

ARTICLES

Assessing Measurement Error in Hypothetical Questions

Adam Kaderabek¹ , Jennifer Sinibaldi² 

¹ Applied Research Business Development, RTI International, Division of Research Services, ² Social Science Research Institute, The Pennsylvania State University, University Park

Keywords: hypothetical questions, behavioral intention, measurement error, response bias, dependent interviewing

<https://doi.org/10.29115/SP-2022-0010>

Survey Practice

The hazard of asking survey participants hypothetical questions is that most often, responses are only an inferential vehicle for understanding preference and may not reflect the truth. Often it is necessary to make a reasoned assumption about the alignment of an individual's attitudes and behaviors. In this paper, we discuss results from a Survey of Doctorate Recipients (SDR) pilot study experiment that found significant response bias to hypothetical questions even when the hypothetical context was unambiguous and included salient properties which were familiar to respondents from prior experience. The experiment asked Treatment group participants to report their reaction to slight alterations in the SDR instrument; meanwhile, those in the Control group were asked hypothetically how they *would react* to the protocol if they had experienced the same changes. We identify the presence of response bias across 10 respondent experience questions using both categorical and quantitative methods. Our results show that the Control group consistently exhibited less preference for the hypothetical protocol than the Treatment group. We believe these differences are representative of the measurement error resulting from the poor correspondence between the Control group's hypothetical attitudes and the Treatment group's reported experience. This work shows that hypothetical bias is a concern in opinion research even when hypothetical conditions may be clearly understood by respondents and similar to their previous experiences.

Background

Methodological research has recognized the use of hypothetical questions as an effective means to investigate respondents' attitudes toward survey participation and data collection to make assumptions about their behavior (Hargittai et al. 2020; Keusch et al. 2019; Olson, Smyth, and Wood 2012; Pinter 2015; Revilla, Couper, and Ochoa 2019; Wenz, Jäckle, and Couper 2019). This relationship between attitude and behavior is particularly important to decision makers because the intentions of stakeholders may be the only data available when looking to implement new policies or protocols. For example, the U.S. Department of Health and Human Services Centers for Disease Control and Prevention's COVID-19 Rapid Community Assessment Guide (2021) includes, "*How likely are you to get a COVID-19 vaccine?*".

^a Corresponding author:
Adam Kaderabek
RTI International
3040 East Cornwallis Road
Haynes Bldg. 423H
Research Triangle Park, NC 27709-2194

Although this example is an important indicator of intention during the COVID-19 pandemic, research on hypothetical questions has identified several risks to response accuracy including social desirability bias (Norwood and Lusk 2011), the length of consideration (Cook et al. 2012) or the framing of consequence associated with the response (Carson and Groves 2007; Howard et al. 2017). But perhaps the most critical influences are task familiarity and context (Keusch et al. 2019; Schlöpfer and Fischhoff 2012; Struminskaya et al. 2020). Yet, even if these factors are (mostly) controlled, it is still challenging to establish a link between an attitude, expressed as an intention at $time_a$, and a behavior observed at $time_b$.

Often it is necessary to make a reasoned assumption about the alignment of an individual's attitudes and behaviors. By identifying a given attitude, and the likelihood that the attitude will contribute to a particular behavioral outcome, researchers can establish a measure of the respondent's intention (Bradburn, Sudman, and Wansink 2004, ch 4). For example, Keusch et al. (2019) presented participants with multiple hypothetical research studies involving passive data collection through a smartphone app and then asked their attitudes toward participating in each study. Based on the participants' responses, they assume under what conditions participants would cooperate. More recently, research has focused on respondents' attitudes toward downloading hypothetical COVID-19 tracking applications (Hargittai et al. 2020). These studies employ intention as a metric for behavioral outcomes by assuming attitudes and behaviors are congruent.

In this paper, we discuss results from an experiment that found significant response bias to hypothetical questions even when the hypothetical context was unambiguous and included salient properties which were familiar to respondents from prior experience. Specifically, some members of a longitudinal panel survey were surveyed using the standard version of the questionnaire that they were used to, while an experimental group received the traditional survey with slight alterations to the presentation of the questions. Upon completion of the survey, all respondents were asked to respond to several attitudes questions in a debriefing section, which we call *respondent experience questions* (REQs).

For all but ten REQs, the question wording was identical for all respondents and assessed respondent attitudes toward the survey (e.g., perceived speed and enjoyment). The remaining ten questions were dependent on group assignment and asked how the respondent either felt, or *would have felt* (i.e., hypothetically), about the modified survey format. Because the experimental format was a change to a salient survey feature and all respondents were equally familiar with the standard survey format from prior rounds, we expect any group differences in the ten conditional REQs to be indicative of response bias between respondents who reported their hypothetical attitudes versus those who reported their attitude based on experience.

Data/Design

The Survey of Doctorate Recipients (SDR), conducted by the National Center for Science and Engineering Statistics (NCSES), is a fixed-panel, longitudinal survey of doctorate recipients in science and engineering fields. The SDR covers recent employment, past employment, and related experiences such as training and educational experiences. The 2019 SDR sample included 120,000 cases and concluded with 80,882 respondents participating (AAPOR RR3 = 67.4%). The SDR has shown a notable trend toward completion via the Web (84% in 2017 and 93% in 2019). With an increasing proportion of the sample completing via a self-administered Web instrument, an experiment was fielded to test a Web-administered, dependent interviewing approach.

Dependent interviewing is the process of reiterating a respondent's prior survey responses during subsequent rounds of data collection and is common in panel surveys (Jäckle 2006). Dependent interviewing can reduce measurement error by reducing burden, aiding memory/recall, and providing panel participants continuity over the course of their enrollment (Pascale and Mayer 2004). The experiment was designed to test these advantages and capture respondents' attitudes toward the use of dependent interviewing.

The SDR pilot design used a stratified systematic random sample of 3,900 eligible respondents. 3,600 cases were selected from 2019 SDR respondents with an additional 300 cases that last participated in either 2015 or 2017. The sampled cases were randomly and evenly distributed into three experimental groups resulting in 1,300 sample cases for each group. The Control group (CTR) received the traditional (i.e., cross-sectional) version of the survey, and two types of dependent interviewing protocols (collectively called TRT) were tested in parallel. Minimal changes were made to the instrument in order to accommodate the dependent interviewing protocol and examples of survey questions for each condition can be found in Appendix A (supplement online). Our analysis focuses specifically on responses to the ten REQs introduced previously, and the differences observed between the attitudes reported by the CTR to the hypothetical modification and attitudes of the TRT who experienced the modification.

Respondent Experience Questions (REQs)

The REQ portion of the questionnaire was similar to a debriefing survey. All respondents were asked several identical questions about their experience with the survey; however, ten conditional questions were altered by group. The TRT respondents were asked direct questions about their experience, such as, *“Do you think that pre-filling some of your answers from <PRIOR CYCLE> and asking you to confirm or update them made this survey... Much more burdensome, A little more burdensome, Neither more or less burdensome, A little less burdensome, Much less burdensome.”* Members of the CTR group were asked a hypothetically equivalent question, i.e., *“If this survey had pre-*

filled your most recent answers and asked you to confirm or update them, would that have made this survey...” followed by the same set of response options available to the TRT group.

The first seven questions, which we call reaction questions, asked respondents to indicate the strength of their reaction (*annoyed*, *appreciative*, *comfortable*, *concerned*, *confused*, *surprised*, and *relieved*) to the experienced/hypothetical protocol using the response options: “Not at all,” “A little,” or “Very much.” The order of these seven questions was randomized for all respondents. The remaining three REQs, which we call efficacy (of TRT) questions, asked whether respondents felt the protocol would have a positive or negative influence on response accuracy or survey burden as well as whether pre-filling the most recent SDR responses would be a “good or bad idea” for future rounds of the survey. Missing responses were comparable across conditions and all ten of our selected REQs achieved response rates of 95% or higher. See Appendix B (supplement online) for screen shots of the respondent experience questions analyzed for both groups.

Methods

As mentioned previously, there were originally two variations of the dependent interviewing protocol. However, for the analyses reported, no statistically significant differences were found between the responses of the two treatment groups. Therefore, we combine both treatments into one TRT group for all presented analyses. Using the CTR and combined TRT, we begin by confirming the conditional dependence of responses on experimental condition by calculating Pearson’s X^2 for the responses to each REQ. We also calculate effect size in the form of odds ratios for each response category. Lastly, we converted the ordinal response categories to discrete numeric values and examined the difference in mean response values by calculating t-tests.

Since the categories for the reaction REQs represented ordinal but asymmetric intervals, we used the values: 0 = “not at all,” 1 = “a little,” and 3 = “very much.” The three efficacy questions were recoded using bipolar scales. The impact on accuracy was coded as: -1 = “less accurate,” 0 = “no impact,” 1 = “more accurate.” Due to the sparsity of some responses, the effect on burden and whether the protocol was a good or bad idea were both collapsed to a three-point scale ranging from -1 to 1, respective of the *negative*, *neutral*, and *positive* response options. All analyses were conducted in R (version 4.1.2, 2021). All findings below result from the analysis of the unweighted data.

Results

The mosaic plots in [Figure 1](#) illustrate the X^2 results by displaying area-proportional distributions using residual-based shading corresponding to the maximum X^2 statistic. The cut-off points ± 2 and ± 4 indicate that the corresponding cells are those with individual residuals significant $\alpha = 0.05$ and $\alpha = 0.0001$, respectively. (Negative values for residuals are indicated by dotted

outlines.) Response distributions trended in the same direction across groups; however, the CTR consistently exhibited more critical attitudes than the TRT. For example, we see that when the question is negative, such as the level of annoyance ([Figure 1](#) top left), responses of “not at all” are most frequent, but the proportion of the CTR reporting “not at all” is smaller than the proportion of the TRT. Conversely, when the question is positive, for example the level of reported appreciation ([Figure 1](#) top center), we see “very much” is the most frequent response but again, the proportion of the CTR reporting “very much” is smaller than the proportion of the TRT reporting the same.

The mosaic plot of the level of reported comfort ([Figure 1](#) top right) shows only 50% CTR participants reported they *would be very comfortable* seeing their answers pre-filled; a rate we would expect less than 0.01% of the time if the conditions were independent. Comparatively, we see that 72% of the TRT reported *being very comfortable* with seeing their prior responses pre-filled; a rate that would be expected less than 5% of the time if the conditions were truly independent. χ^2 tests of independence of the responses across groups indicate significant associations ($p < 0.001$ for all questions) with $\alpha = 0.05$.

The odds ratios of the conditional probabilities for each response further illustrate how significantly the opinions of the TRT deviate from the opinions of the CTR. In [Table 1](#), we see that across almost all questions, the TRT is more than twice as likely as the CTR to report positive reactions. The TRT was more than four times as likely than the CTR to express *no annoyance at all*, over three times as likely to express *no concern at all*, and over five times as likely to report *no confusion at all*. In reporting positive reactions, the TRT was twice as likely to report being *very appreciative* and they were two and a half times as likely to report being *very comfortable*. We find a similar and starker pattern among responses to the efficacy questions with the TRT being less likely to report a negative or neutral response to all questions and far more likely to respond positively. Across all estimated odds ratios, p-values were significant, rising above 0.01 in only two comparisons (i.e., *a little relieved* and *a little surprised*).

Using the discrete values noted in the “Methods” section, we found that t-tests comparing group means similarly resulted in significant associations between the conditions and responses. [Figure 2](#) illustrates the difference in the mean response values for each condition. Although many of the observed differences appear small, all differences were significant with $p < 0.001$. We believe these differences are representative of the measurement error resulting from the poor correspondence between the CTR’s hypothetical attitudes and the TRT’s reported experience. It would also appear that, on average, the mean reactions for negative sentiments are small. For example, the mean level of *annoyance*, *concern*, and *confusion* reported by each group are all below 0.50 which, given the range of 0 to 3, can be loosely interpreted as the mean for each group being close to “not at all.”

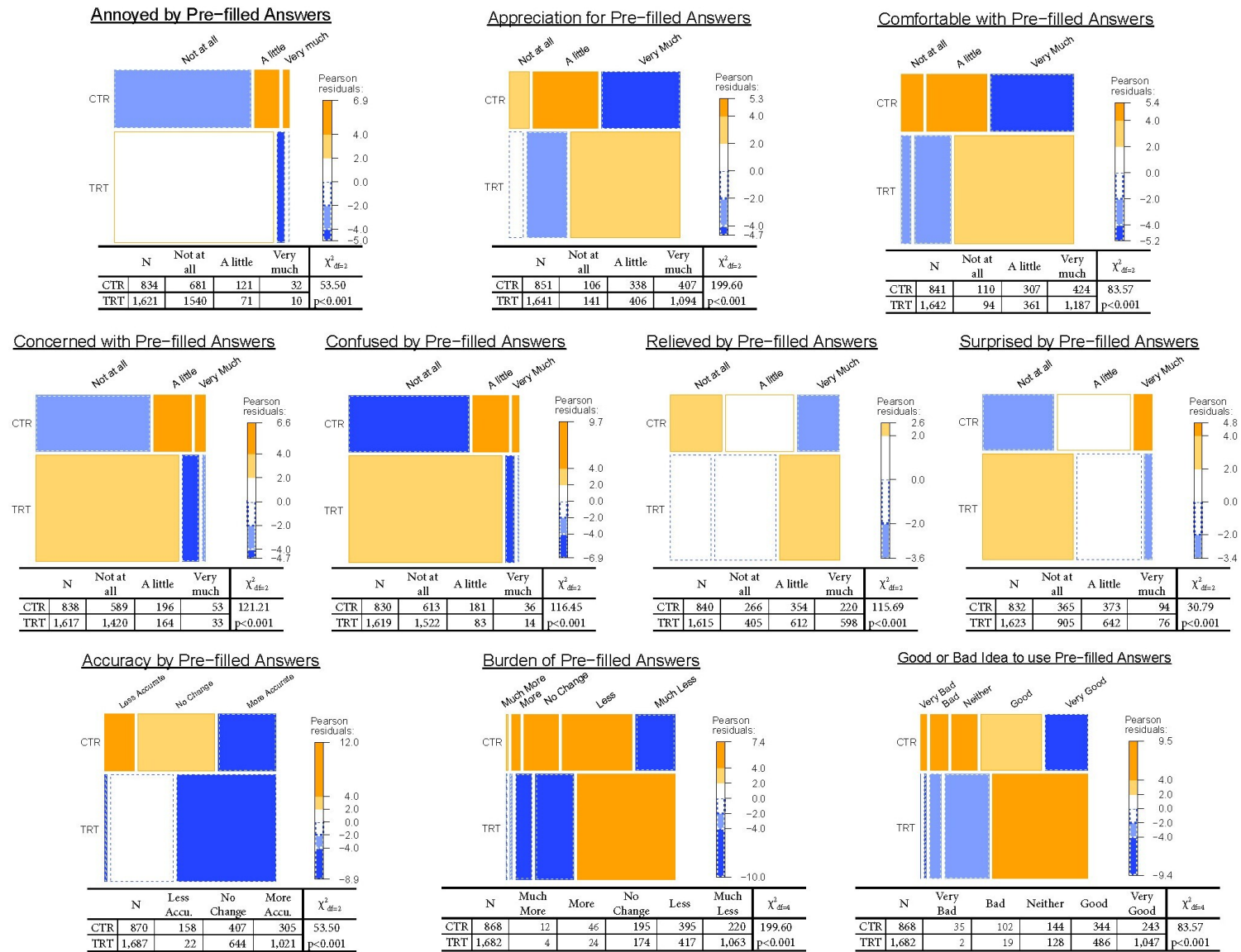


Figure 1. Mosaic Plots of Response Distributions

Table 1. Odds Ratios (TRT vs. CTR)

Reaction	Odds Ratio	95% Confidence Interval		p-value	
		Lower	Upper		
Annoyed					
	Not at all	4.27	3.19	5.75	<0.001
	A little	0.27	0.2	0.37	<0.001
	Very much	0.16	0.07	0.33	<0.001
Appreciative					
	Not at all	0.66	0.5	0.87	<0.01
	A little	0.5	0.42	0.6	<0.001
	Very much	2.18	1.84	2.59	<0.001
Comfortable					
	Not at all	0.4	0.3	0.54	<0.001
	A little	0.49	0.41	0.59	<0.001
	Very much	2.56	2.15	3.06	<0.001
Concerned					
	Not at all	3.05	2.46	3.78	<0.001
	A little	0.37	0.29	0.47	<0.001
	Very much	0.31	0.19	0.49	<0.001
Confused					
	Not at all	5.55	4.27	7.26	<0.001
	A little	0.19	0.15	0.26	<0.001
	Very much	0.19	0.1	0.37	<0.001
Relieved					
	Not at all	0.72	0.6	0.87	<0.001
	A little	0.84	0.7	1	<0.05
	Very much	1.66	1.37	2	<0.001
Surprised					
	Not at all	1.61	1.36	1.92	<0.001
	A little	0.81	0.68	0.96	<0.05
	Very much	0.39	0.28	0.53	<0.001
Accuracy					
	Negative Impact	0.06	0.04	0.09	<0.001
	No Change	0.7	0.59	0.83	<0.001
	Positive Impact	2.84	2.39	3.38	<0.001
Burden					
	Negative Impact	0.24	0.14	0.38	<0.01
	No Change	0.4	0.32	0.5	<0.001
	Positive Impact	3.01	2.44	3.73	<0.001
Good or Bad Idea					
	Bad Idea	0.07	0.4	0.11	<0.001
	Neither Good or Bad	0.41	0.32	0.54	<0.001
	Good Idea	4.92	3.93	6.18	<0.001

Discussion

Hypothetical questions are unique in that they inherently present risk of bias by asking respondents to report how they would behave amid imagined circumstances. This bias occurs when attitudes and resulting intentions are

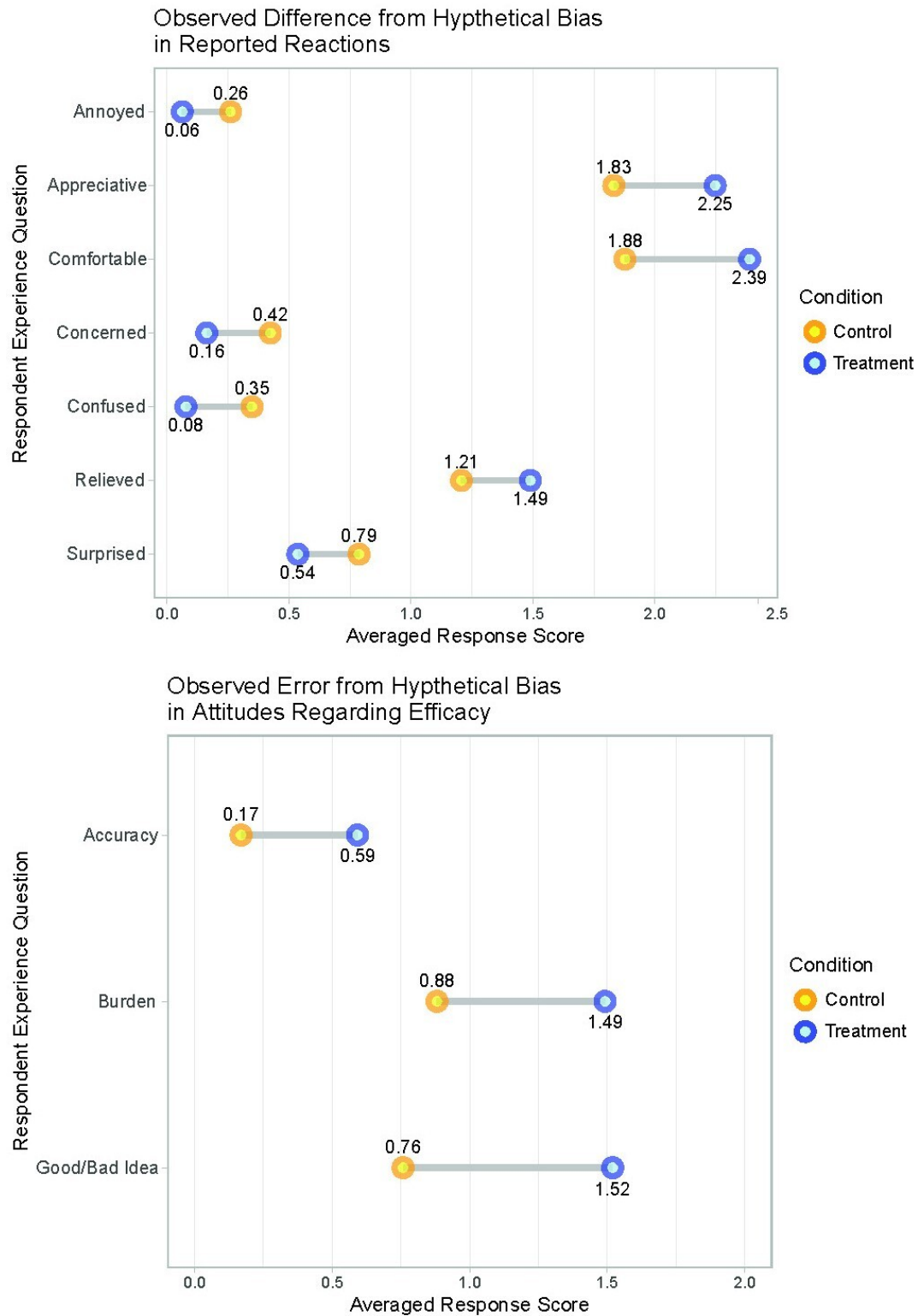


Figure 2. Group Differences in Mean Response by Question

All differences t-tested and found to be significant with $p < 0.001$ ($\alpha = 0.05$)

incongruent with observable behavioral outcomes. In our study, all respondents possessed a common frame of reference, i.e., the traditional version of the survey which they answered in a prior round, and we observed the CTR consistently exhibiting less preference for the hypothetical protocol

than the TRT across several dimensions. We believe this is an illustration of the incongruency between respondents' attitudes and behaviors when conceptualizing a hypothetical survey experience.

Our data allowed us to identify and evaluate aggregate response bias, but we did not have data on respondents' existing attitudes or their understanding of existing social conventions, such as their current attitudes toward privacy and data security. Such data could disentangle response bias stemming from the hypothetical context from other causes of measurement error. For example, if data were available on an individual's attitude toward privacy, it could be analyzed to determine if people more sensitive to privacy issues are less likely to support the treatment before experiencing it, leading to greater bias. We believe that respondents' prior (and ongoing) participation in longitudinal data collection is evidence that, at a minimum, they are not averse to disclosing the information requested in the SDR. This is not to say, however, that respondents' attitudes on privacy would be accepting of any change to the survey format. Without other attitudinal data on respondents, it is impossible to rule out whether or not the TRT (dependent interviewing protocol) could raise concerns over disclosure.

Although we identified significant response bias indicating that the CTR was consistently less receptive to the hypothetical treatment, the intention of this experiment was not to identify measurement error but to make a decision about the future administration of the survey, and hopefully reduce the burden on the panelists. Even without the experiment to implement the dependent interviewing treatment, the responses to the CTR's hypothetical questions would have supported dependent interviewing as an improvement to which respondents would be receptive. So, although we caution about the bias in hypothetical questions, in this study, the bias would not have influenced the outcome (i.e., the decision to implement or not). Perhaps the bias would have been more problematic if the data indicated respondent ambivalence about the treatment or the decision required more exacting evidence.

The field of social surveys is rapidly evolving with new technologies and sources of data offering promises of data collection efficiency and cost reductions. When testing new techniques and collecting data to understand how research participants relate to unfamiliar methods (and their likely adoption of new methods), hypothetical questions may be necessary, but researchers should be aware of the bias and risks they present. We recognize, however, that depending on the specificity of the research question, the risk may be minimal and the bias may not be problematic. We share our results to enrich our understanding of the potential for bias in measuring hypothetical intentions and encourage further work in this area.

.....

Acknowledgements

This paper and the research were made possible by the National Center for Science and Engineering Statistics (NCSES) at the National Science Foundation and the Oak Ridge Institute for Science and Education Fellowship Program. We would also like to thank Dr. Brady West for providing valuable guidance and feedback on our work. This paper was submitted while Adam Kaderabek was a master's student in the Program for Survey and Data Science at the University of Michigan and Jennifer Sinibaldi was the Research Director at NCSES.

Submitted: March 19, 2022 EDT, Accepted: September 01, 2022 EDT

REFERENCES

- Bradburn, Norman M., Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design - for Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: John Wiley & Sons.
- Carson, Richard T., and Theodore Groves. 2007. "Incentive and Informational Properties of Preference Questions." *Environmental and Resource Economics* 37 (1): 181–210. <https://doi.org/10.1007/s10640-007-9124-5>.
- Cook, Joseph, Marc Jeuland, Brian Maskery, and Dale Whittington. 2012. "Giving Stated Preference Respondents 'Time to Think': Results From Four Countries." *Environmental and Resource Economics* 51 (4): 473–96. <https://doi.org/10.1007/s10640-011-9508-4>.
- Hargittai, Eszter, Elissa M. Redmiles, Jessica Vitak, and Michael Zimmer. 2020. "Americans' Willingness to Adopt a COVID-19 Tracking App." *First Monday* 25 (11). <https://journals.uic.edu/ojs/index.php/fm/article/view/11095/9985>.
- Howard, Gregory, Brian E. Roe, Erik C. Nisbet, and Jay F. Martin. 2017. "Hypothetical Bias Mitigation Techniques in Choice Experiments: Do Cheap Talk and Honesty Priming Effects Fade with Repeated Choices?" *Journal of the Association of Environmental and Resource Economists* 4 (2): 543–73. <https://doi.org/10.1086/691593>.
- Jäckle, Annette. 2006. "Dependent Interviewing: A Framework and Application to Current Research." SER Working Paper Series, No. 2006-32. University of Essex, Institute for Social and Economic Research (ISER), Colchester. <https://www.econstor.eu/handle/10419/92099>.
- Keusch, Florian, Bella Struminskaya, Christopher Antoun, Mick P. Couper, and Frauke Kreuter. 2019. "Willingness to Participate in Passive Mobile Data Collection." *Public Opinion Quarterly* 83 (S1): 210–35. <https://doi.org/10.1093/poq/nfz007>.
- Norwood, F. Bailey, and Jayson L. Lusk. 2011. "Social Desirability Bias in Real, Hypothetical, and Inferred Valuation Experiments." *American Journal of Agricultural Economics* 93 (2): 528–34. <https://doi.org/10.1093/ajae/aaq142>.
- Olson, Kristen, Jolene D. Smyth, and Heather M. Wood. 2012. "Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination." *Public Opinion Quarterly* 76 (4): 611–35. <https://doi.org/10.1093/poq/nfs024>.
- Pascale, Joanne, and Thomas S. Mayer. 2004. "Exploring Confidentiality Issues Related to Dependent Interviewing: Preliminary Findings." *Journal of Official Statistics* 20 (2): 357.
- Pinter, Robert. 2015. "Willingness of Online Access Panel Members to Participate in Smartphone Application-Based Research." In *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, edited by Danielle Toninelli, Robert Pinter, and Pablo de Pedraza, 141–56. London: Ubiquity Press. <https://doi.org/10.5334/bar.i>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Revilla, Melanie, Mick P. Couper, and Carlos Ochoa. 2019. "Willingness of Online Panelists to Perform Additional Tasks." *Methods, Data, Analyses* 13 (2): 29.
- Schläpfer, Felix, and Baruch Fischhoff. 2012. "Task Familiarity and Contextual Cues Predict Hypothetical Bias in a Meta-Analysis of Stated Preference Studies." *Ecological Economics* 81: 44–47. <https://doi.org/10.1016/j.ecolecon.2012.06.016>.
- Struminskaya, Bella, Vera Toepoel, Peter Lugtig, Marieke Haan, Annemieke Luiten, and Barry Schouten. 2020. "Understanding Willingness to Share Smartphone-Sensor Data." *Public Opinion Quarterly* 84 (3): 725–59. <https://doi.org/10.1093/poq/nfaa044>.

- U.S. Department of Health and Human Services and Centers for Disease Control and Prevention. 2021. "COVID-19 Vaccine Confidence Rapid Community Assessment Guide."
<https://www.cdc.gov/vaccines/covid-19/vaccinate-with-confidence/rca-guide/#rapid-community-assessment-tools>.
- Wenz, Alexander, Annette Jäckle, and Mick P. Couper. 2019. "Willingness to Use Mobile Technologies for Data Collection in a Probability Household Panel." *Survey Research Methods* 13 (1): 1–22.

SUPPLEMENTARY MATERIALS

Appendix A

Download: <https://www.surveypractice.org/article/38280-assessing-measurement-error-in-hypothetical-questions/attachment/99652.docx>

Appendix B

Download: <https://www.surveypractice.org/article/38280-assessing-measurement-error-in-hypothetical-questions/attachment/99656.docx>
