individual may have a (even weak) opinion on all of them, the scales might be biased, or individuals may not know how to place themselves accurately on each and every one of them. Allowing for the possibility of DK is therefore an implicit acknowledgement of a social fact: There may be individuals who are not know-it-alls, both among the respondents and among the researchers. Second, treating all questions in a survey alike (by not offering a DK option) sends an implicit signal of equality among the questions; there is an understanding that no matter the question, the individual possesses an answer for each and every one of them. When questions are general and/or tapping into strong attitudes, this equality may hold, but the assumption is counter-intuitive when comparing specialized and general questions. In other words, by treating all questions alike and not providing a DK option, researchers may implicitly send a message that they fail to acknowledge that different standards of opinionation apply for different objects, even when this difference is intuitive. Finally, by pressing individuals to give an opinion when they find it difficult (for social or cognitive reasons) to do so amounts to applying pressure on them that in social interactions is often present in relations of power, and it is possible that respondents might feel uncomfortable with this relationship.

The choice of not providing a DK option for questions about obscure or complicated objects (on which respondents may reasonably not have a readily-available attitude) could therefore be construed as a violation of the maxims of quantity and manner, especially if people are not instructed to think more in depth about these objects (e.g., Hippler & Schwarz, 1989). If people perceive the absence of DKs as a conversational rule violation, then we would expect them to update their subsequent response behavior accordingly. More specifically, they should lose some of the motivation to answer subsequent questions accurately, and they should evaluate the overall survey more negatively.

Giving Respondents Instructions: a Conversational Rule Perspective

Sincerity is usually appreciated in social conversations. But what happens when researchers are upfront about the demands that a survey section will put on respondents and include a vignette asking them to invest their time and effort in providing responses as accurate as possible? This declaration will alert some respondents to the difficulties of the survey. Being upfront about the difficulties of the task at hand could signal increased professionalism on behalf of the researchers, as it demonstrates awareness of the potential challenges respondents could face. Thus, this declaration can be construed as a positive enforcement of the maxim of manner. However, for others who intend to do precisely this—answer to the best of their abilities, despite being aware of the task difficulty—this sincere, advanced warning might be redundant and perceived as a signal that the researchers doubt their intentions of answering accurately.

Asking people upfront to give better answers means implicitly that some of them would otherwise be tempted to do the opposite. The extent to which this can be a positive social signal (and not a violation of the maxim of quantity) depends therefore on the task at hand. If the questions following this demand are comparatively speaking “easy,” such as media usage or

ratings of the most prominent politicians, this demand for respondent commitment might be construed as doubting their intentions. But if the questions following the vignette are more difficult to answer, then such instructions could be construed as a friendly warning. Since we want the vignette to be construed as a friendly warning, we place it before a battery of questions tapping into attitudes that are less crystalized, concerning more unusual political objects.

Our interest is in what the inclusion of this warning does to respondents' response strategy for the rest of the survey. If this is a positive social signal (of sincerity or increased professionalism of the researchers), then people should try to reward it by maintaining their focus longer than for the items that it was intended.

Empirical Expectations

Our main expectation is that compliance with the maxims of conversations is encoded as relevant general information which respondents make use of later on in the survey. Moreover, we expect these context-based inferences to direct not as much what particular information is retrieved later on, but rather respondents' motivation to optimize their responses later on in the survey.

We identify four observable indicators of response strategy and overall survey experience. The first indicator is the correlation between conceptually-similar items asked on different pages after a conversational rule has been tampered with. If being given the opportunity to say DK and exposure to the vignette is each construed as compliant with the conversational norms, and if this signal affects response strategy later on, then correlations between conceptually-similar items asked later should be stronger. The second indicator is the time taken to respond to an item. Longer question response latencies have been found to indicate response optimizing (e.g., Callegaro, Yang, Bhola, Dillman, & Chin, 2009) and are consistent with respondents going

through all the steps required by survey response theories (e.g., Tourangeau et al., 2000). Thus, if people interpret the presence of the DK option and the vignette in line with conversational norms, and they adapt their behavior, then they should spend more time on later survey items. A third indicator is the overall survey evaluation: the quality of the survey should be rated higher by those who got the warning and the opportunity to express a lack of opinion on difficult questions. Finally, we also examine the self-evaluated level of attention throughout the survey: Respondents who received either one of the manipulations should rate their attention higher.

The Experimental Design and Data

We test the expectation that violations of conversational norms have lingering effects using two web survey experiments. Both experiments were conducted through the Laboratory of Opinion Research (LORE) at the University of Gothenburg. All respondents were voting-age Swedish residents, and the surveys were conducted in Swedish. The experiments were each embedded in longer surveys. We discuss each of them in turn.

The First Experiment

The experiment was conducted in December 2013 (N = 1503, average age = 51.6, percent females = 48.7, percent with university degree = 41). It had a participation rate of 75% when computed according to RR6 / COMR standards (The American Association for Public Opinion Research, 2015) and consisted entirely of respondents from a probability based web panel (Martinsson, Andreasson, Markstedt, & Riedel, 2013). Field work was closed after 28 days, and one reminder was sent. The experiment followed an unrelated study in the same survey. It had a randomized 2 (*DK options:* no vs. yes) × 2 (*Response instructions vignette:* no vs. yes) between-subjects design, as depicted in Figure 1.

[*Figure 1.* Two-by-two (vignette by "don't know" (DK) response option) experimental design used in both studies reported in this article.]

The 'DK options' factor in the first experiment. We included or omitted a DK response option to 16 questions, all of them on one page. The items were adapted from existing national surveys. The topics were dominated by social and economic issues. The list of questions is provided in Appendix 1.[1] Illustrations of how the survey pages with and without the DK options looked are provided in Appendix 2.

To actively enforce the maxim of quantity, we designed two questions on which a DK option would map true responses for many respondents. For one, respondents had to evaluate their trust in the Fortifications Agency, an obscure governmental agency. Another item asked for opinions on a policy proposal to create spaces where the young and the old could meet. The difficulty of this item came from the absence of concrete examples of such "spaces."

The "response instructions vignette" factor in the first experiment. Prior to the 16-items page, we either showed or omitted an instructional vignette urging respondents to try their best at answering these items, even if some may be difficult. The vignette was balanced in its discussion of response difficulty, to include both the possibility that people might not have an opinion and the invitation to think carefully about each question on the following page before answering. It read [translated]:

> "On the following page you will receive questions about your opinion of different things. We appreciate if you would take time to read these in peace and quiet and carefully consider what your view is on each question. Some people have a very clear opinion on certain questions, but this is not the case for everyone. Sometimes people feel that they have carefully pondered a question but that they still lack a clear opinion. Other times they may after careful consideration find that there is a response option that is closest to their

> view. Please consider each response option carefully before answering each question."

Dependent outcomes in the first experiment. Our first outcome of interest is the correlation between conceptually similar items measured on survey pages after the treatment (and without a DK option). We choose to focus on environmental issues as this topic was only marginally addressed earlier.[2] Interest in environmental issues was measured as part of a larger battery on the first page after the DK manipulation. On the following page, we asked how often respondents performed various environment-friendly activities (six items). If survey design features have a lingering effect on response behavior, then we should observe more optimizing-consistent behavior in the groups that previously received the DK and/or the vignette treatment: In these groups the correlation between interest in the environment and pro-environmental activities should be higher.

Our second dependent outcome is the amount of time individuals spend per item.[3] The vignette treatment was administered immediately prior to the start of the DK treatment, therefore we look at the evolution of the response time on subsequent pages with and without DK. Strictly speaking, the vignette only cautioned individuals for the immediately following page, so any effects we observe on the time taken to answer items on later pages suggests that respondents encoded the message at a more general level.

On the fourth page after the DK factor, we measured two additional dependent outcomes: respondents' subjective evaluation of the suitability of the response options provided in the survey overall and their subjective attention levels throughout the survey. The exact question wordings are provided in Appendix 1. If the survey features consistent with the respect of conversational maxims have lingering effects, then the groups that received the DK option and/or

the vignette treatment should evaluate the suitability of the survey options and their own attention higher.

The Second Experiment

The second experiment was fielded in March 2014 (*N* = 2489, average age = 50.3, percent females = 43, percent with university degree = 52). It had a participation rate of 54% when computed according to RR6 / COMR standards (The American Association for Public Opinion Research, 2015) and consisted mainly of opt-in respondents (Martinsson, Andreasson, Markstedt, & Riedel, 2014). Field work was closed after 32 days and two reminders were sent. The experiment started on the twelfth page of the survey. It also had a fully randomized 2 (*DK options:* no vs. yes) × 2 (*Response instructions vignette:* no vs. yes) between subjects design, as in Figure 1.

The "DK options" factor in the second experiment. We included or omitted a DK response option to 30 items arranged on four pages. The first 14 items were very similar to the first experiment, but were split on two pages. We included the same items on the Fortification Agency and the "meeting spaces" as in the first experiment. The third page contained 10 items from the European Elections Survey. These items probed for the responsibility of the European Union and the national state for various outcomes, and the DK option was deemed as a potentially relevant response alternative given the topic narrowness. The fourth page contained six items from the human values question battery (Schwartz, 1994), and the DK option was deemed necessary due to the unusual response scale (running from -1 to +7).

The "response instructions vignette" factor in the second experiment. Prior to the first DK factor page we included or omitted a vignette identical to the one used in the first experiment.

Dependent variables in the second experiment. Similar to the first experiment, our first dependent outcome in this case is the correlation between the conceptually similar attitudinal and behavioral items asked after the experimental treatment and without a DK option. After the values-related questions, respondents evaluated their willingness to perform various activities for environmental purposes (three items) and, on a subsequent page they rated pro-environmental policy proposals (three items). The full wording is provided in Appendix 1. We expect this correlation between attitudes and behavior to be stronger if respondents received the DK and/or the vignette treatment.

Our other dependent outcomes are similar to the first experiment: time per item following the experimental treatment, survey satisfaction, and survey attention measures. The latter items were identically worded as in the first experiment. They were placed on the seventh page after the end of the DK factor. On all these measures, we expect more positive reactions from the treated groups.

Experimental Manipulations and Conversational Rules: A Validation Check

The inclusion of a DK option and the vignette in the experimental treatment were intended as an enforcement of the conversational maxims or manner and quantity because we expected people to find it hard to have an opinion on all the items (due to the lack of familiarity, lack of question clarity, topic specificity or to the scale construction). While we did not ask respondents directly whether they thought a DK option or the vignette wording were conversationally justified, the statistics in Table 1 serve as a validation check for our manipulations. Table 1 presents the proportions of DK and mid-point responses in the treatment part of both the first and the second study, by experimental condition and question object. If the inclusion of a DK is justified, then people with no DK alternative should have a higher propensity to choose the

midpoint option for difficult items. If the inclusion of the vignette is justified, then people who have the opportunity to say DK for these difficult items should do so even more after exposure to the vignette.

Confirming our expectations, people choose midpoints less when DK is available (in both experiments, and for all question batteries, except the values one). The decrease in midpoint selection is compensated almost entirely by the increase in DK use when such an option exists. For example, the proportion of midpoint use on the Fortifications Agency item decreases significantly with the availability of DK ($z < -10.00$, $p = 0.00$ in both studies). With respect to the vague "spaces for young and old" policy, those who received both the vignette and the DK are less likely to select the midpoint than those who received no experimental treatment ($z < -2.32$, $p < 0.02$ in both studies). In fact, as expected, among those with a DK option, the likelihood of stating no opinion increases with exposure to the vignette, for both items in both studies (three out of the four *z*-statistics are significant at $p = 0.10$ or better). We conclude that both the DK and the vignette were needed for more accurate responses.

[*insert Table 1 about here*]

Post-Manipulation Response Behavior: Results

Table 2 presents our first main result:[4] the impact of the compliance with the conversational rules on the correlations between conceptually close items asked after the experimental manipulation. The top part presents the correlation between environmental interest and reported pro-environmental behavior in the first study, by experimental group. As can be seen, this correlation is greater in all groups that received some form of experimental treatment. Compared to the group that received no vignette and no DK option, the correlation is 32% higher in the group that received both the DK option and the vignette (0.57 compared to 0.43, Jennrich

χ^2 (1) = 6.16, p = 0.05).[5] The bottom part of the table presents the correlations between environmental attitudes and behavior in the second study by experimental group. Compared to those who received no DK option and no vignette, the correlation in the group that received only the vignette treatment is 30% higher (0.56 compared to 0.43, Jennrich χ^2 (1) = 8.43, p = 0.01). Among those who received both treatments the correlation is 20% higher (0.52 compared to 0.43, Jennrich χ^2 (1) = 3.61, p = 0.057). Overall, experiencing some differences in survey design features has quite a sizable impact on further survey response behavior and hence on our estimates.

[*insert Table 2 about here*]

Table 3 presents the evolution of the mean question response time in both experiments,[6] in seconds, starting with the questions placed immediately after the vignette. As can be seen, in both experiments respondents in the treatment groups (with the exception of the DK-only group in the first experiment) spend more seconds per item than those who received no DK and no vignette. In both cases, the effect is stronger for those who received both the vignette and the DK. In the second experiment, the group that received both the DK and the vignette maintains a significantly higher time per question for another five pages (the last two without a DK option). In the first experiment this group also has a higher time per question in the two subsequent pages, but the difference is not significant. Thus, the results in both studies strongly point in the same direction, suggesting that respondents interpret the availability of the vignette and the DK option as an encouragement to optimize their responses to subsequent pages, not just for the one specified in the text.[7]

[*insert Table 3 about here*]

Table 4 presents the mean satisfaction with the survey response options (in the top part) and the mean attention paid to the survey (the bottom part) in each experimental group and in each experiment.[8] Overall, individuals give quite high ratings in both experiments and on both variables, but these ratings are even higher in some groups. Looking first at the general satisfaction with the response options, we observe a significant increase in ratings among those who received both the DK and the vignette treatment in the first experiment compared to those who received no treatment (0.728 compared to 0.692, with the difference significant at $p = 0.05$). A similar trend appears in the second experiment (0.635 compared to 0.615), but this difference only approaches significance at conventional levels ($F(1, 2325) = 2.48$, $p = 0.115$ two tailed). With regard to the subjective attention paid to survey, this is higher among those who received the vignette in the second experiment (0.82 and 0.83), as compared to those who received no treatment (0.79, $F(1, 2324) > 15.59$, $p < 0.01$, for the differences from the control group). In the first experiment the attention estimates do not vary by group. One explanation for the stronger results in the second experiment may lie with the time in between the vignette and the evaluations: In the first experiment, the vignette and the evaluation question were separated by four pages, and respondents had completed a lengthier study before. Thus, it is possible that they may have considered the section containing our manipulation to be too small a part of their experience. In contrast, in the second study, the vignette and the evaluations were separated by eleven pages worth of questions, representing half of the survey. In short, these results suggest that people did take the vignette as a general attention warning, and not just for a single page.

[*insert Table 4 about here*]

The overall conclusion of these results is, we believe, that survey design choices do impact respondents' strategy to optimize later and their satisfaction with the survey. Indicative of such a

strategy, correlations between conceptually related items on a different topic are stronger for the treatment groups and the time spent on subsequent pages is generally higher. Moreover, two other indications that respondents interpret the early survey design features as part of a larger conversation emerge: Subjective perceptions of attention to the survey and satisfaction with the adequacy of survey options in the entire survey are also greater when people have the opportunity so say DK for some items and are made aware of the difficulty of answering them.

Discussion

Our results are consistent with the idea that respondents take cues from the survey design, and, akin to social conversations, use them to make inferences about the quality of the survey (and/or about the quality of researchers as conversation partners). When these design choices are consistent with the rules of social conversations, they perform better throughout. In other words, the results provide evidence that the survey context affects not just what information people bring to mind when they encounter a certain detail, but also their willingness to optimize their responses in its aftermath. Previous research has drawn attention to the conversational asymmetry (Schober, 1999) inherent in surveys: unlike in social conversations, in surveys people rarely have the chance to ask for additional clarifications if they don't understand a question, or if they believe the response options are too restrictive. We believe the insights brought about by our results are particularly important for self-administered surveys as in these surveys the conversational asymmetry is even greater than in other modes such as telephone interviews or face-to-face surveys.

As this is a first study, a number of details remain to be filled. First, we do not test the mediator (or the mechanism) behind these effects. This would have required a different experimental design (Imai, Keele, Tingley, & Yamamoto, 2011). Thus, we do not directly test

the claim that individuals take specific design choices as cues of a general *survey* (or *researcher*) *quality*; in other words, that survey design choices lead to social inferences. Nor do we directly test the possibility that the mechanism responsible for the impact of early survey experiences on later survey performance is *affect*-infused. Individuals might become frustrated by certain choices early on, which in turn puts them in a more negative mood later (but not sufficiently negative for them to break completely with the survey); alternatively, individuals might be positively surprised by certain early features, which improves their mood later on. Presumably, in a social conversation, upon observing the conversation partner behaving in a socially less acceptable way, one experiences both negative emotions, *and* makes inferences about the partner's personality, intentions etc. The mechanisms are both cognitive and affective. Our results indicate that when a survey respects the conversational norms, this affects respondents at a different level than what specific information they retrieve. It affects their motivation to optimize later on for new items and their satisfaction with the survey as whole. The specific cognitive or affective reasons underpinning this impact however, should be determined by future studies.

As a future avenue of research into the mechanism responsible for the lingering effects of design choices, it will be important to measure affective reactions to survey design. This is difficult, as much of the data we have about survey experience rely on self-reported evaluations by respondents, and therefore, these data have a dominant cognitive component. Other sources, such as advanced forms of paradata including physiological reactions of respondents during the survey may be one way to get evidence of the emotional impact, but this is hard to implement at a larger scale.

If surveys are akin to social conversations, a number of claims could be tested further. First, a number of individual-level variables could moderate the effects of early design choices. Individuals who are less socially trustful should be more sensitive to survey design choices. Less knowledgeable individuals should react more strongly when researchers actively comply with social conversation norms, because they are more likely to satisfice in the first place. Our sample, being highly knowledgeable and socially trustful in both studies, prevents us from fully testing this possibility. Second, the impact of design choices could vary with the type of items included in the survey both early and later on. If a DK option or a difficulty warning is not needed for the early items, because they tap into well-established attitudes or behaviors, then having any of these features might be counterproductive. Alternatively, if the later items are easy to answer, the effects of any early survey design choice could be more muted. Third, having a DK option or a vignette are not the only choices that can have a lingering effect. Ideally, it would be necessary to test the subsequent impact of other choices that may go against social rules – like asking sensitive questions in the beginning of the survey, etc.

We believe that our experimental results point to a potentially important, yet previously overlooked aspect of the survey experience. If the survey interview is construed as a flow, then researchers' survey design choices should affect respondents' behavior not just through cognitive mechanisms (e.g. priming), but through other mechanisms as well, such as affect, or social inferences. It is generally considered common sense that much of what one says in a conversation will have social and affective implications not just immediately, but also later on. It is time to get a more precise understanding of these implications in the survey context.

References

Berinsky, A. J. (1999). The two faces of public opinion. *American Journal of Political Science, 43*(4), 1209-1230. doi: 10.2307/2991824

Bishop, G. F. (2005) *The illusion of public opinion: Fact and artifact in American public opinion polls*. Lanham, MD: Rowman & Littlefield Publishers.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004) *Asking questions: the definitive guide to questionnaire design: For market research, political polls, and social and health questionnaires*. San Francisco: Jossey-Bass.

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T. Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *Bulletin de Méthodologie Sociologique, 103*(1), 5-25. doi: 10.1177/075910630910300103

Clark, H. H., & Schober, M. F. (1992) Asking questions and influencing answers. In J. M. Tanur (ed.) *Questions about questions: inquiries into the cognitive bases of surveys* (pp. 15-48). New York: Russell Sage Foundation.

Converse, P. E. (1962). The nature of belief systems in mass publics. In D. Apter (Ed.) *Ideology and discontent* (pp. 201-261). New York: Free Press.

Converse, J. M., & Presser, S. (1986). Survey questions: Handcrafting the standardized questionnaire. Beverly Hills, CA: Sage.

Galesic, M.,& Tourangeau, R.(2007). What is sexual harassment? It depends on who asks! Framing effects on survey responses. *Applied cognitive psychology 21*(2), 189-202. doi: 10.1002/acp.1336

Gilljam, M., & Granberg, D. (1993). Should we take don't know for an answer? *Public Opinion Quarterly 57*(3), 348-357. doi: 10.1086/269380

Hall, C. C., Goren, A., Chaiken, S. & Todorov, A. (2009) Shallow cues with deep effects: Trait judgments from faces and voting decisions. In E. Borgida, J. L. Sullivan, & C. M. Federico (Eds.) *The political psychology of democratic citizenship* (pp. 73-99). Oxford: Oxford University Press.

Hippler, H. J., & Schwarz, N. (1989). 'No opinion'-filters: A cognitive perspective. *International Journal of Public Opinion Research*, *1*(1), 77-87. doi: 10.1093/ijpor/1.1.77

Igou, E. R., & Bless, H. (2007). Conversational expectations as a basis for order effects in persuasion. *Journal of Language and Social Psychology 26*(3), 260-273. doi: 10.1177/0261927X06303454

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(04), 765-789. doi: 10.1017/S0003055411000414

Krosnick, J. A. (1999). Survey research. *Annual review of psychology 50*(1), 537-567. doi: 10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.) *Handbook of survey research* (pp. 263-313). West Yorkshire: Emerald Group.

Krosnick, J. A., Holbrook, A. H., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., … Conoway, M. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly 66*(3), 371-403. doi: 10.1086/341394

Martinsson, J., Andreasson, M., Markstedt, E., & Riedel, K. (2013) *Technical Report Citizen Panel 8 – 2013*. Gothenburg: University of Gothenburg, LORE.

Martinsson, J., Andreasson, M., Markstedt, E., & Riedel, K. (2014) *Technical Report Citizen Panel 9 – 2014.* Gothenburg: University of Gothenburg, LORE.

Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology 29*(8), 1011-1020. doi: 10.1002/(SICI)1099-0992(199912)29:8<1011::AID-EJSP974>3.0.CO;2-A

Schober, M. F. (1999). Making sense of questions: An interactional approach. In M. G. Sirken, D. J. Hermann, S. Schechter, N. Schwartz, J. M. Tanur & R. Tourangeau (eds.), *Cognition and survey research* (pp. 77-93). New York: John Wiley & Sons.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues, 50*(4), 19-45. doi: 10.1111/j.1540-4560.1994.tb01196.x

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M. Zanna (Ed.) *Advances in experimental social psychology* (pp. 123-162). San Diego, CA: Academic Press.

Schwarz, N. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. *International Statistical Review/Revue Internationale de Statistique 63*(2), 153-177. doi: 10.2307/1403610

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

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology 21*(2), 277-287. doi: 10.1002/acp.1340

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*(2), 177-183. doi: 10.1093/ijpor/10.2.177

Smyth, J. D., Dillman, D. A., & Christian, L. M. (2006). Context effects in Internet surveys: New issues and evidence. In A. N. Joinson, K. Y. A. McKenna, T. Postmes & U-D. Reips (Eds.) *Oxford handbook of Internet psychology* (pp.429-446) Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199561803.013.0027

The American Association for Public Opinion Research. (2015). *Standard Definitions : Final Dispositions of Case Codes and Outcome Rates for Surveys. 8th edition*. AAPOR.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly 68*(3), 368-393. doi: 10.1093/poq/nfh035

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science 36*(3), 579-616. doi: 10.2307/2111583

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51-68. doi: 10.1002/acp.1331

Zhang, Y. C., & Schwarz, N. (2012). How and why 1 year differs from 365 days: A conversational logic analysis of inferences from the granularity of quantitative expressions. *Journal of Consumer Research 39*(2), 248-259. doi: 10.1086/662612

Figure 1
Two-by-two (vignette by "don't know" [DK] response option) experimental design used in both studies reported in this article

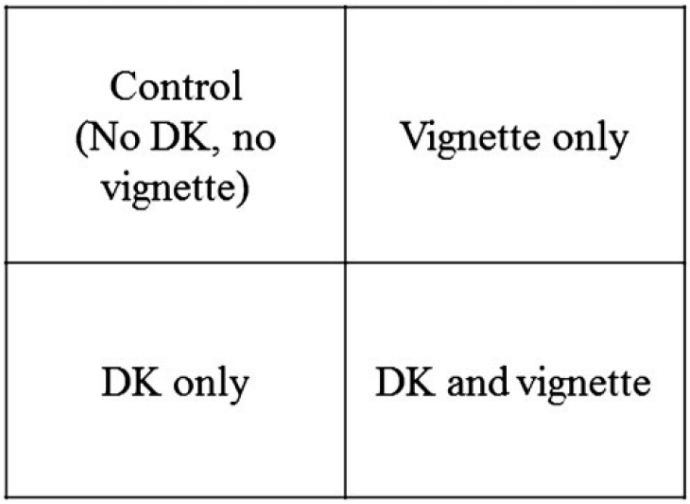

Table 1
Descriptive Statistics of Response Behavior for the Treatment Items, by Study and Experimental Group

Experimental groups	Study #1									
	All treatment items		Policy items		Specific item: Young-old policy		Trust in institutions items		Specific item: Fortification agency	
	Mid-point	DK	Mid-point	DK	Mid-point	DK	Mid-point	DK	Mid-point	DK
Control										
M	0.31	–	0.23	–	0.14	–	0.40	–	0.74	–
SD	0.16	–	0.16	–	0.35	–	0.26	–	0.44	–
N	369	–	369	–	366	–	369	–	366	–
Vignette only										
M	0.30	–	0.21*	–	0.17	–	0.39	–	0.72	–
SD	0.15	–	0.14	–	0.38	–	0.25	–	0.45	–
N	369	–	369	–	367	–	369	–	358	–
DK only										
M	0.22***	0.10	0.18***	0.04	0.13	0.01	0.27***	0.15	0.34***	0.43
SD	0.15	0.12	0.14	0.09	0.34	0.12	0.24	0.20	0.47	0.50
N	372	372	372	372	372	372	372	372	367	367
DK and vignette										
M	0.21***	0.11	0.16***	0.05	0.08***	0.06	0.27***	0.17	0.29***	0.47
SD	0.14	0.13	0.12	0.09	0.26	0.24	0.23	0.22	0.46	0.50
N	370	370	370	370	370	370	370	370	370	370

continued

Table 1
Continued

Experimental groups	Study #2													
	All treatment items		Policy items		Specific item: Young-old policy		Trust in institutions items		Specific item: Fortification agency		Gvt and EU responsibility items		Values items	
	Mid-point	DK	Mid-point	DK	Mid-point	DK	Mid-point	DK	Mid-point	DK	Mid-point	DK	Mid-point	DK
Control														
M	0.23	–	0.22	–	0.20	–	0.37	–	0.72	–	0.12	–	0.20	–
SD	0.11	–	0.16	–	0.40	–	0.24	–	0.45	–	0.14	–	0.21	–
N	598	–	601	–	599	–	600	–	593	–	600	–	598	–
Vignette only														
M	0.23	–	0.22	–	0.20	–	0.37	–	0.74	–	0.12	–	0.21	–
SD	0.11	–	0.15	–	0.40	–	0.24	–	0.44	–	0.15	–	0.21	–
N	586	–	589	–	588	–	588	–	575	–	587	–	586	–
DK only														
M	0.18***	0.07	0.18***	0.06	0.18	0.03	0.24***	0.14	0.30***	0.47	0.10	0.06	0.20	0.01
SD	0.10	0.08	0.15	0.10	0.39	0.18	0.21	0.18	0.46	0.50	0.13	0.14	0.20	0.09
N	571	571	577	577	576	576	576	576	573	573	574	574	573	573
DK and vignette														
M	0.18***	0.07	0.16***	0.06	0.15***	0.05*	0.24***	0.17**	0.25***	0.55***	0.10	0.06	0.20	0.01
SD	0.10	0.08	0.14	0.11	0.36	0.23	0.21	0.18	0.43	0.50	0.14	0.15	0.20	0.05
N	586	586	591	591	591	591	591	591	588	588	587	587	587	587

Note. Mean proportion (*M*) of mid-point and DK use (by item topic), with standard deviations (*SD*) in each experimental group in the treatment section. In the first study the treatment section consisted of 16 items (10 policy items, 6 institutional trust items). In the second study it consisted of 30 items (8 policy items, 6 institutional trust items, 10 responsibility items, 6 value importance items). *** indicates a significant difference from the control group at $p < .01$ (two tailed). ** indicates a significant difference from the control group at $p < .05$ (two tailed). * indicates a significant difference from the control group at $p < .10$ (two tailed). DK = "Don't Know" response choice included.

Table 2
Posttreatment Bivariate Correlations Between Conceptually Similar Items, by Experimental Study and Experimental Group

Experimental groups	Experiment #1		
	DV: Reported environmental behavior		
	IV: Interest in environmental issues		
	Pearson *r*	*p*	*N*
Classic (no vignette, no DK)	**0.429**[a]	.000	364
Vignette only	0.502	.000	370
DK only	0.477	.000	368
DK and vignette	**0.566**[a]	.000	370
Experimental groups	Experiment 2		
	DV: Reported environmental behavior		
	IV: Environmental attitudes		
	Pearson *r*	*p*	*N*
Classic (no vignette, no DK)	**0.432**[b,c]	.000	596
Vignette only	**0.560**[b]	.000	585
DK only	0.494	.000	572
DK and vignette	**0.518**[c]	0.000	584

Note. All the variables are scaled from 0 to 1. In both studies, the dependent variable is based on multiple indicators combined into a scale of environmental behavioral intentions (see the Appendix 1 and text for details). In the first study, interest in environmental issues is measured by one item as part of a more extended topic interest battery. In the second study, the environmental attitudes variable is a scale of pro-environmental policy opinions, as discussed in the text. Correlations with common superscript letters indicate a statistically significant difference. a or b indicate a significant difference at $p < .05$, c indicates significant difference at $p < .10$, two tailed. DK = "Don't Know" response choice included.

Table 3
Treatment Effects on Time Spent per Question (in seconds)

	Experiment 1:				Experiment 2:						
	Survey timeline → → →				Survey timeline → → →						
	↓ Vignette	DK page	Post-DK page 1	Post-DK page 2	↓ Vignette	DK page 1	DK page 2	DK page 3	DK page 4	Post-DK page 1	Post-DK page 2	Post-DK page 3
Experimental groups												
Control group												
M		8.02	4.43	5.79		8.39	5.86	9.51	10.85	6.58	7.92	3.99
SD		3.64	1.84	2.13		3.46	2.71	4.15	5.04	2.84	3.38	1.60
N		350	359	339		571	571	568	571	568	565	574
Vignette only												
M		9.98***	4.37	5.96		9.28***	6.12	9.56	11.07	6.53	8.05	4.03
SD		3.94	1.73	2.15		3.81	2.63	4.08	4.85	2.65	3.19	1.58
N		340	350	353		548	553	553	556	543	561	554
DK only												
M		8.19	4.38	5.77		9.09***	6.06	9.71	11.56**	6.67	7.98	4.11
SD		3.60	1.71	2.20		3.53	2.61	4.14	5.12	2.87	3.41	1.60
N		362	358	358		549	542	540	546	545	542	540
DK and vignette												
M		10.44***	4.56	5.90		9.97***	6.55***	10.25***	11.46**	6.94**	8.27*	4.13
SD		3.77	1.76	2.08		3.73	2.77	4.13	4.91	2.90	3.17	1.62
N		340	360	355		550	550	562	553	567	555	560

Note. We exclude cases with latencies higher than four times the size of the interquartile range from the analysis. *** indicates a significant difference from the control group at $p < 0.01$ (two tailed). ** indicates a significant difference from the control group at $p < 0.05$ (two tailed). * indicates a significant difference from the control group at $p < 0.10$ (two tailed).

Table 4
End-of-Survey Evaluations by Experiment and Experimental Group

Experimental groups	Respondent satisfaction with the response options in the survey					
	Experiment #1			Experiment #2		
	M	*SE*	*N*	*M*	*SE*	*N*
Classic (no vignette, no DK)	**0.692**[a]	0.203	368	**0.615**[b]	0.224	598
Vignette only	0.708	0.180	370	0.630	0.219	583
DK only	0.715	0.186	369	0.629	0.223	566
DK and vignette	**0.728**[a]	0.177	370	**0.635**[b]	0.221	582
Experimental groups	Respondent subjective evaluation of attention to the survey questions					
	M	*SE*	*N*	*M*	*SE*	*N*
Classic (no vignette, no DK)	0.810	0.189	368	**0.783**[c, d]	0.196	596
Vignette only	0.820	0.172	370	**0.832**[d]	0.153	584
DK only	0.790	0.192	370	0.795	0.188	565
DK and vignette	0.808	0.185	371	**0.823**[c]	0.161	583

Note: Cell entries are means (*M*) and standard errors (*SE*). The variables are scaled from 0 to 1. Means with common superscript letters indicate a statistically significant difference. a, c, or d indicate significant difference at $p < .05$, b indicates significant difference at $p = .115$, two tailed. DK = "Don't Know" response choice included.

Endnotes

[1] Respondents were not required to answer all questions in any version of the questionnaire. They were allowed to proceed to the next survey page without responding to all questions if they wished to do so. The same response policy was applied in Study 2.

[2] In the first study, the environment was mentioned explicitly in just one out of 16 questions and was indirectly related to two others. In the second experiment, the environment was never mentioned explicitly and was related to the topic of three out of 30 questions.

[3] While response latencies have been shown to be influenced by such factors as age and expertise (Yan & Tourangeau, 2008), the randomized between group design in our case limits the influence of such factors on the group means. Moreover, as attitude accessibility for unfamiliar or complicated items is expected to be low (such as for obscure institutions and policies, but also for various specific behavioral reports), we should observe longer latencies if people think in depth about them.

[4] The full regression analyses are presented in Appendix 3.

[5] The test for the equality of correlations was performed using the mvtest correlations command in Stata 13.

[6] As response latency distributions can be affected by extreme outliers, we exclude those with latency higher than 4 times the interquartile range from the analysis. This transformation eliminated all extreme outliers identified by the iqr command in Stata 13.

[7] It could be argued that people who receive the DK spend more time per question because they have to read more. This explanation is however inconsistent with the design of the questions. All the questions on every page were arranged in batteries, thus the individual only had to read

"Don't know" once (which would take significantly less than the additional several seconds respondents in the cumulative treatment group spent per page).

[8] The full regression analyses are presented in Appendix 3.