Journal of Official Statistics
vol. 38, 4 (December 2022)

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 929–938, http://dx.doi.org/10.2478/JOS-2022-0040

Preface
Overview of the Special Issue on Respondent Burden

1. Introduction

In an era of increasing nonresponse and attrition rates (Brick and Williams 2013), a better understanding of how to measure and reduce respondent burden is crucial. Respondent burden can contribute to lower response rates, increase attrition in panel surveys (Rolstad et al. 2011), and increase survey nonresponse and nonresponse bias (Crawford et al. 2001). Respondent burden can also affect data quality by increasing item nonresponse (Fricker et al. 2014), breakoffs (Galesic 2006), or satisficing (Krosnick 1999). Despite the negative impact of respondent burden on survey outcomes, few surveys include direct measures of burden, asking respondents to self-report their level of burden or other burden-related constructs such as sensitivity, difficulty, or effort level, and researchers and survey organizations often have different conceptualizations of respondent burden. This lack of common or standardized definition of burden poses challenges for understanding the impact of respondent burden on survey response and how to mitigate burden in the future.

In this special issue on respondent burden, we highlight recent innovations in research on respondent burden, including how burden is conceptualized and measured, survey and respondent characteristics that contribute to burden, and the consequences of burden (for example, impact on response rates, survey attrition, and data quality). To help organize the 12 articles in this issue into a common framework, Table 1 shows the type of burden measured, respondent and survey characteristics contributing to burden, and the outcomes and consequences of burden in each of the 12 articles.

2. Conceptualization and Measurement

As the diverse articles in this special issue demonstrate, there are differences amongst survey organizations and researchers in how burden is conceptualized and measured. Many countries require the monitoring of response burden for official surveys (e.g., European Commission 2017). However, these regulations typically focus on objective burden (for example, time needed to complete a questionnaire) and may allow for considerable freedom in how burden is assessed (e.g., expert estimate, direct measurement). In the U.S., the dominant measure of respondent burden relies on objective metrics such as the survey duration, as well as the time spent reading instructions, gathering and entering data, and reviewing it (Paperwork Reduction Act Guide 2011). In a classic paper on the topic, Bradburn (1978) posited that burden can also

Acknowledgments: The Guest Editors would like to acknowledge the memory of Scott Fricker, and to highlight his great contributions to studying respondent burden in this special issue. We would like to thank Polly Phipps and Nancy Bates for their help and support in making this special issue a reality. We are grateful for your guidance and expertise.

Table 1. Type of burden measured (conceptualization of burden), respondent and survey characteristics (causes of burden), and the outcome and consequences of burden studied in the 12 articles in this issue.

Special issue articles	Type of burden	Respondent characteristics	Survey characteristics	Outcomes/Consequences of burden
1. "Response Burden: A Review and a Conceptual Framework": Yan and Williams	Theoretical discussion of multiple components of burden and the concept of burden at different time points of a survey (initial, cumulative, and continued)	Interest; motivation; attitudes; cognitive ability	Length; frequency; design, incentive; mode; framing of survey request	Subjective perceptions of initial, cumulative, and continued burden, and impact on response and data quality
2. "Testing a Planned Missing Design to Reduce Respondent Burden in Web and SMS Administrations of the CAHPS Clinician & Group Survey (CG-CAHPS)": Brenner, Hargraves, and Cosenza	Survey duration	Self-rated health and demographics (used as controls)	Survey mode (web versus SMS); survey duration	Survey completion and breakoffs
3. "Response burden and dropout in a probability-based online panel study – a comparison between an app and browser-based design": Roberts, Herzing, Asensio Manjon, Abbet, and Gatica-Perez	Survey duration; five-point fully labelled agree/disagree response scale for four statements on interest, question comprehension, and difficulty	Device (PC versus mobile)	Survey mode (app vs. browser)	Breakoff rates; non-participation in a second wave

Table 1. Continued

Special issue articles	Type of burden	Respondent characteristics	Survey characteristics	Outcomes/Consequences of burden
4. "The Effect of Burdensome Survey Questions on Data Quality in an Omnibus Survey": Phillips and Stenger	Inclusion of difficult survey topics and questions (social network and knowledge question modules)	Age and education	Placement of social network and knowledge question modules	Item nonresponse; data quality; acquiescence
5. "Relationship Between Past Survey Burden and Response Probability to a New Survey in a Probability-Based Online Panel": Jin and Kapteyn	Survey length and frequency	Sex; age; race and ethnicity; marital status; education; household income; household size; big five personality traits; unobserved respondent commitment	Number of prior survey invitations and duration of prior surveys in a panel study	Response to new survey request in panel study
6. "The Effects of Response Burden: Collecting Life History Data in a Self-Administered Mixed-Device Survey": Carstensen, Lang, and Cordua	Survey duration; self-reported survey difficulty	Respondent choice of mode; respondent demographics (used as controls)	Survey mode (PC versus mobile); collection of detailed life history module	Breakoffs; panel consent; satisficing; data quality; item nonresponse

Table 1. Continued

Special issue articles	Type of burden	Respondent characteristics	Survey characteristics	Outcomes/Consequences of burden
7. "Your Best Estimate Is Fine. Or Is It?": Timbrook, Olson, and Smyth	Total number of conversational turns; question duration in telephone survey; no direct burden measurement in the mail survey	None	Invitation to estimate (ItE); Interaction of ItE with question characteristics (difficulty and sensitivity)	Quality indicators (item nonresponse; nonsubstantive answers; qualified/range answers; heaped/rounded answers; estimated means)
8. "Analyzing the Association of Objective Burden Measures to Perceived Burden with Regression Trees": Yang and Toth	Survey duration; Four items with five-point response scales on overall burden, difficulty, sensitivity, and length	Number of expenditures; use of expenditure records; call records; demographics (used as controls)	Survey mode (personal visit versus telephone)	None
9. "Modeling the Relationship between Proxy Measures of Respondent Burden and Survey Response Rates in a Household Panel Survey": Earp, Kaplan, and Toth	Survey length and frequency; question effort and sensitivity	Respondent socioeconomic and demographic characteristics (e.g., employment, home ownership, marital status); prior survey participation	Survey length and frequency	Responses rates; attrition over time

Table 1. Continued

Special issue articles	Type of burden	Respondent characteristics	Survey characteristics	Outcomes/Consequences of burden
10. "Exploring Burden Perceptions of Household Survey Respondents in the American Community Survey": Holzberg and Katz	13 subjective burden questions used in conjunction with other discussion questions to collect feedback from respondents in focus groups	Household size and characteristics	Survey mode; length; contact strategy; question difficulty and sensitivity; respondent attitudes, beliefs, and knowledge	None
11. "Determination of the Threshold in Cutoff Sampling Using Response Burden with an Application to Intrastat": Bavdaž, Polanec, and Smith	Survey duration	Respondent business sector (manufacturing, trade, other)	Survey topic (trade flow); applies technique for optimal cut-off sampling and discusses trade-offs between response burden and measurement accuracy	Accuracy of estimates
12. "A User-Driven Method for Using Research Products to Empirically Assess Item Importance in National Surveys": Ong, Schultz, Sinozich, Sinibaldi, West, Wagner, and Finamore	Applied article where information on responses were used in conjunction with information about data use to propose a new method for managing burden	Survey duration and breakoffs	Information about item usage (download statistics; citations; importance of citation source)	Identification of items that may be burdensome and infrequently used as potential candidates for removal; justification for retaining burdensome but important/ frequently cited items

be thought of as a multidimensional concept that includes respondents' subjective perceptions of how effortful the survey is, how sensitive or invasive the questions are, and how long the survey takes to complete. The level of burden can also vary by the mode of data collection, survey topic, demographic characteristics of respondents, and frequency with which individuals or businesses are sampled: a 30 minute in-person survey may seem less burdensome than a 30-minute web survey.

In the nearly 45 years since Bradburn's paper was published, objective measures of respondent burden have remained dominant; little research has assessed both objective and subjective sources of respondent burden to better understand the unique contributions each may have. Based on extensive qualitative research with establishment (business) surveys, Dale et al. (2007) argue that National Statistical Institutes should directly measure and monitor four aspects of burden: (1) Perceived response burden (measured as perception of time, perception of burden); (2) Actual response burden (measured as time to collect information, time to complete the questionnaire); (3) Perceived causes of response burden (measured as reasons for time consumption, conditions for burden) and (4) Motivation (measured as perceived usefulness to business and society). However, these standards have not been implemented widely.

The work of the late Scott Fricker and colleagues (e.g., Fricker et al. 2012, 2014) was amongst the first to use direct measures of respondent burden in a series of models for a U.S. Federal government survey. They asked respondents to report the amount of burden they experienced in the Consumer Expenditure Quarterly Interview Survey, a survey sponsored by the U.S. Bureau of Labor Statistics. This work used the dimensions developed by Bradburn (1978), including effort, survey length, and frequency, to explore respondent burden and its impact on data quality, survey costs, and response rates. Findings from this research showed that burden is a multidimensional construct; some of the objective measures of burden such as length did not impact burden, but subjective factors such as respondents' perception of the survey task and motivation had a significant effect on burden and survey outcomes. Scott Fricker and colleagues advocated for inclusion of subjective burden measures such as these on U.S. Federal surveys and for examining respondents' subjective perceptions of burden rather than just objective measures; this work treated respondent burden as a function of the respondent and the survey and not just the survey alone. This work inspired other survey methodologists at the U.S. Bureau of Labor Statistics, U.S. Census Bureau, and other organizations to pursue research on how to conceptualize respondent burden using subjective measures and strategies to mitigate respondent burden in the future, including the Guest Editors' desire to work on this special issue.

Several articles in this special issue discuss or use subjective burden measures as an important component of burden. Yan and Williams included subjective perceptions of burden as an important part of their overall framework in conceptualizing respondent burden. Yang and Toth leveraged questions on overall burden, difficulty, sensitivity, and length that Fricker et al. worked diligently to include in the Consumer Expenditure Survey to analyze the relationship between objective and subjective burden. Holzberg and Katz also used Consumer Expenditure Survey questions as part of focus group discussions with American Community Survey respondents to learn more about the influence of different survey features on burden. Other papers also incorporated different measures of subjective

burden as a key part of their analyses. Roberts et al. asked respondents completing an online survey to self-report whether they thought the questionnaire was interesting, the length was adequate, the questions were comprehensive, and that filling in the questionnaire presented no difficulty. Differences in these measures were examined by survey mode and device. Cartensen et. al. directly asked respondents who completed a life history questionnaire, "how difficult was it for you to provide the information on your background that we asked you about in this survey?" and used responses as a measure of whether treatments increased response burden.

Overall, there seems to be no consensus on if and how burden should be assessed in surveys. This lack of consensus has led to differences in the purpose of studying burden; for example, studies may focus on accumulated burden on panel respondents over time, the burden of being repeatedly sampled, or the burden of a particular item. These differences are also partly caused by data availability. If no direct measures of burden are available, questionnaire characteristics, such as an assessment of question difficulty, may be taken as indicator of burden. However, an important drawback of this approach is that it does not allow researchers to disentangle causes of burden and burden itself. For example, difficult survey questions may lead to item nonresponse and increased survey response times, but it may be challenging to determine whether the difficult question had a causal impact on these survey outcomes without more direct measures of question difficulty from the respondents' perspective. In addition, as seen in Table 1, there is often overlap between survey characteristics that contribute to burden (such as survey duration) and burden measures, hence sometimes the type of burden measured is also a characteristic of the survey. In the following sections, we outline the measures of burden and the respondent and survey characteristics that were used by each article in this special issue to help determine the outcomes and consequences of burden.

3. Respondent Characteristics

The articles in this issue draw on multiple respondent and survey characteristics that contribute to burden. In their theoretical article, Yan and Williams argue that respondent factors such as interest, motivation, attitudes, and cognitive ability can affect the level of burden experienced. Other respondent characteristics, such as demographics and socioeconomic factors of a household or establishment, can be treated as direct measures or used as controls in investigating respondent burden. For example, Earp, Kaplan, and Toth used respondent characteristics such as employment, marital status, and prior survey participation to predict level of burden and response rates. Several papers used respondent choice of survey mode (as opposed to assignment of mode) as a respondent characteristic that may contribute to burden. For example, Carstensen et al. found that participants who chose to use a mobile response mode showed increased breakoffs and reported fewer life history episodes than non-mobile users, suggesting more burden. Turning to establishment surveys, Bavdaž et al. used the respondent's business sector as a respondent characteristic that may affect level of burden. These and other articles in this issue highlight the types and diversity of respondent characteristics that may play a role in the level of burden experienced and can be used when direct measures (e.g., asking respondents to self-report their level of burden) are not available.

4. Survey Characteristics

Features of surveys such as the survey frequency, mode, question features, or contact strategies were also examined for their impact on respondent burden. For example, Phillips and Stenger investigated question order and placement of burdensome questions to determine if order effects impact level of burden. Carstensen et al. looked at the impact of a long reference period in collection of detailed life history information within a survey module and how the length of the period impacts respondent burden. Brenner et al. investigated whether a planned missing design to reduce survey duration would lead to differences in survey completion and breakoffs. Turning to differences in survey administration, Timbrook et al. looked at the impact of interviewers instructing respondents to allow for estimated answers on reduction of burden.

Several articles also looked at the relationship between survey mode and respondent burden. For example, Roberts et al. showed that respondents using an app were more likely to participate in future survey waves than those using a mobile browser, suggesting less burden when using an app. Yang and Toth assessed survey mode used in the Consumer Expenditure Survey (personal visit versus telephone) and its relationship to objective and subjective burden.

Features of individual questions are also an important survey characteristic that can affect burden. Several articles assessed the impact of survey items that are difficult or sensitive in nature, including Holzberg and Katz, and Earp et al. In addition, several articles leveraged previous knowledge of questions respondents tend to find difficult or sensitive would lead to differences in burden outcomes when more direct measures of burden were not available (e.g., Timbrook et al.; Earp et al.; Phillips and Stenger) as well as survey frequency and participation over time (e.g., Jin and Kapteyn; Earp et al.). These and other articles in this issue highlight the types of survey characteristics that are likely to impact respondent burden.

5. Consequences of Burden

All papers in the special issue were interested in how respondent factors and survey characteristics would ultimately contribute to respondent burden and survey outcomes, including response rates, attrition, breakoffs, and data quality. Survey breakoffs were assessed in several articles as a consequence of burden (e.g., Brenner et al.; Carstensen et al.; Roberts et al.; Ong et al.). The impact of survey frequency and participation in panel surveys, as measured through survey attrition and refusal to participate in future waves, was also assessed in several articles, including Jin and Kapteyn, Earp et al., Roberts et al., and Carstensen et al. Data quality was examined and operationalized in a wide variety of ways across papers, including item nonresponse (e.g., Phillips and Stenger; Timbrook et al.), acquiescence (Phillips and Stenger), satisficing, heaping, and rounding (Carstensen et al.; Timbrook et al.), whether respondents provided codeable answers (Timbrook et al.), and survey response time and future participation (Brenner et al.; Hargraves et al.; Ong et al.).

6. Conclusion

Declining response rates and rising survey costs will likely make the study of burden more important and more common in the future. If survey nonresponse is caused in part by

respondent burden, it is critical to collect direct, self-report measures of respondent burden whenever possible to quantify its effects and help to increase response rates. We anticipate further research into how subjective perceptions of burden are measured, particularly for household surveys. We also envision increased reliance on objective and subjective burden measurements in monitoring efforts and quality assessments. Ideally, researchers will continue to add measures of subjective burden when possible, especially since subjective and objective measures do not always correlate. Because changes to the methodology of a survey can impact respondent burden, it is critical to continuously measure burden to capture these changes.

Burden should also be considered in applications beyond survey data. Many advocates of using administrative records, wearables, and other less traditional forms of data assert that these new methodologies are less burdensome for respondents. We argue that this is an assumption to be further scrutinized. Requests for consent to link data sources may contribute to a different form of respondent burden, for example. Finally, we underscore the importance of assessing burden from all angles, including examination of final survey data. Approaches like the one undertaken by Ong et al. illuminate the importance of systematically re-considering whether respondents need to be asked certain questions at all.

Ultimately, while the articles in this special issue approach the study of burden in diverse ways, all serve as examples of leveraging whatever data is available to study burden. It is important that researchers make it clear how burden is defined and measured, given the myriad of ways it can be operationalized. We look forward to seeing what is next.

Robin L. Kaplan
Jessica Holzberg
Stephanie Eckman
Deirdre Giesen
Guest Editors

7. References

Bradburn, N. 1978. "Respondent burden." In Proceedings of the Survey Research Methods Section of the American Statistical Association 35: 40. Alexandria, VA, USA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf.

Brick, J.M., and D. Williams. 2013. "Explaining rising nonresponse rates in cross-sectional surveys." *The ANNALS of the American academy of political and social science* 645(1): 36–59. DOI: https://doi.org/10.1177/0002716212456834.

Crawford, S.D., M.P. Couper, and M.J. Lamias. 2001. "Web surveys: Perceptions of burden." *Social science computer review* 19 (2): 146–162. DOI: https://doi.org/10.1177/089443930101900202.

Fricker, S., C. Kreisler, and L. Tan. 2012. "An exploration of the application of PLS path modeling approach to creating a summary index of respondent burden." In *JSM Proceedings*: 4141–4155.

Fricker, S., T. Yan, and S. Tsai. 2014. Response burden: What predicts it and who is burdened out. In *JSM Proceedings*: 4568–4577. Available at: https://www.bls.gov/osmr/research-papers/2014/pdf/st140170.pdf.

Dale, T., and J. Erikson, J. Fosen, G. Haraldsen, J. Jones, and Ø. Kleven 2007. *Handbook for Monitoring and Evaluating Business Survey Response Burdens*. European Commission, Eurostat. Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/12-HANDBOOK-FOR-MONITORING-AND-EVALUATING-BUSINESS-SURVEY-RESONSE-BURDEN.pdf/600e3c6d-8e8d-44f7-a8f5-0931c71d9920.

European Commission. 2017. *European Statistics Code of Practice for the National Statistical Authorities and Eurostat (EU statistical authority)*. European Statistical System Committee. Available at: https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7?t = 1528447068000.

Galesic, M. 2006. "Dropouts on the web: Effects of interest and burden experienced during an online survey." *Journal of Official Statistics* 22 (2): 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf.

Krosnick, J.A. 1999. " *Survey research*." *Annual review of psychology* 50(1): 537–567. Office of Personnel Management. Available at: https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/paperwork-reduction-act-guide.pdf (accessed June 2022).

Paperwork Reduction Act Guide. 2011. Office of Personnel Management. Available at: https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/paperwork-reduction-act-guide.pdf.

Rolstad, S., J. Adler, and A. Rydén. 2011. "Response burden and questionnaire length: is shorter better? A review and meta-analysis." *Value in Health* 14 (8): 1101–1108. DOI: https://doi.org/10.1016/j.jval.2011.06.003.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response burden: What is it and what predicts it?" *Advances in Questionnaire Design, Development, Evaluation and Testing*: 193–212. DOI: https://doi.org/10.1002/9781119263685.ch8.

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 939–961, http://dx.doi.org/10.2478/JOS-2022-0041

Response Burden – Review and Conceptual Framework

Ting Yan[1] *and Douglas Williams*[2]

Concerns about the burden that surveys place on respondents have a long history in the survey field. This article reviews existing conceptualizations and measurements of response burden in the survey literature. Instead of conceptualizing response burden as a one-time overall outcome, we expand the conceptual framework of response burden by positing response burden as reflecting a continuous evaluation of the requirements imposed on respondents throughout the survey process. We specifically distinguish response burden at three time points: initial burden at the time of the survey request, cumulative burden that respondents experience after starting the interview, and continuous burden for those asked to participate in a later round of interviews in a longitudinal setting. At each time point, survey and question features affect response burden. In addition, respondent characteristics can affect response burden directly, or they can moderate or mediate the relationship between survey and question characteristics and the end perception of burden. Our conceptual framework reflects the dynamic and complex interactive nature of response burden at different time points over the course of a survey. We show how this framework can be used to explain conflicting empirical findings and guide methodological research.

Key words: Response burden; initial burden; cumulative burden; continuous burden.

1. Introduction

Concerns about the burden that surveys place on respondents have a long history in the survey field. As early as the 1920s, survey researchers and organizations were warned that lengthy interviews would impose excessive burden for respondents (Chapin 1920; Sharp and Frankel 1983). In one of the most influential papers on burden, Bradburn (1977, 49) stated that "the topic of respondent burden is not a neat, clearly defined topic about which there is an abundance of literature". More than four decades later, falling response rates and the establishment of regulations to reduce the time and effort required of respondents to government surveys have contributed to a growing empirical literature on this topic. However, the term "burden" is loosely defined and it is conceptualized and measured in various ways.

As a matter of fact, burden is used as a blanket term throughout the literature referring to both actual or objective burden and subjective or perceived burden. Researchers in the last four decades have used burden (e.g., Groves et al. 1999), respondent burden (e.g., Bradburn 1977), response burden (e.g., Filion 1981), perceived burden (e.g., Hedlin et al. 2008) and reporting burden (e.g., Eurostat 2011) in their work.

[1] Westat, 1600 Research Blvd, Rockville, MD 20850, U.S.A. Email: tyanuconn@gmail.com
[2] U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington D.C. 20212, U.S.A. Email: williams.douglas@bls.gov

While these terms are sometimes used interchangeably, researchers vary in the extent to which they use these terms to refer to actual burden imposed on respondents (e.g., the number of pages in a mail questionnaire) or perceived burden felt by respondents (e.g., the feeling of burdened out). In this article, we define response burden as negative feelings or hardships experienced by survey respondents (Graf 2008; Frankel 1980; Sharp and Frankel 1983). It is subjective burden perceived by respondents. The focus of the article is on factors contributing to, mediating, or moderating subjective burden. Actual or objective burden will be discussed as a cause of subjective burden.

To guide future research on response burden, this article has two goals. First, we review the many ways response burden or components of response burden have been conceptualized and summarize the state of the literature on the measurement of response burden for household surveys. (For business surveys, we refer readers to the Total Business Survey Burden Model described in Haraldsen et al. 2013.) Second, we propose a dynamic conceptual framework of response burden that builds on earlier conceptual frameworks and consolidates and synthesizes empirical research on response burden. Our conceptual framework allows survey researchers to have a comprehensive understanding of factors contributing to survey respondents' perception of response burden as well as the consequences of response burden through a survey fielding cycle. The conceptual framework we propose is an important contribution, providing a foundation for planning and designing future research around a topic that has thus far eluded a clear conceptualization. With this framework, we provide survey researchers a tool for identifying gaps in the existing literature to guide new research designs. Survey practitioners, on the other hand, can use the framework to inform the design of their survey protocols to identify and reduce response burden.

2. A Review of Conceptualization of Response Burden

As one of the earliest conceptualizers of response burden, Bradburn envisioned burden as the product of "an interaction between the nature of the task and the way in which it is perceived by the respondent" (Bradburn 1977, 49). Although he used the term "respondent burden" in his article, his conceptualization of burden is the same as our definition of response burden. Bradburn identified four factors influencing response burden: interview length, the effort required of respondents, the frequency of interviewing, and the amount of stress on respondents. The first three factors are objective burden whereas the fourth one is subjective burden. Bradburn concluded with his conceptualization that response burden is largely mediated by the importance (to the respondent) of the survey data, implying a simplified conceptualization of burden. That is, if only we can convey the importance of the survey, respondents will willingly accept greater burden.

Haraldsen (2004) outlines a model where the subjective perception of burden is shown as a mediating variable explaining the relationship between the various causes of survey burden and data quality (Figure 2). Causes of response burden are further divided into survey properties reflecting objective burden and respondent characteristics (Figure 3). Haraldsen (2004) then expanded the model to serve as a toolbox to decompose and test burden properties of Internet business surveys (Figure 4) and presented qualitative test results to shed light on survey properties and their impact on data quality. However, these

tests were unable to shed light on the interaction between the survey properties and respondent characteristics.

Read (2019) proposes seven factors contributing to response burden: length, effort, emotional stress, frequency, availability/opportunity, ability/competence, and motivation/interest. The seven factors are a combination of objective burden, subjective burden and moderators of burden. In addition, Read (2019) makes a distinction between discrete burden and cumulative burden in the setting of an app use study. Discrete objective burden is the amount of actual burden placed by each individual task whereas cumulative burden is the sum of discrete objective burden across all tasks. Similarly, discrete subjective burden is the respondents' perception of burden for each task or for the same task at a different time point. Cumulative subjective burden refers to the trend of discrete burden over a period of time. Using data from an app use study, he then empirically examined the relationship between objective burden and subjective burden. He found that measures of objective burden (that is, the number of app uses completed, total time spent completing these app uses, average time per app use, and the durations of each app use) are not closely related to measures of subjective burden (likelihood of participation, time/effort well spent, interest in the survey, and perceived difficulty of the survey). Furthermore, there was no consistent change in subjective burden throughout the four weeks of data collection, failing to support his definition of cumulative subjective burden and calling for an updated framework.

Yan et al. (2020) combined Bradburn's and Haraldsen's work and posited a path model in an attempt to quantify the joint impact of survey properties and respondent characteristics on respondents' perception of response burden. They are the first to present empirical results for their conceptual model. They showed, through observational data, that low motivation, difficult tasks, challenging survey effort, and negative perceptions of the survey all contribute to the perception of response burden. Specifically, respondents with lower motivation, harder recall tasks, challenging survey effort, and negative attitudes towards the survey are more likely to experience (or report the perception of) burden than those with higher motivation, easier reporting tasks, and positive perceptions. In addition, a negative perception of the survey mediates the effects of motivation, task difficulty, and survey effort on response burden.

A common weakness with three of the four conceptual models of response burden (Bradburn 1977; Haraldsen 2004; Yan et al. 2020) is that response burden is envisioned as a one-time overall outcome. In reality, the level of perceived burden is likely to fluctuate throughout a survey cycle and may be affected by different factors at different time points. For instance, when faced with a new survey request, respondents may have an initial expectation about the likely level of burden associated with the survey request, which affects their decision to participate in the survey. However, once they start the survey and as they continue progressing through the survey, an extremely difficult or sensitive survey item may sharply increase the level of the perceived burden, causing them to reconsider their decision to stay on or to quit (compare the "sample-decide-reconsider" model of breakoffs proposed in by Tourangeau et al. 2013). Furthermore, when respondents are invited back by the same survey for another interview, experience with the prior interview affects the level of perceived burden about this new survey request. A conceptual

framework that considers response burden at different time points through a survey cycle is much needed to understand the dynamic change of perceived burden.

3. A Review of Measurement of Response Burden

A review of empirical literature shows that researchers have used four different approaches to measure response burden throughout four decades of research (see Table 1 in Yan et al. (2020) for a summary of measurements of and uses of burden in empirical research). The first approach follows the meaning of the term "perceived burden" by directly asking respondents how burdensome the survey was (Bottone et al. 2018; Galesic 2006; Hedlin et al. 2008; Yan et al. 2020). This single measurement of burden is found to be related to respondents' decision to break off a web survey (Galesic 2006), difficulty of contacting and recruiting respondents to take part in a survey (Yan et al. 2020), the amount of item missingness (Yan et al. 2020), and the level of respondents' concerns with a survey request (Yan et al. 2020).

The second approach measures properties of surveys or tasks that are believed to impose response burden. This is primarily the length of an interview, in terms of time spent completing the interview, or number of pages or questions required of the survey, difficulty of the response task and the number of survey requests (Filion 1981; Warriner 1981; Hoogendoorn and Sikkel 1998; Groves et al. 1999; Singer et al. 1999; Hoogendoorn 2004; Rostald et al. 2011; Read 2019; Kleinert et al. 2021). This method identifies likely *sources* of response burden rather than measuring response burden as experienced by respondents. They are essentially "actual burden" or "objective burden". This measurement of burden is typically used in understanding unit nonresponse and attrition (e.g., Groves et al. 1999; Hoogendoorn and Sikkel 1998).

The third approach measures respondents' attitudes and beliefs toward surveys, such as their interest in the survey topic, their views about the importance of the survey, and their perception of the time and effort spent (Sharp and Frankel 1983; Hoogendoorn 2004; Stocke and Langfeldt 2004; Galesic 2006; Fricker et al. 2011, 2012; Geisen 2012; Read 2019; Atkinson et al. 2019). Sometimes, response burden is inferred or extracted from respondents' doorstep concerns, complaints, and verbatim answers (Giesen et al. 2018; Wenemark et al. 2010). These respondent attitudes are potential mediators of the perception of response burden, resulting in differential perceptions of burden for the same survey across respondents, but they are not direct measures of response burden themselves.

The last approach measures response burden through the consequence of 'feeling burdened' such as respondents' willingness to be re-interviewed and their feeling of exhaustion, and so on (Sharp and Frankel 1983; Stocke and Langfeldt 2004).

Although these very different measurements of response burden inform our understanding of the concept of response burden, only the first approach measures subjective burden directly.

4. A Dynamic Conceptual Framework of Response Burden

We expand the conceptual framework of response burden to incorporate and reflect the dynamic and complex interactive nature of response burden at different time points over

the course of a survey. Instead of conceptualizing burden as a static outcome at one point in time, we propose a framework positing burden as reflecting a continuous evaluation of the requirements imposed on respondents throughout the survey process. We specifically distinguish burden at three time points: (1) the initial perception of response burden at the time of a new survey request (or with mail surveys, at the initial receipt of the questionnaire); (2) cumulative subjective burden that respondents experience after starting the interview; and (3) continued burden for those who are asked to participate in a later round of interviews in a longitudinal setting. At each time point, response burden has different causes, moderators or mediators, and outcomes and consequences. We show this in our expanded conceptual framework in Figure 1. (Please note that Figure 1 is not a path diagram as in structural equation modeling. Instead, it is a figure aimed to show perceived burden as well as contributing factors at three time points of a survey cycle.) In addition, at each time point, response burden can be measured continuously, once, or at a few meaningful time points, and directly or indirectly.

Initial burden refers to response burden perceived at the time of a brand new survey request. The survey request could be an advance postcard or an invitation letter sent to a respondent's address, a voice mail left by an interviewer, or an introduction made by an interviewer at the door step. Initial burden changes as respondents learn more about the survey request, when their concerns are addressed, or when the interviewer offers an incentive. The consequence of initial burden is sampled members' decision to participate; those who anticipate response burden to be high are less likely to take part in the survey request and, even if they finally decide to participate, require more contacts and recruitment effort (Yan et al. 2020). Typical outcomes influenced by initial burden include unit nonresponse rates at the survey level and response propensity at the individual level.

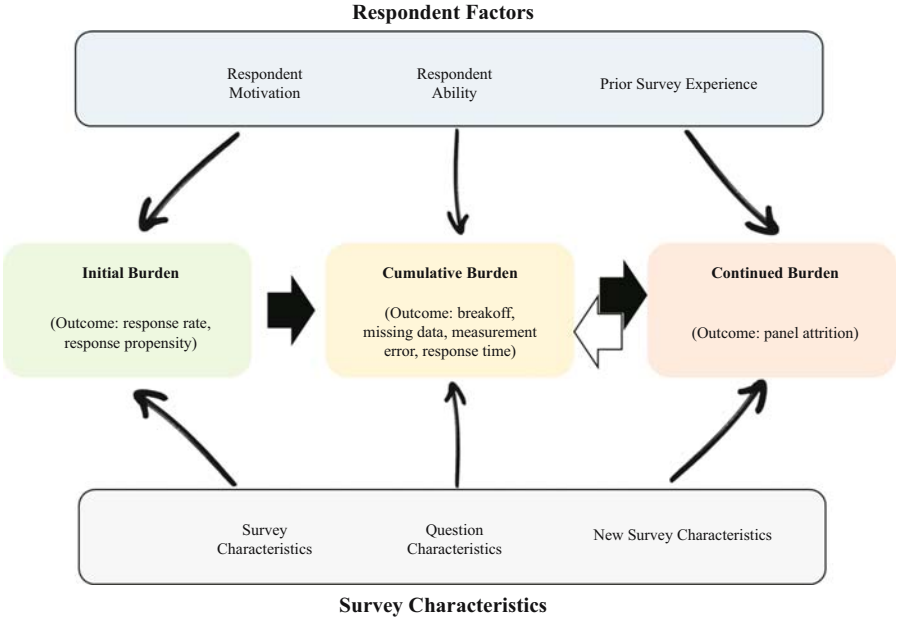


Fig. 1. Conceptual framework of response burden.

Initial burden is rarely measured directly but is often measured indirectly through proxies such as the number of contact attempts required to complete an interview, the number of times a respondent expresses unwillingness to participate in a survey request, and whether additional mailings are needed and so on.

Cumulative burden refers to response burden experienced by respondents as they progress through (i.e., answer) the survey. It is closely associated with the amount of effort exerted by respondents in answering the survey. Cumulative burden changes as respondents progress through different types of survey items in a survey questionnaire. For instance, a complex grid on a web screen or a difficult question asking about the number of doctor visits in the last five years could potentially increase cumulative burden felt by respondents. By contrast, a running tally feature that sums up numbers entered by respondents may decrease the level of cumulative burden perceived by respondents. The outcomes of cumulative burden include breakoffs and satisficing response behaviors. Specifically, those who feel burdened are more likely to break off the interview, as predicted by the "sample-decide-reconsider" model of breakoffs (Tourangeau et al. 2013) and demonstrated empirically by Galesic (2006). They are also more likely to skip questions (Yan et al. 2020), to satisfice by taking mental shortcuts, and to provide lower quality data (Warriner 1981). In countries where respondents are asked explicitly for consent to be contacted for the next round of interviews, consent is another outcome of cumulative burden. Cumulative burden can be measured directly by asking respondents how burdened they are; Galesic (2006) provides a great example of gauging cumulative burden experienced by respondents throughout a web interview. Indirect measurements of cumulative burden include time taken to answer survey questions (e.g., Yan et al. 2015), the amount of pupil dilation (e.g., Neuert 2020; Yan 2021) and the number and duration of fixations (e.g., Lenzner et al. 2010) captured by eye-tracking, the number of times respondents asked for definitions or clarification, and so on.

Continued burden pertains to longitudinal surveys that ask respondents to participate in more than one round of interviews. We specifically define this in the context of longitudinal surveys, where subsequent requests are built upon prior survey request and participation. The outcome of continued burden is panel attrition; that is, panel respondents' decision to participate in multiple rounds of surveying. Although multiple survey requests of farm establishments were not found to have an effect on future survey participation, those who experience a higher level of continuous burden are less likely to participate in the later rounds of interviews (Hoogendoorn and Sikkel 1998; Kleinert et al. 2021; McCarthy et al. 2006). Continued burden can be measured directly (e.g., Yan et al. 2020) or indirectly through similar proxy measurements for initial burden (e.g., the number of contact attempts needed to complete a subsequent interview round).

At each stage or time point, factors external to the survey and question features of the survey affect response burden as the sources or causes of response burden; they can be considered as objective or actual burden. In addition, respondent characteristics may contribute to response burden directly, or they may moderate or mediate the relationship between the external survey and question characteristics (i.e., objective burden) and the perception of response burden at that assessment. As shown in Figure 1 (and further elaborated below and summarized in Table 1), different factors come into play at different time points in the continuum influencing the perception of response burden, leading to different decisions and perceptions on the respondents' part.

Furthermore, initial burden is expected to influence cumulative burden; respondents with a higher level of initial burden are more likely to experience a higher level of cumulative burden than those feeling less burdened at the initial survey request. The response continuum model postulated by Yan and Curtin (2010) provides support to the impact of initial burden on cumulative burden. The response continuum model predicts that reluctant respondents who are more likely to *not* respond to a survey request are also more likely to *not* answer survey items. The model also predicts that respondents who have more missing data are more likely to *not* respond to the next wave of interviews, suggesting that cumulative burden also influences continued burden. Once respondents agree to participate in the next round of an interview, the perception of continued burden will affect the cumulative burden of this new interview round.

This conceptual framework applies to surveys of all topics and of all modes of data collection. That is, survey and question features as well as respondent characteristics are expected to separately and jointly influence response burden across different survey designs. However, specific characteristics influencing response burden could vary by survey and by mode of data collection (Yan et al. 2020).

We draw on existing conceptual models of response burden and empirical literature to identify factors contributing to response burden as well as factors moderating or mediating the perception of response burden. Previous research demonstrates that these factors have either direct relation with response burden or have indirect effects on behavioral outcomes as a result of response burden (e.g., decision to participate in the survey, decision to optimize, and/or decision to take part in the later rounds of the interview). We expand our conceptualization to include and discuss these additional influences below.

4.1. Factors Contributing to Initial Burden

There are three survey characteristics that we identify as the major contributors to respondents' perception of initial burden about the survey request, which directly affects the decision to participate in a survey and are shown in Figure 2. While Figure 1 also recognizes the possible influence of question characteristics, these are only relevant for surveys where respondents can view the questions ahead of time (e.g., mail surveys) and require processing or an investment of effort on the part of respondents. Because of these, we focus on survey design features that are often made available to respondents.

Research demonstrates that the *advertised interview length* is negatively associated with response rates to web surveys (Crawford et al. 2001; Heerwegh and Loosveldt 2006; Marcus et al. 2007; Galesic and Bosnjak 2009; Yan et al. 2011) and to telephone surveys (Collins et al. 1988; Hansen 2007). In addition, advertised interview length is positively associated with breakoffs in the web surveys (Galesic 2006; Galesic and Bosnjak 2009), even though it did not directly affect perceived burden in one non-probability Web study (Galesic 2006). Yu et al. (2015) manipulated the framing of a main study after a screener questionnaire was completed. A random third of respondents were told that they were selected into a long survey and would need to complete more questions. Another random third were told the opposite – they were selected into a short survey and would need to complete fewer questions, whereas the last third were not told anything about the length of the main interview. Not surprisingly, those who were told that they would answer more

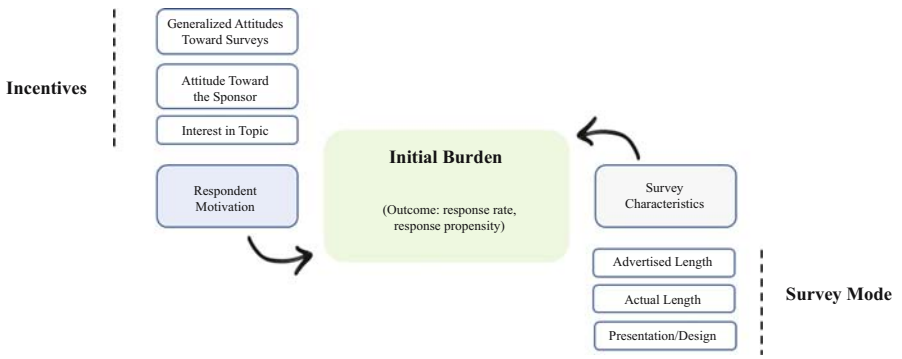


Fig. 2. Survey features and respondent characteristics influencing initial burden.

items found the survey more burdensome than those who were not told about the length of the main interview and those who were told to complete fewer items (Yu et al. 2015).

Actual interview length (in terms of number of pages in a paper questionnaire, number of survey items, or actual time spent on the task) is negatively related to response rates to mail surveys (Burchell and Marsh 1992; Dillman et al. 1993; Bogen 1996; Edwards et al. 2002; Burdein 2013) and positively related to breakoffs in the web surveys (Yan et al. 2011). The number of pages of a questionnaire and the number of survey items is found to be significantly related to respondents' perception of burden (Bottone et al. 2018). But actual time spent is only weakly related to data quality in a business survey (Lund and Haradelsen 2015) and unrelated to subjective burden (Yu et al. 2015; Read 2019).

By contrast, better *survey presentation and design* (e.g., how a mail survey looks in terms of color, spacing, and layout) has been shown to have a positive impact on response rates (e.g., Diaz de Rada 2005; Dillman et al. 2014).

Two survey design features do not necessarily directly lead to initial burden, but we recognize their impact on initial burden through their impact on respondent motivation and ability. *Incentives* are found to successfully increase respondent motivation to participate in a survey request (Mercer et al. 2015). *Mode of data collection* could also affect a respondents' motivation to participate by changing the saliency of the perceived initial burden. For instance, respondents may exhibit differential motivation toward participating in a web survey. The inability to review the survey content (versus a paper survey) requires the respondent to rely on the practitioners' assessment of the survey length or relevance to the survey topic. The lack of being able to preview content in advance may also be beneficial, as the survey content may be unappealing or viewed as not relevant to the survey topic. For interviewer-administered modes, the interviewer may overcome barriers to motivation. However, in both circumstances, mode potentially affects motivation, but does not directly change initial burden. Mode could affect response burden through their interaction with respondent ability. For instance, older respondents who are less comfortable with advanced technology such as smartphone apps; as a result, they are less likely to participate in a survey involving the use of a smartphone app (Jäckle et al. 2019).

Figure 2 shows three respondent level motivation-related characteristics that are hypothesized to affect the perception of the initial burden. Again, Figure 1 recognizes the influence of respondent ability on the initial stages of survey participation (as described in

the prior paragraph). Here, we focus on motivation-related variables that may have a main effect on response burden and interact with advertised interview length, actual survey length, and survey presentation and design in affecting response burden. *Interest* in a survey topic is a significant contributor to sample members' decision to take part in the survey (Baumgartner and Rathbun 1997; Groves et al. 2004; Groves et al. 2006) and is associated with fewer breakoffs in web surveys (Galesic 2006). It may also reduce the impact of actual survey length on the perception of burden. Similarly, a positive *attitude toward the sponsor* of a survey affects respondents' initial decision to participate in the survey; surveys sponsored by government agencies have higher response rates than surveys sponsored by non-government agencies (Heberlein and Baumgartner 1978; Presser et al. 1992). Again, respondents who have a positive view of the survey sponsor may tolerate longer surveys better than those with a negative view. Furthermore, respondents are more likely to respond to surveys when they support the sponsor of the survey (Groves et al. 2012). *Generalized attitudes toward surveys* (such as views about the usefulness and values of surveys, legitimacy of surveys, invasiveness of surveys, and trust in survey organizations and so on) affect potential respondents' likelihood to participate in surveys; those respondents with a positive attitude toward surveys, compared with those who have a more critical position, have been found to have participated more often in surveys in the past (Sharp and Frankel 1983; Goyder 1986; Bergman and Brage 2008; Giesen 2012) and to have a lower level of response burden (Yan et al. 2020). Again, these attitudes may interact with survey characteristics like length and design to affect response burden.

4.2. *Factors Contributing to Cumulative Burden*

The middle part of Figure 1 represents cumulative burden, which influences respondents' decisions to break off, to skip a question, to satisfice or optimize, or to provide poor quality data (Galesic 2006; Yan et al. 2014; Yan et al. 2020). Outcomes of cumulative burden include breakoffs, missing data, measurement error, quality of data, and response times. Cumulative burden is affected by three question characteristics, two survey characteristics, and two respondent characteristics related to ability. These characteristics are shown in Figure 3.

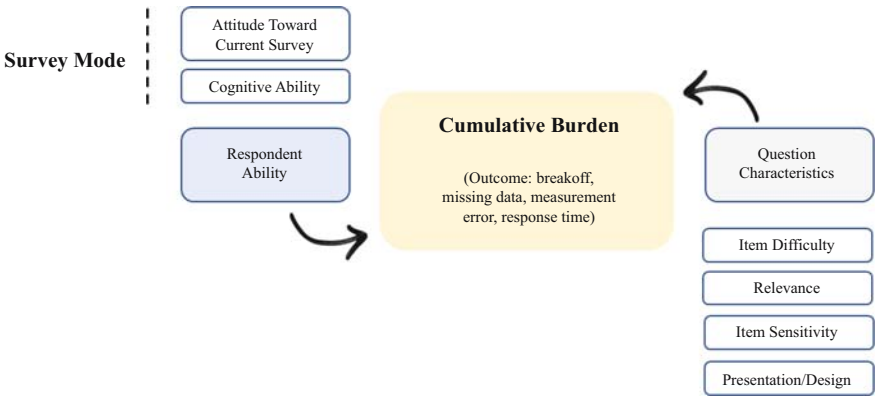


Fig. 3. Question and respondent characteristics influencing cumulative burden.

Item difficulty is measured in many different ways. Longer questions in terms of the number of words, the number of clauses or sentences, and the number of response options take longer to answer (Yan and Tourangeau 2008; Couper and Kreuter 2013), are more prone to response order effects (Schuman and Presser 1996; Holbrook et al. 2007), and are more likely to induce breakoffs (Peytchev 2009). Question text features making a question difficult to understand lead to an increase in cognitive burden (Lenzner et al. 2010). Questions requiring extensive retrieval and integration are subject to longer processing time (Yan and Tourangeau 2008) and more item nonresponse (e.g., Yan et al. 2010). In addition, open-ended questions requiring verbal input are associated with higher breakoffs (Galesic 2006; Peytchev 2009). The difficulty of recall also increases response burden (Yan et al. 2020). On the other hand, psychological studies of task complexity have often found that people are more motivated to complete tasks when they are intricate, challenging, and enriching (e.g., Taylor 1981; Campbell 1988; Maynard and Hakel 1997). Malhotra (2009) demonstrated that a complex ranking task is less prone to response order effects than a simple rating task. *Item sensitivity* tends to increase missing data (Yan et al. 2010; Tourangeau et al. 2000) and measurement error (Tourangeau et al. 2000; Tourangeau and Yan 2007; Kreuter et al. 2008; Sakshaug et al. 2010). Apparent *relevance* of the survey items refers to how strongly the survey items are related to the stated survey topic. Williams and colleagues found that respondents did not mind more survey items in a screener questionnaire when the additional items are clearly relevant to the survey topic (Williams et al. 2016). Apart from their main effects on response burden and other survey outcomes, these variables may interact with the characteristics of the survey to affect response burden.

For self-administered surveys (such as web and mail surveys), poor *survey presentation and design* (e.g., a complex grid on a web screen) has been shown to induce breakoffs (e.g., Peytchev 2009).

We recognize the role that *mode of data collection* has on data quality, but this is not a direct effect on cumulative burden. The effect on data quality is through influencing other direct factors (e.g., respondents' motivation, attitudes about the survey and so on). For instance, a self-administered mode of data collection reduces social desirability reporting than an interviewer-administered survey (Tourangeau and Yan 2007). Further, there is less differentiation in responses in a web survey than a telephone survey (Fricker et al. 2005).

In terms of respondent characteristics related to respondent ability, *cognitive ability* is found to be associated with the level of response burden in answering survey questions (Krosnick 1991; 1999). Everything else being equal, people with a lower level of cognitive ability are more likely to feel burdened than those with a higher level of cognitive ability (Atkinson et al. 2019); they are more likely to take cognitive shortcuts (Krosnick 1991; 1999), take more time to come up with answers (Yan and Tourangeau 2008), are more likely to provide missing data (Yan and Curtin 2010), and are more likely to exhibit response order effects (Holbrook et al. 2007; Malhotra 2008). Age and education are commonly used as measures of cognitive ability (see Salthouse 1991) in the survey literature. *Attitudes toward the current survey* (e.g., perceived difficulty of the survey, perceived effort required, perceived length, and so on) are thought to affect response burden, with more negative attitudes leading to higher levels of response burden (Bradburn 1977; Giesen 2012; Yan et al. 2020). For instance, Yan et al. (2020) showed

that respondents with a negative perception of a survey (e.g., those who found the survey to be less interesting, more difficult, too long, and were asked to do multiple rounds of interviews) are more likely to be "burdened out" than those with less negative perception. In addition, these attitudes mediate the relationship between survey effort and response burden. Furthermore, those with positive attitudes about the current survey tend to have less missing data than those with negative attitudes (Stocke 2006).

4.3. *Factors Contributing to Continued Burden*

The rightmost part of Figure 1 pertains to longitudinal surveys when respondents are invited back to participate in additional rounds of interviews (e.g., panel surveys). Continued burden affects respondents' decision to agree to future survey requests, leading to panel attrition. We hypothesize that two characteristics of the current survey request (comparative length and framing of the survey request) contribute to continued burden, shown in Figure 4.

For many panel surveys, the initial interview is the most burdensome because more information is gathered in that interview. Subsequent interviews may be shorter because the respondent is only requested to update his or her status on many variables since the prior interview. Other information collected such as demographic or household characteristics are unlikely to have changed or can be confirmed in the interview. We suspect that the length of the current round's interview is usually compared with that of the first round (*comparative length*). If respondents' experience with the first round is pleasant (i.e., the first round is short and easy), participation in subsequent requests is more likely, even if the rounds are longer than the initial request. On the other hand, if respondents are told that the length of the new survey is shorter, they may be more likely to participate.

Message *framing* has been used to emphasize the potential losses from not acting to encourage participation. Largely based on prospect theory (Kahneman and Tversky 1979), it has been used to encourage parents to increase physical activity in their children (Drouin et al. 2018), and to participate in additional survey request. Tourangeau and Ye (2009) used this theory to emphasize the loss (versus the gain) in value of the information that the respondent had already provided if the respondent stopped participating to increase the

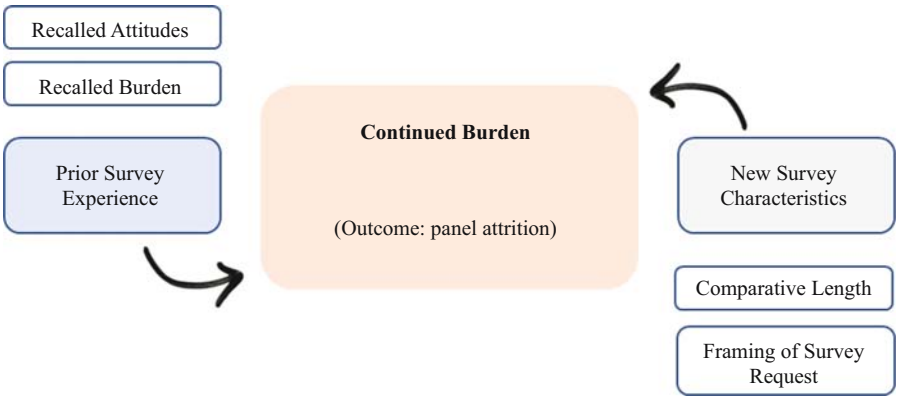


Fig. 4. Survey and respondent characteristics influencing continued burden.

likelihood that the respondent would complete a second interview. Emphasizing these sunk costs may reduce the impact of the anticipated effort in the next round on decisions to take part or not.

The left part of Figure 4 shows that *recalled burden* and *recalled attitudes to the survey*, which are consequences of respondents' prior experience with the survey, affect how they react to a new or subsequent interview request. For instance, respondents who feel overburdened have been shown to be less likely to participate in a new survey request (Hoogendoorn and Sikkel 1998; Martin et al. 2001; Bergman and Brage 2008). In addition, those who were burdened at the previous interview were found to develop more negative attitudes to the survey (Stocke and Langfeldt 2004), which subsequently influences respondents' perception of the level of burden associated with the next round of interviews (Fricker et al. 2012; Yan et al. 2020). Respondents who had a bad experience with the prior interview (such as a bad test score in the last round of interview) were less likely to participate and take the test again in the next interview round (Kleintert et al. 2021).

The peak-end theory explains how recalled burden and recalled attitudes are formed and how they can influence response burden and continued participation in longitudinal surveys. According to peak-end theory, the evaluation of an experience is based on the most extreme moment and the final moments of that experience (Redelmeier and Kahneman 1996). The peak-end theory has been used to explain how the perception of the final moments of the experience can color the entire experience and lead people to prefer longer periods of discomfort that end on a more positive note to shorter periods of discomfort (see Kahneman 2000, for a review). A key component of this theory is that the sheer duration of the event is discounted in the evaluation. This peak-end model is pertinent to the recall of the cumulative burden from an earlier interview round in a longitudinal survey context. Applied to our burden framework, we believe that a persons' evaluation of response burden is constructed based on easily recalled features of the survey experience such as the duration of the survey and the difficulty of the survey items and so on. The final (that is, the most recent) experience and the peak (worst or best) experience with these features in the prior wave of a longitudinal survey affects response burden. As a result, the theory hypothesizes that if the earlier survey begins with the most burdensome questions and ends with relatively easy or interesting questions, that earlier survey will be recalled as less burdensome than if the questions exhibit the opposite organization (starting with easy and ending with more burdensome questions). That difference in recalled burden may affect respondents' perception and the decision to continue with the survey request. As mentioned earlier, Yu et al. (2015) provided some evidence to this model by manipulating, at the end of a short screener interview, how the main study was described. Some were told that they would complete a long survey ("screen in" condition) whereas others told that they would experience a short survey ("screen out" condition). Respondents in the "screen in" condition rated the survey as longer and more burdensome than those in the "screen out" condition, regardless of the actual length of the survey and the actual time spent on the survey. In addition, those in the "screen in" condition were also more likely to breakoff at the main interview than those in the "screen out" condition. Apparently, the negative experience from the ending message of a short screener interview affected how respondents perceived the main interview.

5. Using the Dynamic Conceptual Framework of Response Burden to Unify Existing Research and to Motivate Future Research

Prior research on response burden has relied on incomplete conceptualizations and unsystematic measures of response burden. Consequently, it is not surprising that empirical research on response burden sometimes produces equivocal findings, further making it difficult to draw firm predictions about survey features or respondent characteristics that are most likely to give rise to response burden. For instance, researchers and practitioners too often believe that longer surveys (one of the most frequently used indicators of actual burden) impose greater response burden and yield lower response rates. However, the evidence is at best mixed on the relationship between interview length and response rates. When length is measured in terms of number of pages, only six (out of the 25) empirical studies included in a meta-analysis supported the speculation that longer surveys yield lower response rates (Rolstad et al. 2011). The rest of the 19 surveys failed to bear evidence for the assumed negative impact of survey length on response rates. Survey length is a common measure of actual burden used by regulatory agencies, however, from these findings it is not clear whether or not survey length actually does lead to response burden and lower response rates and what the survey field (and regulators) should do with regard to survey length.

The mixed evidence results from the fact that researchers only looked at the bivariate relationship between interview length as a measure of response burden and response rate (as an outcome of response burden), overlooking the many factors that could mediate and moderate the relationships between interview length (a measure of actual burden), response burden, and response rate as a consequence of response burden. To advance the research on the relationship between response burden and survey participation, our conceptual framework calls for attention to the moderators and mediators and provides suggestions for the additional factors that a researcher could look into in multivariate analyses. Factors displayed in Figure 2 are great candidates. For instance, it is not enough to focus on the bivariate relationships between, for example, interview length and response rates, or survey interest and response rates. The moderators and mediators (such as respondents' motivation, generalized attitudes towards surveys, attitudes towards the sponsor and the survey topic) all need to be taken into consideration. Predictions could be drawn using the conceptual framework. A long survey would be more unbearable to those with negative attitudes towards surveys, lower trust of sponsors, and disinterest in the topic than those with positive attitudes and higher regard of sponsors.

Yan et al. (2020) empirically demonstrated that, although the objective burden (e.g., interview length) has a significant direct impact on response burden, this direct effect is canceled out by the indirect effect of respondents' perception of the survey, producing small and non-significant overall effects on response burden. The seemingly contradictory empirical evidence that a longer questionnaire sometimes leads to a lower response rate and sometimes has no impact on response rate can be accounted for and unified under our dynamic conceptual framework of response burden, which examines the relationship between survey features and respondent characteristics.

The conceptual framework is useful in explaining existing research evidence. More importantly, it is useful in motivating future research. Researchers can choose to focus on

one part of the framework (e.g., what can be done to reduce panel attrition?) or look at the dynamic nature of response burden (e.g., how offering an incentive affects initial burden and continued burden at the next survey request?).

6. Conclusions and Discussion

This article reviews the state of the literature on conceptualization and measurement of response burden in the four decades of research. The review reveals that response burden is under-conceptualized and is measured in various ways, indicating the need for an updated framework. Our theoretical framework builds on the concept of burden as a subjective phenomenon, affected by respondents' psychological responses to various elements of the survey.

Our model has several innovative features. First, it distinguishes three forms of response burden: initial burden; cumulative burden within a survey; and continued burden across waves of a longitudinal survey. Each has distinctive causes, moderators, and mediators, and consequences.

Second, it attempts to synthesize empirical research to more completely account for a full range of variables thought to affect response burden. The impact of these variables on response burden is summarized in Table 1 and can be used to inform new survey designs and to encourage new research.

Third, despite its elaboration and comprehensiveness, the gist of the framework applies to all modes of data collection, all topics, and all samples. Yan et al. (2020) demonstrate that the modes of data collection did not affect the relationship between the underlying factors and perception of response burden; the same set of factors have the same impact on response burden regardless of whether respondents were attempted mostly by phone or in person. Of course, indicators of the latent factors will differ by the nature of the survey.

Fourth, due to its wide application across different contexts, our framework can be used to derive specific testable hypotheses on the impact of various design features and respondent characteristics on response burden. The framework can also be used to summarize and consolidate apparent inconsistent empirical findings.

Finally, the framework also provides practical guidance and generates strategies to reduce response burden. For instance, our model postulates the mediating and moderating effects of respondent perceptions on response burden at different stages, suggesting the importance of paying attention to strategies or methods that are able to improve respondents' perception of the survey, of the sponsor, and of surveys in general, for instance. Yan et al. (2020) advocates for including an evaluation of respondents' perceptions about surveys in the question testing and evaluation stage. Our model provides theoretical support to their suggestion. Another example is to the placement of survey items in a longitudinal setting. Questionnaire design textbooks recommend starting with easy and interesting questions and placing more difficult and sensitive questions at the end of a questionnaire to avoid breakoffs (e.g., Bradburn et al. 2004). However, this recommendation may not be beneficial for longitudinal surveys as predicted by peak-end theory. There may be benefit to manipulating the placement of survey items in order to influence recalled burden.

Our conceptual framework also draws upon psychological theory to understand and form predictions for how response burden is driven by perceptions. However, more

Table I. Summary of impact of variables on response burden.

	Variable	Impact on response burden
Survey design features	Advertised interview length	Announcement of a long interview is shown to increase perception of initial burden
	Actual interview length	Longer interview is shown to increase perception of initial burden
	Survey presentation and design	Better survey presentation and design is shown to reduce initial burden and cumulative burden
	Incentive	Incentives are shown to reduce initial burden through motivation
	Mode	Mode is shown to affect initial burden and cumulative burden through interaction with respondent ability and motivation
	Comparative length	Shorter interview in the earlier round or announced in the new round is expected to reduce continued burden
	Framing of survey request	Loss framing is shown to reduce continued burden
Question characteristics	Item difficulty	Difficult items are shown to increase cumulative burden
	Item sensitivity	Sensitive items are shown to increase cumulative burden
	Item relevance	Relevant items are shown to reduce impact of survey length on cumulative burden
Respondent characteristics	Interest	Interest in survey is shown to reduce initial burden and shown to reduce negative impact of survey characteristics (e.g., survey length) on initial burden and cumulative burden
	Attitude toward the sponsor	Positive attitude toward sponsor is shown to reduce initial burden and is expected to reduce negative impact of survey characteristics (e.g., survey length) on initial burden and cumulative burden
	Generalized attitudes toward surveys	Positive attitudes toward surveys are shown to reduce initial burden and are expected to reduce impact of announced and actual survey length on initial and cumulative burden
	Cognitive ability	Cognitive ability is shown to reduce cumulative burden and shown to interact with mode on initial burden
	Attitudes toward the current survey	Positive attitudes toward the current survey is shown to reduce cumulative burden and shown to mediate the relationship between survey features and cumulative burden
	Recalled burden	Recalled burden is shown to increase continued burden
	Recalled attitudes	Recalled negative attitudes are shown to increase continued burden

research is needed to better understand the limitations of these theories. For instance, in the context of peak-end theory, placing the least cognitively demanding, or the most interesting questions at the end of a survey just to end with a positive experience benefits the recalled burden of a prior survey. But when respondents start a subsequent interview and encounter questions where the burden is now greater than what the recalled experience from the prior interview, will it result in an increase in breakoffs or satisficing strategies? In other words, does influencing a positive evaluation to possibly increase subsequent cooperation simply result in breakoffs or poor data quality? Prospect theory may have broader applications, for instance, influencing initial burden. Using the conceptual framework to structure testable hypothesis will do much to inform the survey field.

Finally, our model assumes that response burden is a measurable subjective phenomenon, supporting the direct question approach that asks people how burdensome the survey is to them. This direct question approach is used in several studies (e.g., Galesic 2006; Yan et al. 2020) and is found to be a useful and promising measure of burden. When direct questioning is not feasible, we encourage researchers to use auxiliary information such as paradata to measure response burden indirectly, and examine correlates and moderators of response burden. However, in that case, researchers need to be explicit as to their exact measurement and conceptualization of response burden.

7. References

Atkinson, T.M., C.E. Schwartz, L. Goldstein, I. Garcia, D.F. Storfer, Y. Li, J. Zhang, B.H. Bochner, and B.D. Rapkin. 2019. "Perceptions of Response Burden Associated with Completion of Patient-Reported Outcome Assessments in Oncology." *Value in Health* 22: 225–230. DOI: https://doi.org/10.1016/j.jval.2018.07.875.

Baumgartner, R., and P. Rathbun. 1997. "Prepaid Monetary Incentives and Mail Survey Response Rates." Paper presented at the annual conference of the American Association for Public Opinion Research, May 15–18, 1997, Norfolk, VA, USA.

Bergman, L.R., and R. Brage. 2008. "Survey Experience and Later Survey Attitudes, Intention, and Behavior." *Journal of Official Statistics* 24 (1): 99–31. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/survey-experiences-and-later-survey-attitudesintentions-and-behaviour.pdf.

Bogen, K. 1996. "The Effect of Questionnaire Length on Response Rates: A Review of the Literature." In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association: 1020–1025. Available at: http://www.asasrms.org/Proceedings/papers/1996_177.pdf (accessed November 2021).

Bottone, M., L. Modugno, and A. Neri. 2018. "Response Burden and Data Quality in Business Surveys: The Experience of Banca d'Italia." *Occasional Papers No. 467*. Available at: https://www.bancaditalia.it/pubblicazioni/qef/2018-0467/QEF_467_18.pdf?language_id=1 (accessed November 2021).

Bradburn, N. 1977. "Respondent Burden." *Health Survey Research Methods Proceedings.* 49–53. Available at: https://hsrmconference.com/sites/default/files/proceedings/HSRMProceedings02.pdf (accessed November 2021).

Bradburn, N., S. Sudman, and B. Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires (Research Methods for the Social Sciences).* Wiley and sons.

Burchell, B., and C. Marsh. 1992. "The Effect of Questionnaire Length on Survey Response." *Quality & Quantity* 26: 233–244. DOI: https://doi.org/10.1007/BF00172427.

Burdein, I. 2013. "Shorter Isn't Always Better." Paper presented at CASRO Online Research Conference, March 7–8, San Francisco, CA, U.S.A.

Campbell, D.J. 1988. "Task Complexity: A Review and Analysis." *Academy of Management Review* 13: 40–52.

Chapin, G. 1920. *Field Work and Social Research*. New York: The Century Co. Commission on Federal Paperwork.

Collins, M., W. Sykes, P. Wilson, and N. Blackshaw. 1988. "Nonresponse, the UK experience." In *Telephone Survey Methodology,* edited by R.M.Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg. New York: Wiley.

Couper, M.P,. and F. Kreuter. 2013. "Using Paradata to Explore Item Level Response Times in Surveys." *Journal of the Royal Statistical Society* 176: 271–286. DOI: https://doi.org/10.1111/j.1467-985X.2012.01041.x.

Crawford, S.D., M.P. Couper, and M.J. Lamias. 2001. "Web Surveys: Perception of Burden." *Social Science Computer Review* 19: 146–162. DOI: https://doi.org/10.1177/089443930101900202.

Diaz de Rada, V. 2005. "Influence of Questionnaire Design on Response to Mail Surveys." *International Journal of Social Research Methodology* 8: 61–78. DOI: https://doi.org/10.1080/1364557021000025991.

Dillman, D.A., M.D. Sinclair, and J.R. Clark. 1993. "Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys." *Public Opinion Quarterly* 57: 289–304. DOI: https://doi.org/10.1086/269376.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* Hoboken, NJ: John Wiley Sons.

Drouin, O., M. Young, and N. King. 2018. "Message framing and counseling of parents on children's physical activity – an experimental study." *Health Psychology and Behavioral Medicine* 6(1): 214–225. DOI: 10.1080/21642850.2018.1515018.

Edwards, P., I. Roberts, M. Clarke, C. DiGuiseppi, S. Pratap, R. Wentz, and I. Kwan. 2002. "Increasing Response Rates to Postal Questionnaires: Systematic Review." *BMJ* 324: 1183–1190. DOI: https://doi.org/10.1136/bmj.324.7347.1183.

Eurostat. 2011. *European statistics code of practice for the national and community statistical authorities*. Adopted by the European Statistical System Committee 28th September 2011. Available at: https://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15 (accessed November 2021).

Filion, F.L. 1981. "Importance of Question Wording and Response Burden in Hunter Surveys." *The Journal of Wildlife Management* 45: 873–882. DOI: https://doi.org/10.2307/3808096.

Frankel, J. 1980. *Measurement of respondent burden: Study Design and Early Findings.* Bureau of Social Science Research, Inc., Washington, D.C. Available at: https://files.eric.ed.gov/fulltext/ED198179.pdf.

Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 69: 370-392. DOI: https://doi.org/10.1093/poq/nfi027.

Fricker, S., J. Gonzalez., and L. Tan. 2011. "Are You Burdened? Let's Find Out." Paper Presented at the Annual Conference of the American Association for Public Opinion Research, May 12–15, Phoenix, AZ, U.S.A.

Fricker, S., C. Kreisler., and L. Tan. 2012. "An Exploration of the Application of PLS Path Modeling Approach to Creating a Summary Index of Respondent Burden." in *JSM Proceedings*, Survey Research Methods Section. 27 July – 2 August Alexandria, VA, USA. American Statistical Association: 4141–4155. Available at: http://www.asasrms.org/Proceedings/ y2012f.html (accessed November 2021).

Galesic, M. 2006. "Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics*, 22 (2): 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf.

Galesic, M., and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73: 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Geisen, D. 2012. "Exploring Causes and Effects of Perceived Response Burden." Paper presented at the International Conference on Establishment Surveys, June 11–14, 2021, Montreal, Quebec, Canada.

Giesen, D., M. Vella., C.F. Brady., P. Brown., D. Ravindra., and A. Vaasen-Otten. 2018. "Response Burden Management for Establishment Surveys at Four National Statistical Institutes." *Journal of Official Statistics* 34: 397–418. DOI: https://doi.org/10.2478/jos-2018-0018.

Goyder, J. 1986. "Surveys on Surveys: Limitations and Potentialities." *Public Opinion Quarterly* 50: 27–41. DOI: https://doi.org/10.1086/268957.

Graf, I. 2008. "Respondent burden." In *Encyclopedia of survey research methods*, edited by P. J. Lavrakas: 739–740. Sage Publications. Available at: https://dx.doi.org/10.4135/9781412963947.n477.

Groves, R., E. Singer, and A. Corning. 1999. "A Laboratory Approach to Measuring the Effects on Survey Participation of Interview Length, Incentives, Differential Incentives, and Refusal Conversion." *Journal of Official Statistics* 15 (2): 251–268. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-laboratory-approach-to-measuring-the-effects-on-survey-participation-of-interview-length-incentives-differential-incentives-and-refusal-conversion.pdf.

Groves, R.M., M.P. Couper, S. Presser, E. Singer, R. Tourangeau, G.P. Acosta, and L. Nelson. 2006. "Experiments in Producing Nonresponse bias." *Public Opinion Quarterly* 70: 720–736. DOI: https://doi.org/10.1093/poq/nfl036.

Groves, R,M., S. Presser., and S. Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly* 68: 2–31. DOI: https://doi.org/10.1093/poq/nfh002.

Groves, R.M., S. Presser, R. Tourangeau, B. West, M.P. Couper, E. Singer, and C. Toppe. 2012. "Support for the Survey Sponsor and Nonresponse Bias." *Public Opinion Quarterly* 76: 512–524. DOI: https://doi.org/10.1093/poq/nfs034.

Hansen, K.M. 2007. "The Effects of Incentives, Interview length, and Interviewer characteristics on Response Rates in a CATI-study." *International Journal of Public Opinion Research* 19: 112–121. DOI: https://doi.org/10.1093/ijpor/edl022.

Haraldsen, G. 2004. "Identifying and Reducing Response Burden in Internet Business Surveys." *Journal of Official Statistics* 20 (2): 393–410. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/identifying-and-reducing-response-burdens-in-internet-business-surveys.pdf.

Haraldsen, G., J. Jones, D. Giesen, and L. Zhang. 2013. "Understanding and Coping with Response Burden." In *Designing and Conducting Business Surveys*, edited by G Snijkers, G. Haraldsen, J. Jones, and D.K. Willimack: 219–252. Hoboken, NJ: John Wiley & Sons.

Heberlein, T.A., and R.M. Baumgartner. 1978. "Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature." *American Sociological Review* 43: 447–462. DOI: https://doi.org/10.2307/2094771.

Hedlin, D., H. Lindkvist, H. Bäckström, and J. Erikson. 2008. "An Experiment on Perceived Survey Response Burden Among Businesses." *Journal of Official Statistics*, 24 (2): 301–318. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-experiment-on-perceived-survey-response-burden-among-businesses.pdf.

Heerwegh, D., and G. Loosveldt. 2006. "An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey Sponsor Logos in Web Surveys." *Journal of Official Statistics* 22 (2): 191–210. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-experimental-study-on-the-effects-of-personalization-survey-length-statements-progress-indicators-and-survey-sponsor-logos-in-web-surveys.pdf.

Holbrook, A.L., J.A. Krosnick, D. Moore, and R. Tourangeau. 2007. "Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes." *Public Opinion Quarterly* 71: 25–348. DOI: https://doi.org/10.1093/poq/nfm024.

Hoogendoorn, A.W. 2004. "A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing." *Journal of Official Statistics*, 20 (2): 219–232. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-questionnaire-design-for-dependent-interviewing-that-addresses-the-problem-of-cognitive-satisficing.pdf.

Hoogendoorn, A.W. and D. Sikkel. 1998. "Response Burden and Panel Attrition." *Journal of Official Statistics* 14 (2): 189–205. Available at: https://www.scb.se/contentassets/-ca21efb41fee47d293bbee5bf7be7fb3/response-burden-and-panel-attrition.pdf.

Jäckle, A., J. Burton, M. Couper, and C. Lessof. 2019. "Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: Response rates and response biases." *Survey Research Methods* 13(1): 23–45. DOI: https://doi.org/10.18148/srm/2019.v1i1.7297.

Kahneman, D. 2000. "Evaluation by Moments, Past and Future." In *Choices, Values and Frames*, edited by Kahneman, D., and Tversky A: 693–708. New York: Cambridge University Press and the Russell Sage Foundation.

Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Decisions under Risk." *Econometrica* 47: 263–291. DOI: https://doi.org/10.2307/1914185.

Kleinert, C., Christoph, B., and Ruland, M. 2021. Experimental Evidence on Immediate and Long-term Consequences of Test-induced Respondent Burden for Panel Attrition. Sociological Methods & Research, 50(4): 1552–1583. DOI: https://doi.org/10.1177/0049124119826145.

Kreuter, F., S. Presser, and R. Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly* 72: 847–865. DOI: https://doi.org/10.1093/poq/nfn063.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. DOI: https://doi.org/10.1002/acp.2350050305.

Krosnick, J.A. 1999. "Survey Research." *Annual Review of Psychology* 50: 537–567. DOI: https://doi.org/10.1146/annurev.psych.50.1.537.

Lenzner, T., L. Kaczmirek, and A. Lenzner. 2010. "Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment." *Applied Cognitive Psychology* 24: 1003–1020. DOI: https://doi.org/10.1002/acp.1602.

Lund, K., and G. Haraldsen. 2015. "On the Relationship between Response Burden and Data Quality: A Complementary Study of Response Burden and Data Quality in Official Establishment Surveys." Paper Presented at the QUEST 2015 Workshop, April 28–30, Helsinki, Finland. Available at: https://wwwn.cdc.gov/qbank/QUest/2015/s121_karianne_lund_quest2015.pdf.

Malhotra, N. 2008. "Completion Time and Response Order Effects in Web Surveys." *Public Opinion Quarterly* 72: 914–934. DOI: https://doi.org/10.1093/poq/nfn050.

Malhotra, N. 2009. "Order Effects in Complex and Simple Tasks." *Public Opinion Quarterly* 180–198. DOI: https://doi.org/10.1093/poq/nfp008.

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schutz. 2007. "Compensating for Low Topic Interest and Long Surveys: A Field Experiment on Nonresponse in Web Surveys." *Social Science Computer Review* 25: 372–383. DOI: https://doi.org/10.1177/0894439307297606.

Martin, E., D. Abreu, and F. Winters. 2001. "Money and Motive: Effect of Incentives on Panel Attrition in the Survey of Income and Program Participation." *Journal of Official Statistics*, 17 (2): 267–284. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/money-and-motive-effects-of-incentives-on-panel-attrition-in-the-survey-of-income-and-program-participation.pdf.

Maynard, D.C., and M.D. Hakel. 1997. "Effects of Objective and Subjective Task Complexity on Performance." *Human Performance* 10: 3–31. DOI: https://doi.org/10.1207/s15327043hup1004_1.

McCarthy, J.S., D.G. Beckler, and S.M. Qualey. 2006. "An Analysis of the Relationship Between Survey Burden And Nonresponse: If we bother them more, are they less cooperative?" *Journal of Official Statistics* 22 (1): 97–112. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-analysis-of-the-relation-

ship-between-survey-burden-and-nonresponse-if-we-bother-them-more-are-they-less-cooperative.pdf.

Mercer, A., A. Caporaso, D. Cantor, and R. Townsend. 2015. "How Much Gets You How Much? Monetary Incentives and Response Rates in Household Surveys." *Public Opinion Quarterly* 79: 105–129. DOI: https://doi.org/10.1093/poq/nfu059.

Neuert, C. 2020. "How Effective Are Eye-Tracking Data in Identifying Problematic Questions?" *Social Science Computer Review* 38: 793–802. DOI: https://doi.org/10.1177/0894439319834289.

Peytchev, A. 2009. "Survey Breakoff." *Public Opinion Quarterly* 73: 74–97. DOI: https://doi.org/10.1093/poq/nfp014.

Presser, S., J. Blair, and T. Triplett. 1992. "Survey Sponsorship, Response Rates, and Response Effects." *Social Science Quarterly* 73: 699–702. Available at: https://www.jstor.org/stable/i40108860.

Read, B. 2019. "Respondent Burden in a Mobile App: Evidence from a Shopping Receipt Scanning Study." *Survey Research Methods* 13: 45-71. DOI: https://doi.org/10.18148/srm/2019.v1i1.7379.

Redelmeier, D.A., and D. Kahneman. 1996. "Patient's Memories of Painful medical Treatments: Real-time and Retrospective Evaluations of Two Minimally Invasive Procedures." *Pain* 66: 3–8. DOI: https://doi.org/10.1016/0304-3959(96)02994-6.

Rostald, S., J. Adler, and A. Ryden. 2011. "Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis." *Value in Health* 14: 1101–1108. DOI: https://doi.org/10.1016/j.jval.2011.06.003.

Sakshaug, J.W., T. Yan, and R. Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-mode Survey of Sensitive and Non-sensitive Items." *Public Opinion Quarterly* 74: 907–933. DOI: https://doi.org/10.1093/poq/nfq057.

Salthouse, T.A. 1991. *Theoretical Perspectives on Cognitive Aging*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sharp, L.M., and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 47: 36–53. DOI: https://doi.org/10.1086/268765.

Schuman, H., and S. Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Format, Wording, and Context.* Sage Publications.

Singer, E., J. van Hoewyk, N. Gebler, T. Raghunathen, and K. McGonagle. 1999. "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." *Journal of Official Statistics* 15 (2): 217–230. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/the-effect-of-incentives-on-response-rates-in-interviewer-mediated-surveys.pdf

Stocke, V. 2006. "Attitudes toward Surveys, Attitude Accessibility and the Effect on Respondents' Susceptibility to Nonresponse." *Quality & Quantity* 40: 259–288. DOI: https://doi.org/10.1007/s11135-005-6105-z

Stocke, V., and B. Langfeldt. 2004. "Effects of Survey Experience on Respondents' Attitudes Towards Surveys." *Bulletin de Methodologie Sociologique* 81: 5–32. DOI: https://doi.org/10.1177/075910630408100103.

Taylor, M.S. 1981. "The Motivational Effects of Task Challenge: A Laboratory Investigation." *Organizational Behavior and Human Performance* 27: 255–278. DOI: https://doi.org/10.1016/0030-5073(81)90049-0.

Tourangeau, R., F.G. Conrad, and M.P. Couper. 2013. *The Science of Web Surveys*. New York, NY: Oxford University Press.

Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. New York, NY: Cambridge University Press.

Tourangeau, R., and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin*,133(5): 859–883. DOI: https://doi.org/10.1037/0033-2909.133.5.859.

Tourangeau, R., and C. Ye. 2009. "The Framing of the Survey Request and Panel Attrition." *Public Opinion Quarterly* 73: 338–348. DOI: https://doi.org/10.1093/poq/nfp021.

Warriner, G.K. 1981. "Accuracy of self-reports to the burdensome question: survey response and nonresponse error trade-off." *Quality & Quantity* 25: 253–269. DOI: https://doi.org/10.1007/BF00167531.

Wenemark, M., G.H. Frisman, T. Svensson, and M. Kristenson. 2010. "Respondent Satisfaction and Respondent Burden among Differently Motivated Participants in a Health-Related Survey." *Field Methods* 22: 378–390. DOI: https://doi.org/10.1177/1525822X10376704.

Williams, D., J.M. Brick, J. Montaquilla, and D. Han. 2016. "Effects of screening questionnaires on response in a two-phase postal survey." *International Journal of Social Research Methodology* 19: 51–67. DOI: https://doi.org/10.1080/13645579.2014.950786.

Yan, T. 2021. "An Eye-tracking Study of Scale Direction Effects." Paper presented at the annual conference of American Association of Public Opinion Research, May 11–14, Virtual conference. Available at: https://www.researchgate.net/publication/363541472_An_Eye-tracking_Study_of_Scale_Direction_Effect/citations#fullTextFileContent.

Yan, T., F.G. Conrad, R. Tourangeau, and M.P. Couper. 2011. "Should I Stay or Should I Go: The Effects of Progress Feedback, Promised Task Duration, and Length of Questionnaire on Completing Web Surveys." *International Journal of Public Opinion Research* 23: 131–147. DOI: https://doi.org/10.1093/ijpor/edq046.

Yan, T., and R. Curtin. 2010. "The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective." *International Journal of Public Opinion Research*, 22:535–551. DOI: https://doi.org/10.1093/ijpor/edq037.

Yan, T., R. Curtin, and M. Jans. 2010. "Trends in Income Nonresponse over Two Decades." *Journal of Official Statistics* 26 (1): 145–164. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/trends-in-income-nonresponse-over-two-decades.pdf.

Yan, T., S. Fricker, and S. Tsai. 2014. "The Impact of Response Burden on Data Quality in a Longitudinal Survey." Paper presented at the International Total Survey Error Workshop, October 1–3, Washington, DC, USA. Available at: https://www.researchgate.net/publication/363541311_The_Impact_of_Response_Burden_on_Data_Quality_in_a_Longitudinal_Survey.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response Burden: What is It and What Predicts It?" In *Advances in Questionnaire Design, Development, Evaluation and Testing,* edited

by P. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, and A. Wilmot, Hoboken, NJ: Wiley: 193–212.

Yan, T., L. Ryan, S. Becker, and J. Smith. 2015. "Assessing Quality of Answers to a Global Subjective Well-being Question Through Response Times." *Survey Research Methods* 9: 101–109. DOI: https://doi.org/10.18148/srm/2015.v9i2.6107.

Yan, T., and R. Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience, and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22: 51–68. DOI: https://doi.org/10.1002/acp.1331.

Yu, E., S. Fricker, and B. Kopp. 2015. "Can Survey Instructions Relieve Respondent Burden?" in *JSM Proceedings*, Survey Research Methods Section. 8–13 August Alexandria, VA: American Statistical Association: 4317–4326. Available at: http://www.asasrms.org/Proceedings/y2015f.html (accessed November 2021).

Received September 2020
Revised June 2021
Accepted December 2021

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 963–986, http://dx.doi.org/10.2478/JOS-2022-0042

Testing a Planned Missing Design to Reduce Respondent Burden in Web and SMS Administrations of the CAHPS Clinician and Group Survey (CG-CAHPS)

Philip S. Brenner[1], *J. Lee Hargraves*[2], *and Carol Cosenza*[2]

We test a planned missing design to reduce respondent burden in Web and SMS administrations of the CAHPS Clinician and Group Survey (CG-CAHPS), a survey of patient experiences widely used by health care providers. Members of an online nonprobability panel were randomly assigned to one of three invitation and data collection mode protocols: email invitation to a Web survey, SMS invitation to a Web survey, or SMS invitation to an SMS survey. Within these three mode protocols, respondents were randomly assigned to a planned missing design, which shortened the survey by about 40%, or to a control group that received the survey in its entirety. We compare survey duration, breakoff and completion rates, and five key patient experience measures across conditions to assess the effect of the planned missing design across the three modes. We found that a planned missing design worked well with our Web survey, reducing survey duration and breakoff without changing estimates relative to the full-survey control condition. However, mixed findings in the SMS survey suggest that even shortened, 15-item surveys may be too long to substantially reduce respondent burden. We conclude with recommendations for future research.

Key words: Online data collection; text messaging; health.

1. Introduction

The high demand for cost-effective survey designs has been an impetus for methodological innovation (Schonlau and Couper 2017; De Leeuw 2018). Modes that contact, recruit, and measure electronically, such as Web surveys and text message surveys, fit this bill, allowing cost-effective designs that can leverage institutional resources (e.g., licenses for Web and text survey applications) to dramatically reduce costs (Dillman et al. 2014). Yet, taking advantage of these innovations to reduce costs may encounter an unintended consequence: high respondent burden (Crawford et al. 2001; Mavletova et al. 2018). Although respondent burden lacks a clear conceptualization, it likely results from multiple factors, including, but not limited to, survey length, complexity, and the effort required for completion (Crawford et al. 2001; Mavletova et al. 2018; Yan et al. 2020). These factors, however, do not determine respondent burden as they may be balanced or even outweighed by the positive characteristics of the survey, such as the respondent's interest in or perceived importance of the topic (Bradburn 1978; Sharp and Frankel 1983).

[1] Department of Methodology and Statistics, Padualaan 14, 3584 CH Utrecht, The Netherlands. Email: p.s.brenner@uu.nl.
[2] Center for Survey Research, University of Massachusetts Boston, 100 Morrissey Blvd. Boston, MA 02125 (617) 287-7200 U.S.A. Emails: lee.hargraves@umb.edu and carol.cosenza@umb.edu

High respondent burden generated by long and complex self-administered surveys such as those conducted on the Web or by short message service (SMS), may cause fatigue and resultant breakoffs that potentially harm data quality (Galesic and Bosnjak 2009; Revilla and Ochoa 2017). While cutting questions, condensing questionnaires, and curtailing complexity may help to reduce respondent burden in this case (Mavletova and Couper 2015; Toepoel and Lugtig 2018), this remedy may cause a side-effect. Even in a shortened and simplified survey questionnaire, important and complex concepts must be adequately measured for the survey to accomplish its goals (Keller et al. 2005). If too many questions are cut or if cuts are made haphazardly and without consideration of their effect on resulting data quality, reliability and validity, especially concurrent and discriminant validities, may be harmed (Ng et al. 2016). A promising solution – a planned missing design – randomly assigns respondents to answer only a subset of questions to shorten the survey and reduce respondent burden (Johnson et al. 2013).

Patient surveys provide an illustrative example of these cross-cutting demands that may be met with a planned missing design. Data from patient surveys are used by government agencies and insurance companies in reimbursement and pay-for-performance computations and are made available to consumers for their medical decision-making (Elliott et al. 2016; Holt 2019; Spranca et al. 2000). However, many healthcare providers and professionals view patient surveys as tertiary to providing care and see high data collection costs as an unnecessary burden on their already strained budgets (Alemi and Jasper 2014). As a result, many hospitals and clinics are demanding shorter self-administered patient surveys that take advantage of methodological and technological innovations to improve cost-effectiveness and allow increased flexibility and customizability (Keller et al. 2005; Lee et al. 2013; Stucky et al. 2016).

Therefore, we test the use of a planned missing design with the CAHPS Clinician and Group Survey (CG-CAHPS), a survey of patient experiences widely used by healthcare providers. We invite members of an online panel by email or SMS to complete a survey administered in one of two modes, by Web or SMS. Panelists who have had a visit with a physician in the past six months are assigned to one of three shortened modules or to the full version of the survey. Comparisons are made between the different invitation and data collection mode conditions and planned missing modules of the survey to assess the effects of these designs on respondent burden, operationalized as survey duration and completion and breakoff rates, and key patient experience outcomes.

2. Background

2.1. Respondent Burden and Web Surveys

Web surveys can be designed to reduce respondent burden (Mavletova and Couper 2015; Toepoel and Lugtig 2018). As a computerized data collection method, Web surveys can be programmed to simplify the respondent's task by streamlining their path through the questionnaire, using survey software to hide complex routing generated by filter questions and skip patterns (Dillman et al. 2014). However, respondent burden is not necessarily alleviated by computerization and the potential for problems remains (Crawford et al. 2001). If questions are too many or too complex, respondents may cope with the high

cognitive demands by engaging in counterproductive behaviors that reduce data quality, such as straight-lining, satisficing, skipping questions, or breaking off from the survey before completion (Conrad et al. 2017; Kim et al. 2019; Zhang and Conrad 2018).

Thus, keeping Web surveys short and simple helps to reduce respondent burden and improve data quality (Mavletova and Couper 2015; Toepoel and Lugtig 2018). A planned missing design fits this need well as an effective way to reduce survey length and complexity (Peytchev and Peytcheva 2017). This approach reduces survey length by randomly assigning respondents to answer only subsets of questions (Enders 2010; Johnson et al. 2013; Graham et al. 2006). Given random assignment, unasked and unanswered questions generate nonresponse patterns in data sets that are known to be missing completely at random (Johnson et al. 2013). As a result, planned missing designs do not bias results, although they may reduce statistical efficiency for some types of analyses as they can shrink analytic sample sizes (Rhemtulla et al. 2016) – the price paid to shorten survey length, relieve respondent burden, and prevent resultant breakoffs. Thus, this study applies a planned missing design to reduce respondent burden in a Web survey.

2.2. Respondent Burden in SMS Surveys

An innovative extension to Web survey methods allows recruitment and data collection to be shifted to SMS. This approach takes advantage of many Web survey platforms' capabilities to recruit and measure using text messaging. SMS shows promise during recruitment to a Web survey, improving response when used as a pre-notification contact preceding an email invitation (Bosnjak et al. 2008) and yielding higher response than emailed invitations when used as a recruitment mode (Mavletova and Couper 2014). Importantly, using SMS as a recruitment or invitation mode in the United States requires an appropriate sampling frame that includes a cell number and permission to contact via text.

SMS also offers some potential benefits as a data collection mode including respondents' increased willingness to disclose sensitive information and a higher response rate compared to interviewer administration (Schober et al. 2015; West et al. 2015). However, SMS data collection also extends the duration of the survey due to the asynchronous nature and norms of texting, which may increase respondent burden and the potential for breakoffs (West et al. 2015). Much like for Web surveys, best practices suggest keeping SMS surveys short and simple to prevent breakoffs and errors resulting from respondent fatigue and high respondent burden (Lau et al. 2019). Therefore, SMS surveys may also benefit from a planned missing design to reduce survey length and respondent burden. Thus, this study novelly tests a planned missing design in an SMS survey.

2.3. Respondent Burden and Patient Surveys

Surveys of patients are well-positioned to benefit from innovations in recruitment and data collection, such as planned missing designs in Web and SMS surveys. One such survey, the CAHPS Clinician and Group Survey (CG-CAHPS), is used by general practitioners and specialists to assess patient experiences after office visits and outpatient procedures. CG-CAHPS collects information about patients' experiences with their provider and other medical and clinic staff to make comparisons across medical practices, giving providers a report card and patients the insights they need to make informed choices (Dyer et al. 2012;

Elliott et al. 2016; Holt 2019; Spranca et al. 2000). Data are commonly, though not exclusively, collected using mail surveys as a primary data collection mode, with telephone interviewers following up with nonrespondents (Elliott et al. 2009; Stucky et al. 2016). However, technological and methodological innovation may change how CG-CAHPS and other surveys of patients recruit and measure (Lee et al. 2013) in our increasingly online and cell-only society (Blumberg and Luke 2020; Pew Research Center 2019). Many medical practices request patients' email addresses and cell numbers to interact with them via medical portals and send them appointment reminders and confirmations (Garrido et al. 2016). Given the multiple types of contact information on patient lists, they provide rich sampling frames that offer multiple options for recruitment and measurement. Using this contact information to shift survey recruitment and data collection to Web and SMS may reduce costs and increase timeliness of response for CG-CAHPS surveys relative to mail and telephone data collection. Yet, without simplification and abbreviation of the questionnaire, respondent burden may yield relatively high rates of breakoffs, lower rates of completion, and other negative effects on data quality.

Therefore, we test the use of a planned missing design in a survey of patient experiences. Members of an online panel were invited by email or text message to complete a Web survey or an SMS survey. We assess and compare evidence of respondent burden across four versions of the questionnaire: three abbreviated questionnaires applying a planned missing design and a control group that includes the full survey instrument. We operationalize respondent burden in three ways: more burdensome modules will take longer to complete, yield more breakoffs, and, accordingly, lower completion rates. We also compare survey progress at interim stages of the recruitment process – click (on URL), screening, and eligibility rates – to assess differences in recruitment and survey progress between modes and modules given the negative effect of respondent burden (Galesic and Bosnjak 2009; Revilla and Ochoa 2017). We then assess the effect of respondent burden and the planned missing design on patient experience outcomes in the CG-CAHPS survey in both Web and SMS survey modes. A lack of differences in patient experience outcomes from the shortened questionnaires and the full-survey control condition would provide evidence for the effectiveness and appropriateness of using a planned missing design to reduce respondent burden with CG-CAHPS in these modes. To our knowledge, this is the first study using a planned missing design to examine the differential effects of respondent burden in Web and SMS modes.

3. Methods

3.1. Sample Design and Justification

Although the intended outcome of the planned missing design is higher quality data (Peytchev and Peytcheva 2017), removing questions may potentially cause unintended errors. Omitted questions may change the context in which subsequent questions are asked, introducing artifactual differences between questionnaire versions (Morgan and Poppe 2015; Swain 2015). These artifactual differences generated by methodological experiments could expose healthcare providers to negative outcomes by invalidating comparisons between providers and reducing the usefulness of CG-CAHPS data which are

used by government agencies and insurance companies in reimbursement and pay-for-performance computations and by consumers for their medical decision-making (Elliott et al. 2016; Holt 2019; Spranca et al. 2000). Thus, it is critical that if alternative methods of data collection are employed, such as planned missing designs in Web and SMS surveys, they not differentially affect CG-CAHPS estimates. Therefore, we test this design using a sample from a nonprobability online panel rather than nesting it within a production survey using a clinic's patient list as a sampling frame. Fifty-thousand members of Qualtrics' online nonprobability panel were invited to participate in a survey using their standard, generic invitation message. Panelists were offered an incentive given their pre-stated preferences (e.g., frequent flyer program miles, hotel brand loyalty points).

3.2. Invitation and Data Collection Modes

Combining the mode of invitation (email or SMS) with the mode of data collection (Web or SMS) resulted in three conditions: email invitation to a Web survey, SMS invitation to a Web survey, and SMS invitation to an SMS survey. Note that a fourth combination, an email invitation to an SMS survey, is possible. However, this design would require an additional step not required by the other designs. An invitation to a Web survey, either by email or by SMS, only requires clicking a link to open a browser. An SMS invitation to an SMS survey can simply be responded to. These approaches seem to us to be relatively seamless. An email invitation to an SMS survey would require a more active switch on the part of the respondent, moving from an email application to a texting application, entering a telephone number or shortcode, and sending a message to this new number. While shifting from email recruitment to SMS data collection is possible and has been used previously (Brenner and DeLamater 2013), we did not foresee it being a useful design for a CAHPS survey and did not test it in this study.

Email invitations were sent to 20,000 panelists that included a URL that forwarded to the survey. Another 20,000 panelists were invited by SMS to complete a Web survey and received a link identical to those sent via email. The remaining 10,000 panelists were invited by SMS to complete an SMS survey. These panelists received questions sequentially by SMS after agreeing to participate by responding to the initial message. The difference in sample sizes is partially due to the number of panelists available and partially due to cost. SMS surveys encounter extra costs in addition to the Web survey platform license, including an additional cost for SMS survey capabilities and per-message transmission costs for both outgoing questions and incoming answers. Data collection began in August 2019 and was completed in September 2019.

3.3. Planned Missing Experiment and Respondent Burden

In each mode, panelists were randomly assigned to one of four survey modules. Three of these modules used a planned missing design to reduce the length of the survey. Each of the first three survey modules included a subset of questions. Module A included questions assessing access to the provider; module B assessed communication with the provider; and module C assessed coordination of care. All three of these survey modules also included two questions about their provider's office staff, an overall provider rating, patient demographic questions, and self-rated health. These shortened modules included 14, 15, or

16 questions; shorter if the respondent answered "no" to one or more filter questions. The final module, module D, served as a control group and included all questions asked in each of the first three modules; a total of 25 questions, shorter if the respondent answered "no" to one or more filter questions.

The planned missing design is used to manipulate respondent burden as the shortened questionnaire modules are hypothesized to relieve respondent burden relative to the full-length questionnaire module. Following previous theory and research, respondent burden is operationalized as survey duration, and reflected in breakoff rates and, accordingly, completion rates (Antoun and Cernat 2020; Bradburn 1978; Mavletova and Couper 2015; Peytchev 2009; Sharp and Frankel 1983; Steinbrecher et al. 2015). Breakoff rates are computed as the percentage of eligible respondents who started the survey but did not complete it. Completion rates are computed as the percent of invited panelists who complete the survey.

3.4. Dependent Variables

Panelists were first screened for a doctor's visit in the past six months to be eligible for the survey. Subsequent questions refer to the provider identified in the screening question. Five dependent variables, created by combining question responses into composites, are measured and analyzed. The first is a composite combining three items, each measuring the patient's perception of the accessibility of their provider. The second combines four items, assessing how often the provider communicated effectively with the patient. The third combines three items, each assessing coordination of patient care. The fourth combines two items, measuring the patient's assessment of the courteousness of the medical office staff. The final dependent variable is a single item measure of the patient's assessment of his or her provider.

The four composites use scales demonstrated to have good internal consistency (Dyer et al. 2012). Each measured variable was recoded following common practice with CAHPS measures (never (0), sometimes (3.33), usually (6.67), and always (10)). For each scale, these values were then averaged and rounded to the nearest integer to create 11-point (0–10) composite measures (AHRQ n.d.). For scales with embedded filter questions, values for unasked follow-up items were missing if respondents responded "no" to the filter question. Scales values were computed using remaining answered scale items. This procedure puts these first four dependent variables on a scale similar to that of the fifth dependent variable, a 0 to 10 overall rating of the provider. See Table 1 for the full question text, response options, question order, and use across planned missing modules.

The survey was optimized for mobile devices, including only item specific scales and using no question grids, drop down menus, or sliders. Nearly all of the questions were successfully adapted for SMS, shortened to 160 characters or fewer, a US-based limit for text messages. Unfortunately, five questions could not be shortened enough without changing their meaning or excluding important information and, thus, exceeded this character limit. Note that most smartphones and carriers reassemble messages exceeding this limit before they appear on the device, and therefore we do not expect that this small subset of longer questions caused a problem for SMS respondents (Ayers et al. 2014).

Table 1. Question order, wordings, and response options for each CAHPS scale, by module.

Scale	Question	Response options	In modules			
Screening question	In the last six months, did you get medical care from a health care provider (physician, nurse practitioner, or physician's assistant)? Do not include dental care or overnight stays in a hospital.	YN	A	B	C	D
	Is this the provider you usually see if you need medical care? If you saw more than one provider, think about the one you saw the most.	YN	A	B	C	D
Provider accessibility	In the last six months, did you contact this provider's office to get an appointment for an illness, injury, or condition that needed care right away?	YN	A			D
Provider accessibility	In the last six months, when you contacted this provider's office to get an appointment for care you needed right away, how often did you get an appointment as soon as you needed?	NSUA	A			D
Provider accessibility	In the last six months, did you make any appointments for a check-up or routine care with this provider?	YN	A			D
Provider accessibility	In the last six months, when you made an appointment for a check-up or routine care with this provider, how often did you get an appointment as soon as you needed?	NSUA	A			D
Provider accessibility	In the last six months, did you contact this provider's office with a medical question during regular office hours?	YN	A			D
Provider accessibility	In the last six months, when you contacted this provider's office during regular office hours, how often did you get an answer to your medical question that same day?	NSUA	A			D
Effective communication	In the last six months, how often did this provider explain things in a way that was easy to understand?	NSUA		B		D
Effective communication	In the last six months, how often did this provider listen carefully to you?	NSUA		B		D
Care coordination	In the last six months, how often did this provider seem to know the important information about your medical history?	NSUA			C	D
Effective communication	In the last six months, how often did this provider show respect for what you had to say?	NSUA		B		D
Effective communication	In the last six months, how often did this provider spend enough time with you?	NSUA		B		D
Care coordination	In the last six months, did this provider order a blood test, x-ray, or other test for you?	YN			C	D

Table 1. Continued

Scale	Question	Response options	In modules			
Care coordination	In the last six months, when this provider ordered a blood test, x-ray, or other test for you, how often did someone from this provider's office follow up to give you those results?	NSUA			C	D
Provider rating	Using any number from 0 to 10, where 0 is the worst provider possible and 10 is the best provider possible, what number would you use to rate this provider?	0-10	A	B	C	D
Care coordination	In the last six months, did you take any prescription medicine?	YN			C	D
Care coordination	In the last six months, how often did you and someone from this provider's office talk about all the prescription medicines you were taking?	NSUA			C	D
Office staff rating	In the last six months, how often were clerks and receptionists at this provider's office as helpful as you thought they should be?	NSUA	A	B	C	D
Office staff rating	In the last six months, how often did clerks and receptionists at this provider's office treat you with courtesy and respect?	NSUA	A	B	C	D

Note: NSUA stands for "Never, Sometimes, Usually, Always"; YN stands for "Yes, No."

3.5. Control Variables

Self-rated health was included as a standard question that asked "In general, how would you rate your overall health?" using a five-point scale: excellent, very good, good, fair, or poor. Self-rated health is included in these analyses as it is predictive of patient experiences and healthcare utilization and is used as a patient-mix adjustment variable for comparing healthcare organizations (DeSalvo et al. 2004; Elliott et al. 2009; Paddison et al. 2013). Mental health was also measured using a question that asked: "In general, how would you rate your overall mental or emotional health?" and used the same response scale as the general health question. This measure of mental health is included as it is predictive of patient experiences and healthcare utilization (Ahmad et al. 2014). Both health variables have been recoded so that higher values reflect better health.

A series of demographic controls, including education (less than a high school degree, high school degree or GED, some college or two-year degree, four-year college degree, or more than a college degree), sex (male or female), and age (18–24, 25–34, 35–44, 45–54, 55–64, and 65 or older) are also included as categorical variables. Descriptive statistics for these independent variables and significance tests comparing across modes and modules are available in Table 2.

Table 2. Independent variables descriptive statistics and comparsion by mode and module

Mode	**Email/Web**					**SMS/Web**					**SMS/SMS**					**Mode diff.**
Module	Overall	A	B	C	D	Overall	A	B	C	D	Overall	A	B	C	D	
Self-rated health (%)																$X^2(8)=47.4$ ***
Excellent	14.0	14.1	14.3	13.3	14.4	12.1	13.0	11.4	12.8	11.1	7.8	8.3	5.9	8.1	8.9	
Very good	30.8	35.6	27.9	29.6	30.2	25.2	27.9	22.7	24.5	25.7	23.3	22.9	19.3	25.9	25.2	
Good	33.2	30.4	34.8	31.1	36.9	36.1	33.0	39.0	37.2	35.3	36.9	34.9	32.6	43.7	35.8	
Fair	18.3	17.3	18.8	21.6	15.1	21.3	20.7	20.3	21.3	25.7	26.9	28.4	35.6	17.8	26.0	
Poor	3.7	2.6	4.2	4.4	3.5	5.3	5.3	6.6	4.3	11.1	5.2	5.5	6.7	4.4	4.1	
Module diff.				$X^2(12)=15.9$					$X^2(12)=7.8$					$X^2(12)=14.1$		
Mental health (%)																$X^2(8)=53.2$ ***
Excellent	19.7	18.6	19.8	20.1	20.5	15.2	16.2	14.6	14.7	15.2	15.6	17.6	10.4	16.4	18.7	
Very good	27.1	29.4	28.9	23.9	26.5	23.1	24.7	23.0	21.6	22.9	20.2	21.3	19.3	26.1	13.8	
Good	27.8	26.8	29.1	28.6	26.7	28.3	28.2	30.0	27.7	27.4	27.2	30.6	28.9	23.9	26.0	
Fair	20.2	19.8	17.3	21.9	21.8	24.5	22.9	24.1	26.1	25.0	26.8	21.3	28.1	27.6	29.3	
Poor	5.1	5.4	4.9	5.6	4.5	9.0	8.0	8.5	9.9	9.6	10.2	9.3	13.3	6.0	12.2	
Module diff.				$X^2(12)=7.7$					$X^2(12)=3.6$					$X^2(12)=15.4$		
Sex (%)																$X^2(2)=127.9$ ***
Female	63.8	67.7	58.7	64.8	63.6	78.2	76.1	77.9	80.5	78.4	84.6	82.2	84.3	84.1	86.7	
Module diff.				$X^2(3)=7.6$					$X^2(3)=2.2$					$X^2(3)=1.0$		
Age (%)																$X^2(10)=69.2$ ***
18-24	11.2	11.3	10.4	11.6	11.4	11.6	8.2	10.6	16.2	11.2	14.0	8.3	12.6	10.5	21.7	
25-34	25.9	23.3	22.8	28.3	29.2	25.5	25.8	26.5	24.2	25.5	29.6	35.2	29.6	30.1	25.5	
35-44	25.1	27.1	24.3	24.7	24.3	26.4	27.1	29.4	24.2	25.0	27.2	30.6	28.1	24.1	26.7	
45-54	16.0	16.3	16.8	15.6	15.3	19.7	19.7	19.0	17.0	22.9	14.9	11.1	14.8	16.5	16.1	
55-64	11.5	13.7	13.4	9.4	9.7	12.8	15.2	10.0	14.1	12.0	10.2	9.3	13.3	9.8	8.7	
65 and older	10.3	8.3	12.4	10.5	10.2	4.1	4.0	4.5	4.3	3.5	4.1	5.6	1.5	9.0	1.2	
Module diff.				$X^2(15)=16.3$					$X^2(15)=22.2$					$X^2(15)=30.9$ **		
Education (%)																$X^2(8)=49.4$ ***
Less than HS	3.3	4.9	1.5	4.2	2.2	3.7	4.5	2.4	4.8	3.2	4.0	1.0	1.5	5.4	7.0	
HS grad./GED	23.4	21.2	23.8	23.6	25.0	22.8	22.9	19.0	23.9	25.3	21.6	21.9	20.1	16.9	26.6	
Some coll., 2yr deg.	35.9	36.0	33.7	36.1	37.9	44.0	41.1	47.6	47.2	40.3	47.8	45.7	58.2	43.8	43.7	
4-yr coll. degree	26.9	28.2	28.2	25.4	26.0	19.3	18.1	21.2	17.4	20.3	17.8	21.0	13.4	17.7	19.6	
More than 4yr deg.	10.5	9.6	12.9	10.7	8.9	10.2	13.3	9.8	6.7	10.9	8.7	10.5	6.7	16.2	3.2	
Module diff.				$X^2(12)=17.0$					$X^2(12)=21.6$ *					$X^2(12)=33.2$ ***		
Race and ethnicity (%)																$X^2(6)=10.7$
White	68.1	69.9	69.4	67.0	66.1	69.9	72.0	70.1	68.3	69.1	74.8	76.2	75.9	76.2	71.6	
Black	14.4	12.7	12.9	15.8	16.1	13.4	11.7	14.0	13.1	14.9	9.6	7.6	7.5	11.1	11.6	
Latino/a, any race	12.0	11.8	13.7	11.4	11.1	11.0	12.3	8.7	13.1	10.1	10.2	10.5	11.3	9.5	9.7	
Asian or other race	5.5	5.6	4.0	5.8	6.7	5.7	4.0	7.1	5.6	5.9	5.4	5.7	5.3	3.2	7.1	
Module diff.				$X^2(9)=7.7$					$X^2(9)=9.3$					$X^2(9)=4.5$		

Note: *p<.05; **p<.01; ***p<.001

3.6. Analysis

We first compare screening and eligibility rates by predicting a series of nested logistic regression models. The first models are estimated with dummy variables for respondents' randomly assigned invitation and data collection mode (email/Web, SMS/Web, SMS/SMS) as the sole predictor. Model fit is compared and assessed using likelihood ratio X^2 tests. Given significant improvement in fit, pairwise comparisons between categorical predictors are presented. The second models add dummy variables for planned missing design module (Modules A, B, C, and D) as a predictor. Given that screening for eligibility takes place before the respondent is exposed to their randomly assigned questionnaire module, we expect that module will have no effect on screening and eligibility rates. Finally, the interaction between mode and module is tested.

We similarly assess measures of respondent burden by estimating nested logistic regression models that predict the propensity of a case to be completed or to breakoff using mode and module as predictors. We then assess survey duration, the final measure of respondent burden, by estimating a median regression predicting survey duration using mode and module as predictors. Given our expectation of a large difference in completion times between Web and SMS surveys, as suggested in the existing research (Schober et al. 2015) they are analyzed separately.

We then examine the distributions of the demographic and self-reported health variables between modes and modules using chi-square tests. These demographic and health variables are used as covariates in the final set of models assessing the effect of the planned missing design on patient experience outcomes, the key measures of the CG-CAHPS survey. A set of nested ordinary least squares regression models are estimated and adjusted means are presented for each outcome: provider accessibility, communication with provider, coordination of care, assessment of the provider's office staff, and the overall provider assessment. The baseline model adjusts for demographic and health variables. We then estimate two additional sets of models. First, we add dummy variables for mode of invitation and data collection, comparing this model to the prior baseline model using an *F*-test to test the statistical significance of including mode as a predictor of the dependent variable. Where the *F*-test indicates by-mode differences, predicted values of the dependent variables are compared by mode. We then add dummy variables for the four questionnaire modules (A, B, C, and D), comparing this model to the prior model including mode. Where the *F*-test indicates by-module differences, predicted values for the dependent variable are compared by module. Finally, we test the interaction of mode and module by comparing a model with this interaction term to the prior model using an *F*-test.

4. Results

4.1. URL Click, Screening, Eligibility, Breakoff, and Completion Rates

About 14% of those receiving email invitations to a Web survey (2,788 panelists) started the survey by clicking on the URL. Of those panelists who clicked the link to the survey, 82% (2,272 panelists) completed the first question, a screener for a recent (in the past six months) doctor visit. Three-quarters (1,721 panelists) of those answering the screening question were eligible, and nearly all (97%) of these eligible cases completed the survey

(1,665 respondents), yielding a 3.3% breakoff rate and an 8.3% completion rate in the email/Web condition. Sample sizes, eligibility, screening, breakoff, and completion rates by survey mode and module are available in Table 3.

In the condition using an SMS invitation to a Web survey, 2,131 (11%) of those receiving a text invitation started the survey by clicking the link in their SMS messaging application. Of those panelists who clicked the link they received by SMS, 1,991 (93%) answered the screening question. Over 80% (1,635 panelists) of those answering the screening question were eligible, and 90% (1,473 respondents) of these eligible panelists completed the survey, yielding a 9.9% breakoff rate and a 7.4% completion rate in the SMS/Web condition.

Nearly 10% (977) of those invited by SMS to the SMS survey started the survey by replying to the screener question. Note that this condition lacks a URL to click. Thus, the survey begins with screening: an initial text message that asks a question to determine eligibility. Responding to this initial question functions both as the initial "URL click" of a sort and an eligibility screen. Of the panelists who answered the screening question, 787 (81%) reported an eligible doctor's visit. Of the eligible cases, 70% or 550 respondents, completed the survey yielding a 30% breakoff rate and a completion rate of 5.5% in the SMS/SMS condition.

We estimated logistic regression models predicting screening and eligibility using mode as an independent variable. Differences emerge in screening ($X^2 = 27.6$, $p < .001$) and eligibility rates ($X^2 = 27.6$, $p < .001$) between modes during recruitment (see Table 4). Given the significant X^2 statistics, we compared modes pairwise. When predicting screening, significant differences emerge between email/Web and both of the other modes, SMS/Web ($z = 4.55$; $p < .001$) and SMS/SMS ($z = 4.27$; $p < .001$). The small difference between the screening rate in email/Web (11%) and that in SMS/Web and SMS/SMS (10%) is likely significant given the very large sample size, all 50,000 panelists. Predicting eligibility, significant differences also emerge between email/Web (76%) and both of the other modes, SMS/Web (82%; $z = 5.12$; $p < .001$) and SMS/SMS (81%, $z = 3.09$; $p < .01$). We then added module to these models. For neither screening ($X^2 = 1.1$) nor eligibility ($X^2 = 5.0$) does adding module to these models improve fit. Finally, we tested the interaction of mode and module in these models. For neither model does adding the interaction improve model fit.

4.2. Modeling Respondent Burden: Completion Rates

Next, we estimate logistic regression models to examine differences in respondent burden between components of the study design. Survey completion is used as a first (negative) indictor of respondent burden (Bradburn 1978; Sharp and Frankel 1983). First, we predict survey completion using mode as the sole predictor, then add module as a predictor, and finally their interaction. The first model suggests that completion rates differ significantly between the three modes ($X^2 = 77.9$, $p < .001$). Pairwise comparisons show significant differences between these three rates. Completion is highest in email/Web (8.3%) compared with SMS/Web ($z = 3.5$, $p < .001$) and SMS/SMS ($z = 9.4$, $p < .001$) and lowest in SMS/SMS (5.5%), compared with SMS/Web (7.4%; $z = 6.4$, $p < .001$) which has a completion rate that falls in between the other two modes.

Table 3. Samples sizes, screening, eligibility, breakoff, and completion rates, by condition

Mode		All Modules		Module A		Module B		Module C		Module D	
		#	rate	#	rate	#	rate	#	rate	#	rate
Email/Web	# Invited	20000		5000		5000		5000		5000	
	# Clicked URL (% of invited)	2788	13.9%	733	14.7%	622	12.4%	805	16.1%	628	12.6%
	# Screened (% of invited)	2272	11.4%	594	11.9%	548	11.0%	583	11.7%	547	10.9%
	(% of URL clicked)		81.5%		81.0%		88.1%		72.4%		87.1%
	# Eligible (% of screened)	1721	75.7%	436	73.4%	413	75.4%	455	78.0%	417	76.2%
	# Breakoff (% of eligible)	56	3.3%	19	4.4%	9	2.2%	6	1.3%	22	5.3%
	# Completed / (% of invited)	1665	8.3%	417	8.3%	404	8.1%	449	9.0%	395	7.9%
SMS/Web	# Invited	20000		5000		5000		5000		5000	
	# Clicked URL (% of invited)	2131	10.7%	551	11.0%	570	11.4%	488	9.8%	522	10.4%
	# Screened (% of invited)	1991	10.0%	510	10.2%	494	9.9%	478	9.6%	509	10.2%
	(% of URL clicked)		93.4%		92.6%		86.7%		98.0%		97.5%
	# Eligible (% of screened)	1635	82.1%	411	80.6%	413	83.6%	402	84.1%	409	80.4%
	# Breakoff (% of eligible)	162	9.9%	50	12.2%	32	7.7%	27	6.7%	53	13.0%
	# Completed / (% of invited)	1473	7.4%	361	7.2%	381	7.6%	375	7.5%	356	7.1%
SMS/SMS	# Invited	10000		2500		2500		2500		2500	
	# Screened (% of invited)	977	9.8%	224	9.0%	242	9.7%	242	9.7%	269	10.8%
	# Eligible (% of screened)	787	80.6%	186	83.0%	194	80.2%	197	81.4%	210	78.1%
	# Breakoff (% of eligible)	237	30.1%	63	33.9%	50	25.8%	45	22.8%	79	37.6%
	# Completed / (% of invited)	550	5.5%	123	4.9%	144	5.8%	152	6.1%	131	5.2%

Table 4. Predicted recruitment (screening and eligibility rates) and respondent burden (completion and breakoff rates and survey duration), modes and modules

		Overall		**Adding mode**			**Adding module**								**Testing interaction**
							Module A		Module B		Module C		Module D		
		%/*M*	SE	%/*M*	SE	LR X^2 / *F* *z*	%/*M*	SE	%/*M*	SE	%/*M*	SE	%/*M*	SE	LR X^2 / *F* *z*	LR X^2 / *F*
Recruitment																
Screened	Overall	0.10	0.00			$X^2(2)= 27.57^{***}$									$X^2(3)= 1.09$	$X^2(6)=8.55$
	Email/Web (1)			0.11	0.00	$z_{1\text{-}2}=4.55^{***}$	0.12	0.00	0.11	0.00	0.11	0.00	0.11	0.00		
	SMS/Web (2)			0.10	0.00	$z_{1\text{-}3}=4.27^{***}$	0.10	0.00	0.10	0.00	0.10	0.00	0.10	0.00		
	SMS/SMS (3)			0.10	0.00	$z_{2\text{-}3}=0.51$	0.10	0.00	0.10	0.00	0.10	0.00	0.10	0.00		
Eligible	Overall	0.79	0.01			$X^2(2)= 27.55^{***}$									$X^2(3)= 4.99$	$X^2(6)=4.58$
	Email/Web (1)			0.76	0.01	$z_{1\text{-}2}=5.12^{***}$	0.74	0.01	0.76	0.01	0.78	0.01	0.75	0.01		
	SMS/Web (2)			0.82	0.01	$z_{1\text{-}3}=3.09^{**}$	0.81	0.01	0.82	0.01	0.84	0.01	0.81	0.01		
	SMS/SMS (3)			0.81	0.01	$z_{2\text{-}3}=1.02$	0.79	0.02	0.81	0.02	0.82	0.01	0.80	0.02		
Respondent burden																
Complete	Overall	0.07	0.00			$X^2(2)= 77.88^{***}$									$X^2(3)= 5.63$	$X^2(6)=3.55$
	Email/Web (1)			0.08	0.00	$z_{1\text{-}2}=3.50^{***}$	0.08	0.00	0.08	0.00	0.09	0.00	0.08	0.00		
	SMS/Web (2)			0.07	0.00	$z_{1\text{-}3}=9.35^{***}$	0.07	0.00	0.07	0.00	0.08	0.00	0.07	0.00		
	SMS/SMS (3)			0.06	0.00	$z_{2\text{-}3}=6.36^{***}$	0.05	0.00	0.06	0.00	0.06	0.00	0.05	0.00		
Breakoff	Overall	0.11	0.00			$X^2(2)=344.80^{***}$									$X^2(3)=37.22^{***}$	$X^2(6)=1.34$
	Email/Web (1)			0.03	0.00	$z_{1\text{-}2}=7.46^{***}$	0.04	0.01	0.03	0.00	0.02	0.00	0.05	0.01	$z_{A\text{-}B}=3.19^{***}$	
	SMS/Web (2)			0.10	0.01	$z_{1\text{-}3}=15.72^{***}$	0.12	0.01	0.08	0.01	0.07	0.01	0.13	0.01	$z_{A\text{-}C}=4.26^{***}$	
	SMS/SMS (3)			0.30	0.02	$z_{2\text{-}3}=11.26^{***}$	0.35	0.03	0.25	0.02	0.22	0.02	0.38	0.03	$z_{B\text{-}D}=4.13^{***}$	
															$z_{C\text{-}D}=5.22^{***}$	
Survey duration	Overall (Web)	137	1.54			$F_{(1,3177)}=0.11$	141	3.03	116	3.07	113	2.99	186	3.07	$F_{(3,3175)}=121.0^{***}$	$F_{(3,3171)}=0.88$
	Email/Web (1)			138	2.07										$z_{A\text{-}B}=5.72^{***}$	
	SMS/Web (2)			137	2.19										$z_{A\text{-}C}=6.73^{***}$	
															$z_{A\text{-}D}=10.30^{***}$	
															$z_{B\text{-}D}=15.93^{***}$	
															$z_{C\text{-}D}=17.05^{***}$	
	SMS/SMS (3)	1845	191				1835	430	1975	366	1684	391	1884	390	$F_{(3,482)}=0.2$	

Note: *M* is used to represent the median. $^{***}p<.001$; $^{**}p<.01$

In the second model, we add planned missing modules as a predictor. Accounting for module does not significantly improve the fit of the model ($X^2 = 5.6$). A subsequent test including the interaction of mode and module also unsurprisingly fails to significantly improve the fit of the model ($X^2 = 3.55$).

4.2.1. Modeling Respondent Burden: Breakoff Rates

Survey breakoff is used as a second indictor of respondent burden (Peytchev 2009; Steinbrecher et al. 2015). We estimate logistic regression models to further examine differences in breakoff rates between components of the study design, first including only mode as a predictor, then adding module, and finally testing their interaction. The first model suggests that mode is a significant predictor of breakoffs ($X^2 = 344$, $p < .001$). Given the significant X^2 statistic, we compared modes pairwise, finding significant differences between all three modes, with email/Web (3.3%) prompting a lower breakoff rate than SMS/Web (9.9%; $z = 7.5$; $p < .001$) and SMS/SMS (30%) resulting in a higher breakoff rate than email/Web ($z = 15.7$; $p < .001$) and SMS/Web ($z = 11.3$; $p < .001$).

In the second model, we add planned missing modules as a predictor. Adding module significantly improves the fit of the model ($X^2 = 37.2$; $p < .001$). Given the significant X^2 statistic, we compared modules pairwise. Full length module D increases breakoffs compared to modules B ($z = 4.1$; $p < .001$) and C ($z = 5.2$; $p < .001$). The longest of the shortened modules, module A, also shows a higher breakoff rate than modules B ($z = 3.2$; $p < .001$) and C ($z = 4.3$; $p < .001$). Notably, modules A and D and modules B and C do not differ in their breakoff rates. Finally, the interaction of mode and module was tested but was not statistically significant ($X^2 = 1.3$).

4.2.2. Modeling Respondent Burden: Survey Duration

Survey duration is used here as the final measure of respondent burden (Antoun and Cernat 2020; Mavletova and Couper 2015). We examined differences in survey duration between modes and modules by estimating quantile (median) regression models. These models predicted survey duration in seconds, first introducing mode as a predictor, then adding module to the model. Models are estimated separately for Web (email/Web and SMS/Web) and SMS surveys given our expectation of a large difference in survey duration between data collection modes (Schober et al. 2015).

Median duration did not differ between the two Web surveys ($F_{1,3177} = 0.1$): 138 seconds in email/Web and 137 seconds in SMS/Web. As no significant difference emerged between these two modes, we pool the Email/Web and SMS/Web data to test differences across modules. Median survey duration differed between the four modules ($F_{3,3175} = 121.0$, $p < .001$). Given the significant F statistic, we compared modules pairwise. Median duration for the pooled web-based surveys differs between each pair of modules except for modules B and C for which duration fails to differ. Time to complete Module D exceeds all other modules: A ($z = 10.3$, $p < .001$), B ($z = 15.9$, $p < .001$) and C ($z = 17.1$, $p < .001$) and time to complete module A exceeds modules B ($z = 5.7$, $p < .001$) and C ($z = 6.7$, $p < .001$). Finally, the interaction of mode and module was tested but was not statistically significant ($F_{1,3171} = 0.9$).

In the SMS/SMS survey, median duration was over 30 minutes (1,845 seconds), ranging from 1,668 seconds for Module C to 1975 for Module B. However, survey duration did not significantly differ between modules in the SMS/SMS condition ($F_{3,482} = 0.2$).

4.3. Demographic and Self-Reported Health Covariates

A number of significant differences emerge in the distributions of the four demographic variables and two self-reported health variables between the three modes (see Table 2). Respondents answering the survey in the SMS/SMS condition report relatively poorer health and mental health. SMS respondents are also more likely female and are relatively younger than their counterparts completing Web surveys. Those invited by email to complete a Web survey, however, are relatively more educated than those invited by SMS to take either a Web or SMS survey and report better mental health. Relatively few differences emerge across modules within modes. Differences emerge in the SMS/Web survey on education, and in the SMS/SMS survey on age and education. However, no clear and consistent patterns emerge from these comparisons across modules.

4.4. Predicting Patient Experience Outcomes

The first set of regression models provide a baseline restricted model predicting the five outcomes controlling only for demographic and health covariates. The adjusted means from these models for the five outcomes range from 7.0 for provider accessibility to 8.1 for effective communication and the overall provider rating (see Table 5).

The next set of regression models include mode as an independent variable. *F*-tests comparing each model to the corresponding previously estimated baseline model show statistically significant effects for mode in models for three of the five outcomes, effective communication ($F_{2,1794} = 10.5$, $p < .001$), office staff rating ($F_{(2,3640)} = 13.2$, $p < .001$), and provider rating ($F_{(2,3635)} = 7.8$, $p < .001$), and was marginally significant in a fourth model for provider accessibility ($F_{(2,1712)} = 2.6$, $p < .10$).

For each model with a significant *F*-statistic, we compare adjusted means for these outcomes by mode. Email/Web responses differ from SMS/SMS for four outcomes, provider accessibility (6.8 and 7.3; $t = 2.14$, $p < .05$), effective communication (7.8 and 8.5; $t = 3.96$, $p < .001$), office staff rating (7.7 and 8.3; $t = 5.03$, $p < .001$), and provider rating (8.0 and 8.4; $t = 3.50$, $p < .001$), in each case the SMS survey yielding the higher value. Email/Web responses differ from SMS/Web for three outcomes, effective communication (7.8 and 8.3; $t = 3.64$, $p < .001$), office staff rating (7.7 and 7.9; $t = 2.79$, $p < .01$), and provider rating (8.0 and 8.2; $t = 2.81$, $p < .01$), in each case the SMS/Web condition yielding the higher value. For only one outcome, the office staff rating, do SMS/Web (7.9) and SMS/SMS (8.3) responses differ ($t = 3.12$, $p < .01$).

The next set of models include module as an independent variable. *F*-tests comparing each model to the corresponding previously estimated model show statistically significant effects for module for just one of the five outcomes, care coordination ($F_{2,1794} = 10.5$, $p < .001$), with module C yielding lower values of the outcome than module D.

The final set of models include the interaction of mode and module. *F*-tests comparing each model to the corresponding previously estimated main-effect only models show no statistically significant effects for the interaction terms.

Table 5. Adjusted means for outcomes (CAHPS composites), by module and mode types

		Demographic controls			Adding mode				Adding module									Adding interaction
									Module A		Module B		Module C		Module D			
Outcomes		mean	SE	N	mean	SE	N	F (mode) t (mode diff.)	mean	SE	mean	SE	mean	SE	mean	SE	F (module)	F(mode*module)
Provider accessibility	Overall	6.96	0.06	1730				$F_{(2,1712)}$=2.6†	6.95	0.09					6.98	0.09	$F_{(1,1711)}$=0.1	$F_{(2,1709)}$=1.2
	Email/Web (1)				6.83	0.10	806	$t_{1\text{-}3}$=2.14*	6.81	0.11					6.84	0.12		
	SMS/Web (2)				7.03	0.10	710		7.01	0.12					7.04	0.12		
	SMS/SMS (3)				7.28	0.19	214		7.26	0.20					7.29	0.20		
Effective communication	Overall	8.09	0.05	1812				$F_{(2,1794)}$=10.5***			8.04	0.07			8.14	0.07	$F_{(1,1793)}$=0.9	$F_{(2,1791)}$=0.4
	Email/Web (1)				7.83	0.08	806	$t_{1\text{-}3}$=3.96***			7.78	0.09			7.88	0.09		
	SMS/Web (2)				8.25	0.08	751	$t_{1\text{-}2}$=3.64***			8.20	0.10			8.29	0.10		
	SMS/SMS (3)				8.48	0.14	255				8.43	0.15			8.52	0.15		
Care coordination	Overall	7.30	0.06	1846				$F_{(2,1828)}$=0.7					7.12	0.08	7.49	0.09	$F_{(1,1827)}$=9.2**	$F_{(2,1825)}$=2.0
	Email/Web (1)				7.24	0.09	851						7.06	0.11	7.42	0.11		
	SMS/Web (2)				7.33	0.09	747						7.16	0.11	7.52	0.11		
	SMS/SMS (3)				7.44	0.16	248						7.26	0.18	7.62	0.18		
Provider rating	Overall	8.12	0.03	3653				$F_{(2,3635)}$=7.8***	8.07	0.06	8.21	0.06	8.05	0.06	8.17	0.06	$F_{(3,3632)}$=1.4	$F_{(6,3626)}$=1.0
	Email/Web (1)				8.00	0.05	1673	$t_{1\text{-}3}$=3.50***	7.95	0.07	8.08	0.07	7.93	0.07	8.05	0.07		
	SMS/Web (2)				8.20	0.05	1494	$t_{1\text{-}2}$=2.81**	8.14	0.07	8.28	0.07	8.12	0.07	8.24	0.07		
	SMS/SMS (3)				8.35	0.09	486		8.30	0.10	8.43	0.10	8.28	0.10	8.40	0.10		
Office staff rating	Overall	7.88	0.04	3658				$F_{(2,3640)}$=13.2***	7.87	0.08	7.86	0.08	7.93	0.08	7.88	0.08	$F_{(3,3637)}$=0.2	$F_{(6,3631)}$=1.1
	Email/Web (1)				7.70	0.06	1675	$t_{1\text{-}3}$=5.03***	7.69	0.09	7.68	0.09	7.75	0.09	7.70	0.09		
	SMS/Web (2)				7.94	0.06	1497	$t_{1\text{-}2}$=2.79**	7.93	0.09	7.92	0.09	7.99	0.09	7.93	0.09		
	SMS/SMS (3)				8.33	0.11	486	$t_{2\text{-}3}$=3.12**	8.31	0.13	8.30	0.13	8.37	0.13	8.32	0.13		

Note: ***p<.001; **p<.01; *p<.05; †p<.10;

5. Discussion

5.1. Respondent Burden and Planned Missing Modules

That shorter surveys reduce respondent burden (Johnson et al. 2013; Toepoel and Lugtig 2018; West et al. 2015) motivated our novel test of a planned missing design in Web and SMS patient surveys. We operationalized respondent burden as (longer) survey duration, (lower) completion rates, and (higher) breakoff rates in line with prior theory and research (Antoun and Cernat 2020; Bradburn 1978; Mavletova and Couper 2015; Peytchev 2009; Sharp and Frankel 1983; Steinbrecher et al. 2015). Taken together, findings suggest that the planned missing design did reduce respondent burden, especially in both Web (email/Web and SMS/Web) surveys. Module D, the full-length version of the web survey questionnaire, took significantly longer for Web respondents to answer compared with the three shortened questionnaire modules. Of the planned missing modules, Module A asked more questions than did the others (Modules B and C) and took respondents longer to answer than the other shortened versions. In the SMS/SMS mode, however, no significant differences in survey duration emerged between modules. This is likely due to two important factors: (1) reduced statistical power given the smaller number of completed SMS surveys, and (2) the relative difference between the very brief amount of time needed to answer questions and the longer spans of time between active question answering. Finally, although completion rates did not significantly vary between modules, Modules B and C did yield significantly lower breakoff rates compared to Modules A and D.

5.2. Respondent Burden and Invitation and Data Collection Modes

The three invitation and data collection modes clearly varied by respondent burden as operationalized in the current study. SMS as a data collection mode appeared to increase the burden placed on respondents relative to the two Web surveys. Panelists assigned to SMS were less likely to complete the survey and those who do start an SMS survey are more likely to breakoff, at a rate in line with prior research (Lau et al. 2019). Moreover, those SMS panelists who persist through the survey spend an order of magnitude more time completing it compared to the two groups of Web survey respondents regardless of how they were invited (i.e., via email or SMS). In comparison with Web surveys, SMS surveys may also present respondent burden that differs in kind as well as degree. In the current study, Web surveys are completed by most respondents in a single session, answering the first question and the last within a few minutes of each other. Conversely, SMS respondents stretch out the question-answering process, distributing the burden of responding over a longer period of time (Schober et al. 2015; West et al. 2005) and increasing the potential for breakoffs (Lee et al. 2013).

Respondent burden in the SMS/Web condition fell in between the SMS survey with the SMS invitation and the other Web survey using an email invitation. This finding follows extant research demonstrating that SMS invitations can lead to higher respondent burden, including higher breakoffs and reduced completion rates (Mavletova and Couper 2014).

5.3. Effect of Respondent Burden on Outcomes

Although the planned missing design appears to reduce respondent burden, few differences emerge in patient experience outcomes. This finding suggests that partitioning

the survey instrument as part of the planned missing design did not change how respondents answered questions. For only one outcome, coordination of care, did including module in the model significantly improve model fit. Respondents in full-length module D report better experiences with care coordination than respondents in shortened module C, the only shortened module with this question included. This single difference between modules is likely a context effect (Swain 2015). In module D, the full survey, care coordination questions were interleaved with questions about the effectiveness of the provider's communication and the overall provider rating (see Table 1). Questions about the provider explaining things well, listening carefully, showing respect, and spending adequate time may encourage module D respondents to feel more positively about their provider. Thus, respondents answering those questions may give higher ratings of care coordination (knowing the patient's medical history, providing test results) than module C respondents for whom care coordination questions are the first set answered after the initial screening items. In sum, while the longer survey modules arguably increased respondent burden, illustrated in longer survey duration and higher breakoff rates, respondent burden seemingly did not alter measured outcomes. This finding provides some initial evidence supporting the use of the planned missing design with CG-CAHPS.

5.4. *Effect of Invitation and Data Collection Mode on Outcomes*

Invitation and data collection modes also influenced how respondents answered in four of the five outcomes. Email/Web respondents reported poorer experiences with their provider than did respondents invited by SMS. For only one outcome did SMS/Web respondents differ from SMS/SMS respondents, the former reporting poorer experiences with their provider's office staff than the latter. While the source of these by-mode differences in outcomes is uncertain, the pattern of findings suggests a potential cause. The consistent differences in four of five outcomes between email/Web and SMS/SMS, the consistent lack of differences in four of five outcomes between SMS/Web and SMS/SMS, and that outcomes from SMS/Web consistently fell in between those from the other two mode protocols, suggest a mode of invitation effect generating these differences. Panelists who respond to an invitation by SMS, or are even willing to receive an invitation by SMS, clearly differ from panelists who receive and respond to email invitations.

5.5. *Limitations and Future Directions*

In summary, the planned missing design was successful for the Web surveys, reducing respondent burden – shorter survey duration and fewer breakoffs – without changing patient experience estimates. While we are optimistic about the use of a planned missing design with CG-CAHPS surveys in the future, further study is needed, specifically with a sample of patients from a provider or healthcare organization. We used a sample from an online non-probability panel as a first test. It is possible that online panelists may differ from a sample of patients from a provider list in key ways that could alter response patterns and patient experience estimates. Given their greater experience with answering surveys, panelists may be more adept at survey response than the average patient (Toepoel et al. 2008). Thus, the applicability of these findings may be limited if panelists answer these questions in ways that differ systematically from a sample from a clinic, hospital, or

physician patient list; a concern that is not unique to this study (Hillygus et al. 2014). However, given that we screened panelists for a physician visit in the last six months, and this is not a rare population, these findings are arguably likely to be similar to what would be found using a patient sample. Future research should validate these findings with a sample generated from a list of patients with a recent visit to a healthcare provider.

The planned missing design yielded relatively more mixed results in the SMS survey, reducing breakoff rates but failing to shorten survey duration. This finding suggests that respondent burden may be linked to qualities of the planned missing design other than time to completion, such as the sheer number of questions asked. Respondents may fear that the SMS survey in particular is never-ending as every answer sent leads to a new question received Indeed, SMS achieved the lowest completion rate and highest breakoff rate in comparison with both Web survey modes. While SMS also yielded far and away the longest survey duration, our measure of duration lacks detail, an important limitation to be addressed in future research. Ideally, researchers would record the time that a question is sent, the message is opened, and the answer is received. With these additional details, future research may better understand how respondents interact with SMS surveys and whether and how the number of questions influences respondent burden.

Brief indicators of progress (e.g., "question 10 of 15" or "66% complete") sent by text in combination with or in addition to the survey questions may help motivate SMS survey respondents to continue with the survey and resist the desire to breakoff. However, these progress indictors may counterproductively extend the length of the text message. Moreover, progress indicators may reduce rather than increase response rates in Web surveys (Villar et al. 2013). Therefore, future research should carefully evaluate their use in SMS surveys.

6. Conclusion

This article contributes to the literature on respondent burden by testing a planned missing design – randomly assigning respondents to answer only a subset of questions to shorten the survey and reduce respondent burden (Johnson et al. 2013) – in Web and SMS surveys. To our knowledge, this is the first test of this approach to reduce respondent burden in an SMS survey.

Three shortened versions of the CG-CAHPS patient experience survey applying a planned missing design reduced the number of questions by about 40% in comparison to the full survey module. We invited members of an online panel using one of two invitation modes, email or SMS, to complete a survey in one of two data collection modes, Web or SMS surveys. We found that a planned missing design reduced respondent burden for Web survey respondents, yielding fewer breakoffs and shorter survey duration. Most importantly, we reduced respondent burden without significantly altering the way that respondents answered most of the questions about patient experience. While findings should be investigated further using samples from clinic patient lists, the findings from the Web survey are encouraging: a planned missing design can make our respondents' jobs easier while keeping survey statistics consistent.

The findings for the SMS survey were less clear. Although the shortened versions of the SMS surveys tended to reduce the breakoff rate relative to the full-length control version, they did not yield a reduction in the time it took respondents to complete the survey.

Arguably, both the full 25-item CG-CAHPS survey and the shortened 15-item surveys may simply be too long to expect respondents to complete by text message. Moreover, the piecemeal way questions are sent after previous questions are answered may prevent the planned missing design from perceptibly reducing the burden of the task. Thus, the nature of the mode combined with the number of questions may simply exhaust even those respondents initially willing to complete the survey, as reflected in a lower completion rate among eligible respondents and a breakoff rate that is 3-10 times higher than that in the Web surveys. The solution may be to further reduce the number of questions asked in an SMS survey, perhaps to around five or so, and make more explicit ways for respondents skip questions they do not wish to answer (Lau et al. 2019). Moreover, sample members who are capable of and willing to complete an SMS survey may be very different from those without the ability and willingness to respond via SMS. While this may change over time, this hurdle to completing SMS surveys with general populations may present a significant barrier to moving surveys into this mode for data collection.

We also found evidence for a mode of invitation effect. Respondents able and willing to even receive an SMS invitation may be substantially different from those without the capability and willingness. Those respondents invited by SMS to participate in Web survey responded to the survey in a way that split the difference between those who were invited by email to complete a Web survey and those invited by SMS to an SMS survey. Future research should further investigate how using SMS as an invitation mode may alter who responds.

In conclusion, the planned missing design can clearly be applied successfully in Web surveys. Our original full-length survey was already relatively brief, about 25 questions. Thus, other survey researchers may find additional reductions in response burden yielding even larger increases in completion or response rates and reductions in breakoffs and survey duration. However, work remains to determine if and how a planned missing design can be successfully adapted to reduce respondent burden in SMS surveys. The next step in applying a planned missing design in SMS may be to further reduce respondent burden by further shortening the survey, or in other words, planning for more missing.

7. References

Ahmad, F., A.K. Jhajj, D.E. Stewart, M. Burghardt, and A.S. Bierman. 2014. "Single Item Measures of Self-Rated Mental Health: A Scoping Review." *BMC Health Services Research* 14: 398. DOI: https://doi.org/10.1186/1472-6963-14-398.

AHRQ, Agency for Healthcare Research and Quality. n.d. "How to Report Results of the CAHPS Clinician & Group Survey." Available at: https://www.ahrq.gov/sites/default/-files/wysiwyg/cahps/surveys-guidance/cg/cgkit/HowtoReportResultsofCGC-AHPS080610FINAL.pdf (accessed September 2021).

Alemi, F., and H.M. Jasper. 2014. "An Alternative to Satisfaction Surveys: Let the Patients Talk." *Quality Management in Health Care* 23(1): 10–19. DOI: https://doi.org/10.1097/QMH.0000000000000014.

Antoun, C., and A. Cernat. 2020. "Factors Affecting Completion Times: A Comparative Analysis of Smartphone and PC Web Surveys." *Social Science Computer Review* 38(4): 477–489. DOI: https://doi.org/10.1177/0894439318823703.

Ayers, R., S. Brothers, and W. Jansen. 2014. "Guidelines on Mobile Device Forensics." National Institute of Standards and Technology, Special Publication 800-101r1. Washington, DC: US Department of Commerce. DOI: http://dx.doi.org/10.6028/NIST.Sp. 800-101r1.

Blumberg, S.J., and J.V. Luke. 2020. "Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2019." National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/nhis.htm. (accessed September 2021).

Bosnjak, M., W. Neibarth, M.P. Couper, W. Bandilla, and L. Kaczmirek. 2008. "Prenotification in Web-Based Access Panel Surveys." *Social Science Computer Review* 26(2): 213–223. DOI: https://doi.org/10.1177/0894439307305895.

Bradburn, N. 1978. "Respondent Burden." In Proceedings of the Survey Research Methods Section of the American Statistical Association: 35–40. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y1978f.html (accessed September 2021).

Brenner, P.S., and J.D. DeLamater. 2013. "Paradata Correlates of Data Quality in an SMS Time Use Study–Evidence from a Validation Study." *Electronic International Journal of Time Use Research* 10(1): 38–54. DOI: dx.doi.org/10.13085/eIJTUR.10.1.38-54.

Conrad, FG., M.P. Couper, R. Tourangeau, and C. Zhang. 2017. "Reducing Speeding in Web Surveys by Providing Immediate Feedback." *Survey Research Methods* 11(1): 45–61. DOI: https://doi.org/10.18148/srm/2017.v11i1.6304.

Crawford, S.D., M.P. Couper, and M.J. Lamias. 2001. "Web Surveys: Perception of Burden." *Social Science Computer Review* 19(2): 146–162. DOI: https://doi.org/10.1177/089443930101900202.

De Leeuw, E.D. 2018. "Mixed-Mode: Past, Present, and Future." *Survey Research Methods* 12(2): 75–89. DOI: doi:10.18148/srm/2018.v12i2.7402.

DeSalvo, K.B., V.S. Fan, M.B. McDonell, and S.D. Fihn. 2004. "Predicting Mortality and Healthcare Utilization with a Single Question." *Health Services Research* 40(4): 1234–1246. DOI: https://doi.org/10.1111/j.1475-6773.2005.00404.x.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* New York: Wiley.

Dyer, N., J.S. Sorra, S.A. Smith, P. Cleary, and R. Hays. 2012. "Psychometric Properties of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) Clinician and Group Adult Visit Survey." *Medical Care* 50: S28. DOI: https://doi.org/10.1097/MLR.0b013e31826cbc0d.

Elliott M.N., M.K. Beckett, W.G. Lehrman, P. Cleary, C.W. Cohea, L.A. Giordano, E.H. Goldstein, and C.L. Damberg. 2016. "Understanding The Role Played By Medicare's Patient Experience Points System In Hospital Reimbursement." *Health Affairs* 35(9): 1673–1680. DOI: https://doi.org/10.1377/hlthaff.2015.0691.

Elliott, M.N., A.M. Zaslavsky, E. Goldstein, W. Lehrman, K. Hambarsoomians, M.K. Beckett, and L. Giordano. 2009. Effects of survey mode, patient mix, and nonresponse on CAHPS® hospital survey scores. *Health Services Research*, 44 (2p1): 501–518. DOI: https://doi.org/10.1111/j.1475-6773.2008.00914.x.

Enders, C.K. 2010. *Applied Missing Data Analysis*. New York: Guilford.

Galesic, M., and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73(2): 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Garrido, T., B. Raymond, and B. Wheatley. 2016. "Lessons From More than a Decade In patient Portals." *Health Affairs Blog*. DOI: https://doi.org/10.1377/hblog20160407.054362.

Graham, J.W., B.J. Taylor, A.E. Olchowski, and P.E. Cumsille. 2006. "Planned Missing Data Designs in Psychological Research." *Psychological Methods* 11(4): 323–343. DOI: https://doi.org/10.1037/1082-989X.11.4.323.

Hillygus, D.S., N. Jackson, and M. Young. 2014. "Professional respondents in non-probability online panels." In *Online panel research: A data quality perspective,* edited by M. Callegaro, R. Baker, J. Bethlehem, A.S. Góritz, J.A. Krosnick and P.J. Lavrakas: 219–237. New York: Wiley.

Holt, J.M. 2019. "Patient Experience in Primary Care: A Systematic Review of CG-CAHPS Surveys." *Journal of Patient Experience* 6(2): 93–102. DOI: https://doi.org/10.1177/2374373518793143.

Johnson, D.R., V. Roth, and R. Young. 2013. "Planned Missing Data Designs in Health Surveys." In Proceedings of the Tenth Conference on Health Survey Research Methods, November 2013: 43–53. Hyattsville, MD: Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/data/hsrmc/hsrmc_10th_proceedings_2011.pdf (accessed February 2020).

Keller, S., A.J. O'Malley, R.D. Hays, R.A. Matthew, A.M. Zaslavsky, K.A. Hepner, and P.D. Cleary. 2005. "Methods Used to Streamline the CAHPS® Hospital Survey." *Health Services Research*, 40: 2057–2077. DOI: https://doi.org/10.1111/j.1475-6773.2005.00478.x.

Kim, Y., J. Dykema, J. Stevenson, P. Black, and D.P. Moberg. 2019. "Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys." *Social Science Computer Review* 37(2): 214–233. DOI: https://doi.org/10.1177/0894439317752406.

Lau, C.Q., H. Sanders, and A. Lombaard. 2019. "Questionnaire Design in Short Message Service (SMS) Surveys." *Field Methods* 31(3): 214–229. DOI: https://doi.org/10.1177/1525822X19862506.

Lee, S.S.S., X. Xin, W.P. Lee, E.J. Sim, B. Tan, M.P.G. Bien, A.S.T. Lau, and J. Thumboo. 2013. "The Feasibility of Using SMS as a Health Survey Tool: An Exploratory Study in Patients with Rheumatoid Arthritis." *International Journal of Medical Informatics* 82(5): 427–434. DOI: https://doi.org/10.1016/j.ijmedinf.2012.12.003.

Mavletova, A. and M.P. Couper. 2014. "Mobile Web Survey Design: Scrolling versus Paging, SMS versus E-mail Invitations." *Journal of Survey Statistics and Methodology* 2: 498–518. DOI: https://doi.org/10.1093/jssam/smu015.

Mavletova, A., and M.P. Couper. 2015. "A Meta-Analysis of Breakoff Rates in Mobile Web Surveys." In *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, edited by D. Toninelli, R. Pinter, and P. de Pedraza: 81–98. London: Ubiquity Press.

Mavletova, A., M.P. Couper, and D. Lebedev. 2018. "Grid and Item-by-Item Formats in PC and Mobile Web Surveys." *Social Science Computer Review* 36(6):647-668. DOI: https://doi.org/10.1177/0894439317735307.

Morgan, S.L., and E.S. Taylor Poppe. 2015. "A Design and a Model For Investigating the Heterogeneity of Context Effects in Public Opinion Surveys." *Sociological Methodology* 45(1): 184–222. DOI: https://doi.org/10.1177/0081175015570095.

Ng, J.H., E. Henry, T. Oberlander, P. Shi, S.H. Scholle. 2016. "Shortening a Patient Experiences Survey for Medical Homes." *Healthcare* 4:1. DOI: https://doi.org/10.3390/healthcare4010001.

Paddison, C.A.M., G.A. Abel, M.O. Roland, M.N. Elliott, G. Lyratzopoulos, and J.L. Campbell. 2013. "Drivers of Overall Satisfaction with Primary Care: Evidence from the English General Practice Patient Survey." *Health Expectations* 18: 1081-1092. DOI: https://doi.org/10.1111/hex.12081.

Pew Research Center. 2019. *Mobile Technology and Home Broadband.* Available at: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/06/PI_2019.06.13_Mobile-Technology-and-Home-Broadband_FINAL2.pdf. (accessed September 2021).

Peytchev, A. 2009. "Survey Breakoff." *Public Opinion Quarterly* 73(1): 74–97. DOI: https://doi.org/10.1093/poq/nfp014.

Peytchev, A. and E. Peytcheva. 2017. "Reduction of Measurement Error due to Survey Length: Evaluation of the Split Questionnaire Design Approach." *Survey Research Methods* 11(4): 361–368. DOI: https://doi.org/10.18148/srm/2017.v11i4.7145.

Revilla, M. and C. Ochoa. 2017. "Ideal and Maximum Length for a Web Survey." *International Journal of Market Research* 59(5): 557–565. DOI: https://doi.org/10.2501/IJMR-2017-039.

Rhemtulla, M., V. Savalei, and T.D. Little. 2016. "On the Asymptotic Relative Efficiency of Planned Missingness Designs." *Psychometrika* 81(1): 60–89. DOI: https://doi.org/10.1007/s11336-014-9422-0.

Schober, M.F., F.G. Conrad, C. Antoun, P. Ehlen, S. Fail, A.L. Hupp, M. Johnston, L. Vickers, H.Y. Yan, and C. Zhang. 2015. "Precision and Disclosure in Text and Voice Interviews on Smartphones." *PLoSONE* 10. DOI: https://doi.org/10.1371/journal.pone.0128337.

Schonlau, M. and M. Couper. 2017. "Options for Conducting Web Surveys." *Statistical Science* 32(2): 279–292. DOI: https://doi.org/10.1214/16-STS597.

Sharp, L.M., and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 47: 36–53. DOI: https://doi.org/10.1086/268765.

Spranca, M., D.E. Kanouse, M. Elliott, P.F. Short, D.O. Farley, and R.D. Hays. 2000. "Do Consumer Reports of Health Plan Quality Affect Health Plan Selection?" *Health Services Review* 35(5): 933–947. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089177. (accessed September 2021).

Steinbrecher, M, J. Roßmann, and J.E. Blumenstiel. 2015. "Why Do Respondents Break Off Web Surveys and Does it Matter? Results from Four Follow-up Surveys." *International Journal of Public Opinion Research* 27(2): 289–302. DOI: https://doi.org/10.1093/ijpor/edu025.

Stucky, B.D., R.D. Hays, M.O. Edelen, J. Gurvey, and J.A. Brown. 2016. "Possibilities for Shortening the CAHPS Clinician and Group Survey." *Medical Care* 54(1): 32–37. DOI: https://doi.org/10.1097/MLR.0000000000000452.

Swain, M.S. 2015. "The Effects of a Planned Missingness Design on Examinee Motivation and Psychometric Quality." PhD Dissertation, James Madison University. Available at: https://www.proquest.com/docview/1685390254 (accessed September 2021).

Toepoel, V., and P. Lugtig. 2018. "Modularization in an Era of Mobile Web: Investigating the Effects of Cutting a Survey Into Smaller Pieces on Data Quality." *Social Science Computer Review*: 1–15. DOI: https://doi.org/10.1177/0894439318784882.

Toepoel, V., M. Das, and A. van Soest. 2008. "Effects of Design in Web Surveys: Comparing Trained and Fresh Respondents." *Public Opinion Quarterly* 72(5): 985–1007. DOI: https://doi.org/10.1093/poq/nfn060.

Villar, A., M. Callegaro, and Y. Yang. 2013. "Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys." *Social Science Computer Review* 31(6): 744–762. DOI: https://doi.org/10.1177/0894439313497468.

West, B.T., D. Ghimire, and W.G. Axinn. 2015. "Evaluating a Modular Design Approach to Collecting Survey Data Using Text Messages." *Survey Research Methods* 9:111–123. DOI: https://doi.org/10.18148/srm/2015.v9i2.6135.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response Burden: What Is It and What Predicts It?" In Advances in Questionnaire Design, Development, Evaluation, and Testing, edited by P.C. Beatty, D. Collins, L. Kaye, J.L. Padilla, G. B. Willis, and A. Wilmot: 193–212. New York: Wiley.

Zhang, C., and F.G. Conrad. 2018. "Intervening to Reduce Satisficing Behaviors in Web Surveys: Evidence From Two Experiments on How it Works." *Social Science Computer Review* 36(1): 57–81. DOI: https://doi.org/10.1177/0894439316683923

Received May 2020
Revised July 2021
Accepted December 2021

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 987–1017, http://dx.doi.org/10.2478/JOS-2022-0043

Response Burden and Dropout in a Probability-Based Online Panel Study – A Comparison between an App and Browser-Based Design

Caroline Roberts[1], *Jessica M.E. Herzing*[2], *Marc Asensio Manjon*[1], *Philip Abbet*[3], *and Daniel Gatica-Perez*[4]

Survey respondents can complete web surveys using different Internet-enabled devices (PCs versus mobile phones and tablets) and using different software (web browser versus a mobile software application, "app"). Previous research has found that completing questionnaires via a browser on mobile devices can lead to higher breakoff rates and reduced measurement quality compared to using PCs, especially where questionnaires have not been adapted for mobile administration. A key explanation is that using a mobile browser is more burdensome and less enjoyable for respondents. There are reasons to assume apps should perform better than browsers, but so far, there have been few attempts to assess this empirically. In this study, we investigate variation in experienced burden across device and software in wave 1 of a three-wave panel study, comparing an app with a browser-based survey, in which sample members were encouraged to use a mobile device. We also assess device/software effects on participation at wave 2. We find that compared to mobile browser respondents, app respondents were less likely to drop out of the study after the first wave and the effect of the device used was mediated by subjective burden experienced during wave 1.

Key words: App-based survey; attrition; mobile survey; usability; user satisfaction.

1. Introduction

Mobile Internet technologies presents numerous opportunities for survey research, as well as some important challenges (Link et al. 2014). Respondents can now access online surveys via web browsers on a number of different Internet-enabled devices (notably, desktop PCs and laptops, tablets and smartphones (Callegaro 2010; De Bruijne and Wijnant 2014a; Lugtig and Toepoel 2016; Peytchev and Hill 2010; Struminskaya et al.

[1] Institute of Social Sciences, Faculty of Social and Political Sciences University of Lausanne, Bâtiment Géopolis, Quartier Mouline, 1015 Lausanne, Switzerland. Emails: Caroline.Roberts@unil.ch and marc.asensiomanjon@unil.ch
[2] Interfaculty Centre for Educational Research (ICER), University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland. Email: jessica.herzing@unibe.ch
[3] Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland. Emails: philip.abbet@idiap.ch and gatica@idiap.ch.
[4] EPFL (Ecole polytechnique fédérale de Lausanne), School of Engineering and College of Humanities, Inn building, Station 14, 1015 Lausanne, Switzerland. Email: daniel.gatica-perez@epfl.ch
Acknowledgments: This work was supported by the EPFL and University of Lausanne "Collaborative Research on Science and Society" (CROSS) 2019 Program, as part of the project "Leveraging on-device smartphone inference to address resistance to participate in social surveys (LOIS)"; by FORS, the Swiss Centre of Expertise in the Social Sciences; and by the Faculty of Social and Political Sciences at the University of Lausanne. We would like to thank Oliver Bornet (Idiap) for technical discussions about the mobile platform. We also thank all participants in the study.

2015)). While this range of access options means that long-standing coverage problems associated with web surveys are diminishing (Couper et al. 2017; Lee et al. 2019), research has highlighted difficulties around ensuring data quality where multiple response devices are used in the same survey (Antoun et al. 2017; Callegaro 2013; Lee et al. 2019; Revilla et al. 2016). There is evidence that compared to on PCs, answering questionnaires on mobile devices can take respondents longer (Couper et al. 2017; Couper and Peterson 2017), increase breakoff risk (Buskirk and Andrus 2012; Callegaro 2010; Couper et al. 2017; Mavletova and Couper 2015; Peytchev 2009), and affect measurement quality (Antoun 2015; Mavletova 2013; Mavletova and Couper 2013, 2015). Optimising online surveys for mobile devices in ways that motivate participation, engage respondents, and promote conscientious questionnaire completion has, therefore, become a key priority for survey practitioners (Antoun et al. 2018; De Bruijne and Wijnant 2014b; Mavletova et al. 2018; Peytchev and Hill 2010). This is especially important given the growing use of web-based data collection in longitudinal surveys, where mitigating panel attrition is a central, ongoing challenge (De Leeuw and Lugtig 2014).

Understanding the reasons why response behaviour and engagement vary across different access options is key to improving the design of future online surveys. While these reasons are manifold and intrinsically linked to the characteristics of respondents using them (Lugtig and Toepoel 2016), device and software usability and experienced burden have been identified as parts of the Equation (Callegaro et al. 2015; Couper et al. 2017). Web questionnaires designed for PCs are not always well-adapted to browsers on smaller touch-screen devices, and mobile internet connections (and being physically mobile) are less conducive to sustained concentration on questionnaires over long periods (Antoun et al. 2018; Callegaro et al. 2015; Couper et al. 2017) making the response task more demanding, and time-consuming (Antoun et al. 2018). Mobile software applications (apps) designed for hosting survey questionnaires offer ways to address some of these constraints, as well as other ways to make mobile-web surveys less burdensome and more engaging (Link et al. 2014). Most research investigating the utility of apps for surveys to date, however, has focused on the new measurement tools they offer and respondent willingness to use them to complete alternative data collection tasks (e.g., Jäckle et al. 2019; Keusch et al. 2019; Revilla et al. 2017; Wenz et al. 2019). Few studies have explicitly investigated whether perceptions and experiences of response burden among app respondents are more positive compared to respondents using mobile browsers, and if so, whether this, in turn, affects response behaviour and participation decisions.

In this study, we address this knowledge gap by investigating variation in experienced burden in the first wave of a three-wave panel study, across chosen response devices (PCs (desktop and laptops) verses mobiles (smartphones and tablets) and software (web browser versus mobile app), and the extent to which burden mediates device and software effects on willingness to participate in the subsequent panel wave. Specifically, we address the following research questions:

RQ1: To what extent does experienced response burden vary as a function of response device and software?

RQ2: Does willingness to participate at wave 2 of an online panel study vary as a function of the response device/ software used at wave 1?

RQ3: Does experienced response burden at wave 1 mediate device/software effects on willingness to participate at wave 2?

Before describing the research design and analytic approach in detail, we first review relevant literature relating to design challenges in mixed device web surveys and the problem of response burden – particularly in a panel setting. We also describe some of the opportunities apps offer for optimising web survey design for participants on mobile devices.

2.1. Design Challenges in Web Surveys and the Problem of Response Burden

While the proliferation of access options available for completing web surveys represents good news on the one hand for survey practitioners, on the other, research into its implications for data quality has highlighted areas for concern (De Bruijne and Wijnant 2014b; Maslovskaya et al. 2019; Toepoel and Lugtig 2015). Like modes of data collection, different devices have their own error properties (Couper et al. 2017), which can affect data comparability in mixed device settings (Toepoel and Lugtig 2015). While potentially problematic, measurement differences between devices are generally viewed as a lesser cause for concern than differences resulting from device-related selection errors (Couper et al. 2017; Keusch and Yan 2017; Toepoel and Lugtig 2018; Antoun et al. 2019; Struminskaya et al. 2015), that is, errors resulting from non-coverage and nonresponse which affect who is selected into the response sample (Klausch et al. 2015), and hence, its representativeness.

There are multiple explanations for differences in selection error in estimates based on data gathered from respondents on different devices. Firstly, different devices (and different brands of device and operating systems) tend to be used by different socio-demographic groups, who may, in turn, be more or less likely to use those devices to participate in web surveys. Secondly, in *mobile* web surveys, in which participation is only allowed on a mobile device, unit nonresponse rates are generally higher than for PC web surveys, suggesting an overall negative impact on response propensity of mobile modes. This has been attributed to both respondent and device characteristics (e.g., respondents' level of familiarity with the device and how they habitually use it versus its technological features, like the speed and reliability of internet connection (Couper et al. 2017). Thirdly, user characteristics, device features and environmental influences can all affect respondents' motivation to finish answering the questionnaire once started, as any difficulties experienced (e.g., technical problems, distractions) may cause them to want or have to stop (Link et al. 2014, 19). Indeed, there is consistent evidence that respondents on mobile devices have a higher breakoff rate than those responding on desktop and laptop computers (Buskirk and Andrus 2014; Couper and Peterson 2017; Couper et al. 2017; Guidry 2012; Wells et al. 2013) and that those on smartphones are more likely to quit the survey compared to those on tablets (where breakoff rates for the latter are closer to those on PCs (Guidry 2012; Wells et al. 2013). Less is known about whether, for the same reasons, mobile respondents in web-based panel studies are more likely to drop out of the study in subsequent waves of data collection, but there are reasons (reviewed below) to assume this may well be the case.

To mitigate non-comparability in data quality across response devices in web surveys it is essential to address the underlying mechanisms responsible for negative outcomes for mobile respondents. Explanations for differences in response behaviour by device (besides user characteristics) frequently invoke concepts relating to device usability and response burden. For example, Mavletova and Couper's (2015, 93) meta-analysis found breakoff

rates for mobile respondents to be highest for longer questionnaires with complex design elements (e.g., grids, sliders and images) and for questionnaires designed for completion on a PC (see also Callegaro 2010). The latter require respondents to make additional effort on smaller screens (e.g., to scroll down the screen to read long questions or long lists of response options) and data input using a touch screen may be less comfortable and more error prone than on larger devices (Link et al. 2014).

Studies have also consistently found differences in completion times for respondents answering questionnaires on different devices, with mobile respondents taking longer, on average, than PC respondents (Antoun and Cernat 2020; Couper et al. 2017; Couper and Peterson 2017; Keusch and Yan 2017; Mavletova and Couper 2015). Like breakoffs, longer completion times have been attributed to device-related factors such as the greater need to scroll due to smaller screen sizes, the demands of text input without a keyboard, and increased transmission times due to connection speed (Antoun and Cernat 2020; Couper et al. 2017; Couper and Peterson 2017; Keusch and Yan 2017; Mavletova and Couper, 2013; 2015).

Contextual factors have also been alluded to, however, including mobile respondents being on the move when completing questionnaires, being in the presence of other people or exposed to other distractions, and multi-tasking (Couper et al. 2017; Wenz 2021). Longer completion times should indicate increased objective burden to the extent that they imply greater time sacrifice to complete the survey task and more effort required to finish (Office of Management and Budget 2006). However, the relationship with the respondent's subjective experience of burden is less clear cut, as taking longer to complete a questionnaire may also imply more task engagement (Lynn 2014; Read 2019). For this reason, it is important to take account of the interplay between subjective perceptions of burden and objective hindrances to participation that may influence willingness to continue participating in a survey (Couper et al. 2017), particularly in a longitudinal research setting.

In longitudinal surveys, response burden has also been cited as a common reason for attrition (Hoogendoorn and Sikkel 1998; Laurie 2008; De Leeuw and Lugtig 2014; Kleinert et al. 2019). Attrition refers to respondents dropping out of a panel study, either temporarily or permanently, and can be attributed to a variety of causes, including variation in intrinsic motivations to participate in and commitment to a study, as well as extrinsic factors such as incentives (Lynn 2008; Lugtig 2014). However, experienced burden plays a role in so-called 'panel fatigue' (Laurie et al. 1999; Lemay 2010) and is assumed to accumulate over the course of panel participation (especially with frequent data collection), leading for some to the decision to drop out (Lipps 2009; Watson and Wooden 2009; Lemay 2010). Burden may also play a key part in experienced 'shocks' that can lead to dropout (Lemay 2010; Lugtig 2014; Kleinert et al. 2019), for example, when objective burden in a particular survey wave is greater due to design features such as questionnaire content, length, difficulty, or due to stress or frustration provoked by response tasks (see e.g., Dillman et al. 1993; Galesic and Bosnjak 2009; Marcus et al. 2007; Lynn 2014; Lugtig 2014; and Kleinert et al. 2019). Perceptions (and recall) of such features can vary by response mode (Couper et al. 2017; Gummer and Daikeler 2020), so it is likely that device-related differences may arise in Internet panel studies also.

Further evidence that burden plays a role in panel attrition comes from the fact that attrition is often highest among people for whom the level of burden is objectively greater (Lugtig 2014). This includes, for example, people with lower levels of education or cognitive skills who may find it harder to complete questionnaires (Loosveldt and Carton 2001; Freese and Branigan 2012) and people from ethnic minorities, who may experience increased burden due to language difficulties (Lipps 2009). However, it may also be due to divergent subjective experiences of burden. For example, people with extravert personalities appear to be more susceptible to boredom and panel fatigue (Lugtig 2014). More generally, respondents who gain less enjoyment from completing surveys or respondents who experienced difficulties at the previous survey wave are also more likely to drop out of panels (e.g., Hill and Willis 2001; Kalton et al. 1990; Laurie et al. 1999; Lepkrowski and Couper 2002; Loosveldt and Carton 2001; Olsen 2005; Lugtig 2014). Finally, longer interview times (Hill and Willis 2001) and poor response quality attributed to fatigue (Loosveldt et al. 2002) have also been found to be predictive of attrition. These findings reinforce the hypothesis that the risk of attrition in an online panel study will be greater for respondents completing questionnaires on mobile device browsers if participation is experienced as especially burdensome.

2.2. *Opportunities Offered by Survey Apps*

To mitigate response burden for mobile-web respondents, it is recommended to adapt web questionnaires designed for PCs (or other modes) to accommodate constraints imposed by mobile devices (Buskirk and Andrus 2014; Peytchev and Hill 2010; Antoun et al. 2017; Couper et al. 2017; Herzing 2019). Mobile-optimised survey design has been shown to reduce required effort for mobile device respondents, leading to lower breakoff rates (Mavletova and Couper 2015; Stapleton 2013). Optimised designs can also help to improve respondents' interest and enjoyment, found to be important in motivating (ongoing) participation in mobile web surveys (Bosnjak et al. 2010; Galesic 2006).

Guidelines for adapting web questionnaires for mobile devices apply equally across browser and app-based platforms. However, apps offer additional ways to optimise web surveys to mobile devices that can potentially reduce burden and enhance enjoyment for mobile respondents (Jacobsen and Kühne 2021), which should, in principle, make them even more appealing than conventional browser-based surveys. In an app, the questionnaire resides locally on the phone, so it imposes fewer demands for a persistent internet connection, which may reduce breakoff risk (Link et al. 2014). Questionnaires can more easily be administered to respondents in shorter modules (simultaneously or over time), potentially offering them greater control over the duration of participation in any given response session, which may also mitigate the likelihood of breakoff and perceptions of burden (Johnson et al. 2012; Toepoel and Lugtig 2018). Apps also offer the possibility to contextualise the timing of survey requests (e.g., triggered by location or event) so they are received when participation is more relevant or convenient (Jäckle et al. 2018; Kreuter et al. 2018).

Apps also offer opportunities to potentially reduce response burden by replacing self-report measures with alternative types of data capture, including visual data and passively collected data (e.g., GPS, browser or app logs, or other sensor data (Keusch et al. 2019; Revilla et al. 2019). For example, in a study by Jäckle et al. (2019), respondents were

asked to download an app in order to scan receipts for purchases, allowing them to save time and effort compared to entering details manually. Incorporating multimodal data collection (e.g., capturing photos, videos and audio) alongside conventional measures may also help to increase respondent engagement by making participation in surveys more enjoyable and varied (Link et al. 2014). Apps should, therefore, improve mobile respondents' participation experience, engagement, and motivation to continue participating and to optimise the response process. Assessing these possible benefits is key to finding ways to better optimise web-based surveys for existing and future mobile respondents. If response burden is indeed lower, then app-based data collection may be especially suitable for mobile respondents in online panel studies; just as the benefits of using apps are maximised in longitudinal research designs (Lugtig 2021).

It is important to note that despite their potential advantages, studies to date have generally found low levels of stated and actual willingness to participate in surveys via mobile apps. Here again, factors relating to burden appear among hypothesised explanatory variables (Keusch et al. 2019; Wenz et al. 2019). For example, Wenz et al. (2019, 4) identify potentially burdensome task characteristics (see also Keusch et al. 2019; Revilla et al. 2019), including having to download and install the app to begin with, and whether the survey involves active or passive data collection. Consistent with this, willingness to respond via an app varies both as a function of survey design features as well as respondent characteristics, with those for whom burden to complete tasks is likely lowest responding at higher rates. For example, studies have found younger respondents, and more generally, those who are more familiar with and experienced using mobile devices and other apps, are more willing to respond via a survey app (Pinter 2015; Scherpenzeel 2017; Elevelt et al. 2019; Jäckle et al. 2019; Keusch et al. 2019; Mulder and de Bruijne 2019; Wenz et al. 2019; Lawes et al. 2021; Struminskaya et al. 2021).

This complicates the task of comparing experiences of burden across response devices and software, as selection effects influence who responds via different access options and can bias estimates of burden derived from those samples. It also implies a need to acknowledge that burden may be a more complex phenomenon in the context of mixed device web surveys than originally conceptualised (e.g., by Bradburn 1978; Haraldsen 2004) and operationalised by many researchers (see Yan et al. 2019; Galesic and Bosnjak 2009; Link et al. 2014). This means that efforts to optimise surveys for mobile devices that focus solely on improving questionnaire design features may be insufficient for guaranteeing positive response experiences, as other factors beyond researchers' control (e.g., respondent experience, device usability and environmental factors), may mean that device or software-related differences persist. Nevertheless, to the extent that burden for mobile respondents can be alleviated through optimal design, apps may well be effective at reducing some of the negative effects of mobile devices on survey participation, and by extension, help to improve data quality.

3. Methods

3.1. Data

The data come from a three-wave online panel survey called 'Selects-Civique', conducted in Switzerland during the 2019 Swiss Federal Election campaign, alongside the Swiss

Election Studies ('Selects') (Tresch et al. 2020). The survey included an experimental design (illustrated in Figure 1) to compare a (mobile-optimised) browser-based design with an app-based design. As for the 'Selects' survey, the target population for Selects-Civique was Swiss adults (18 and over) with the right to vote in federal elections, but the sample was restricted to people resident in French-speaking cantons only (due to resource constraints). A sample of 2,175 individuals was drawn from the Federal Statistical Office's sampling frame based on population registers maintained by municipalities and randomly assigned to one of two treatment groups. Group 1 (n = 1,088) received a mailed invitation to participate in the panel via a web browser, while Group 2 (n = 1,087) received a mailed invitation to participate in the panel via a mobile application called 'Civique.org', a pre-existing data collection platform intended as a citizen science initiative to gather multimodal data relevant to local civic causes (www.civique.org – first developed in 2015 and updated over time at Idiap Research Institute, Switzerland by D. Gatica-Perez, J.-I. Biel, O. Bornet, P. Abbet,, and D. Santani, at Idiap Research Institute, Switzerland).

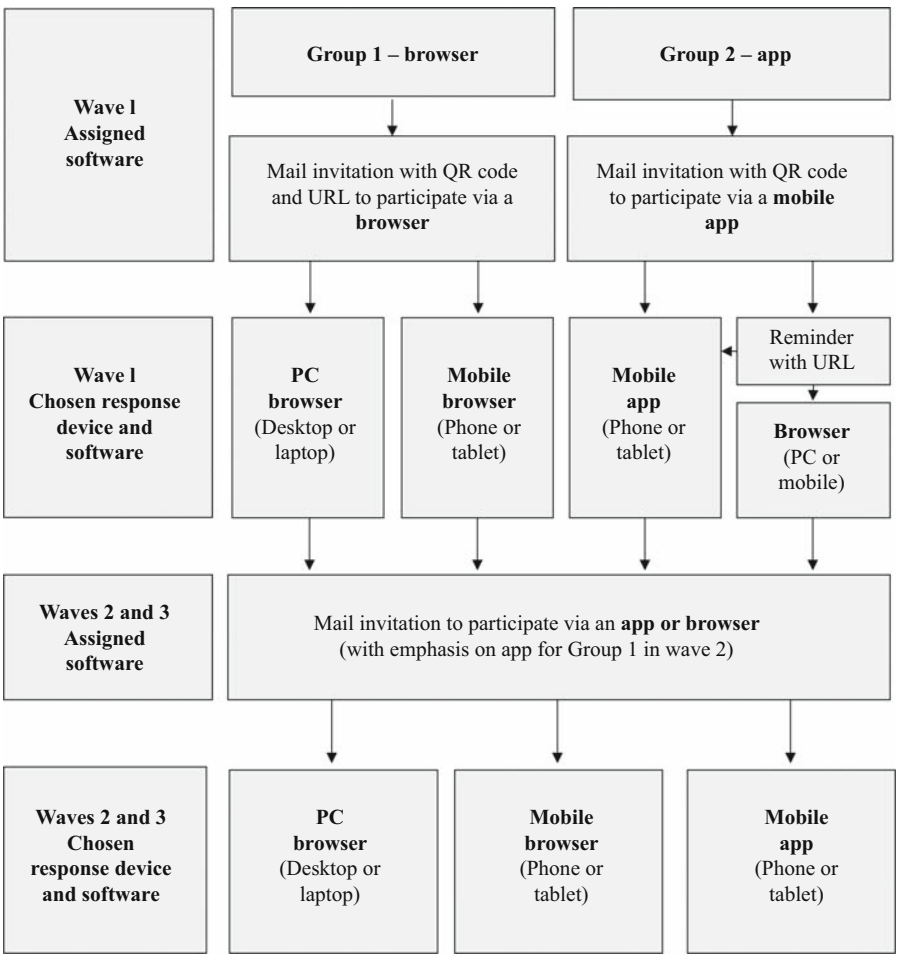


Fig. 1. Research design – Assigned software and chosen response device.

At wave 1 (fielded in May 2019), sample members received an unconditional incentive of USD 10. In both groups, the possibility to use a mobile device was emphasized in the invitation letter by displaying a Quick Response (QR) code, which directly forwarded respondents to the survey landing page (group 1) or the Civique app in either the Google Play or Mac App store depending on which operating system was detected (the app is available for Android and iOS mobile operating systems (but not for PCs)). To enable sample members without mobile devices to participate, the invitation also included a Uniform Resource Locator (URL) link to the survey landing page. Nonrespondents in group 2 were not informed about the browser version until the first reminder of wave 1. At wave 2 (fielded in October 2019), all wave 1 respondents from both groups 1 and 2 (irrespective of response device/ software used) were sent a postal invitation to participate via the app but were simultaneously offered the option of responding via a web browser. Participation via the app was emphasized visually and incentivised with a potentially higher conditional incentive (up to USD 20 for the app participation versus USD 10 for the browser).

Thus, within groups, respondents could self-select their preferred response device and software. At wave 1, group 1 respondents could *only* participate via a web browser but had the choice whether to respond on a PC (desktop or laptop) or on a mobile device (smartphone or tablet) – using the QR code or by typing in the URL. Group 2 respondents could participate via the app (on a smartphone or tablet, but not a PC) or via a web browser on a PC or mobile device (see Figure 1). At wave 2, all respondents could complete the survey using their preferred access option. Our analysis compares mobile respondents using a browser at wave 1 with (a) mobile respondents using the app, and (b) PC respondents using a browser and we address the self-selection into the device response groups using weighting (described further below).

All app users were required to sign a consent form within the app detailing the data privacy and data protection policy (also available to browser respondents on the study website).

3.2. Questionnaires

The questionnaires mainly included questions about political attitudes, voting behaviour, media consumption, social network usage, and socio-demographics. In both the browser and app groups we used an optimal design strategy (Hox et al. 2017) aimed at maximizing data quality in each of the access options. The browser version (programmed in Qualtrics) used a mobile-optimised design, adjusting to screen size, with only one item displayed per screen. The intended completion time was around 20 minutes, the content dictated in part by the design of the parallel "Selects" study.

The app version followed a modular design process that used the available technical features of the platform to improve user experience. This involved splitting the browser version of the questionnaire into thematic sections, each taking one to five minutes to complete. At wave 1, there were nine modules, which were presented in the same order as in the browser questionnaire but were all made available at once, so respondents were free to choose the order of completion. It was not possible to skip questions within modules in the app, but modules could be left out or abandoned. In the browser, respondents were

required to respond to all applicable questions (in the intended, fixed order), but where appropriate, options were provided to allow respondents to withhold their answer.

3.3. Indicators of Response Burden

As response burden can be evaluated both objectively and subjectively, we draw on indicators of both dimensions (Read 2019) – see the Online Supplemental Material (Table S1) for an overview. To measure *subjective response burden*, we analyse respondents' evaluations of experienced burden in the wave 1 questionnaire, based on four items that were asked at the end of the questionnaire (in the last module in the app). The items included statements presented with a five-point fully labelled Agree/Disagree response scale, where 1 meant 'completely agree' and 5 meant 'completely disagree'. The statements were: (1) 'The questionnaire was interesting', (2) 'The length of the questionnaire was adequate', (3) 'The questions were comprehensible', and (4) 'Filling in the questionnaire presented no difficulty'. We assumed that if respondents agreed with the statements, their overall experience at wave 1 was positive and their experienced burden low. Item 1 is relevant to respondent motivation to participate (Groves et al. 2004) and continue participating in the panel (Galesic 2006). Items 2, 3 and 4 are more direct indicators of perceived response burden, relating to length, effort and respondent ability/competence (Read (2019), originally proposed by Bradburn (1978) and Haraldsen (2004)). The items are also akin to common usability metrics used in user experience research, relating to satisfaction, enjoyment, engagement, experienced burden and task success (Geisen and Bergstrom 2017).

After confirming the inter-item correlations were positive and significant, we created a composite measure of subjective burden based on respondents' mean scores for all four items for use in preliminary pairwise comparisons between the groups of interest. Cronbach's alpha for the scale was .71 (a principal components analysis extracting one factor with an eigenvalue greater than one supported this decision). However, two of the items (comprehensibility of the questions and difficulty filling in the questionnaire) had lower correlations with the other two and did not differentiate respondents or device groups well (and indeed, were negatively correlated with the second factor, which had an eigenvalue just below one – output available in the See Online Supplementary Material). For this reason, we also computed a mean score based only on the other two items (questionnaire length and interest) for use in the subsequent (mediation) analyses.

To measure *objective response burden*, we calculated wave 1 completion times based on module completion timings from the app (summing the time taken to complete all the modules), and screen-by-screen timing variables from Qualtrics, which were summed across the items according to how they were grouped in each app module, and then across all the modules. We report mean completion times in minutes, normalised based on +-2 standard deviations (SD) from the mean. Completion times for a total of 16 cases (nine for the app, seven for the browser) were two SD above the mean, and we substituted these with the average time taken by the remaining respondents (following Revilla and Ochoa 2015), as these were likely indicative of interruptions rather than necessarily due to slower pace.

The number of items that were applicable to all respondents was slightly lower in the app version (91 questions applicable to all respondents) than in the browser version for

respondents with a smartphone (96 questions applicable to all), and slightly higher than for browser respondents without a smartphone (83 questions applicable to all). The differences were in a module of questions on smartphone use, in which certain items were deemed not to be relevant to those already responding via the app (a question about the smartphone's operating system; a question on smartphone skills; and questions on willingness to complete a questionnaire on a mobile phone; to, download a survey app and to share GPS location) or to those without a smartphone, who answered a subset of questions about activities completed on the Internet instead. In addition, the *actual* number of questions answered varied by respondent as a function of whether follow-up questions to filters were asked. For this reason, we also compare groups on the number of answers given in wave 1 as an indicator of objective burden. However, as this variable was not normally distributed, highly associated with the response device and correlated with the response duration variable, we excluded it from the subsequent multivariate analyses.

3.4. Analytic Approach

Based on the reviewed literature, we hypothesise that both response burden and dropout will be greatest for mobile-browser respondents, compared with mobile-app respondents and PC-browser respondents. Underpinning this hypothesis is the assumption that variation in wave 2 drop-out by device and software is mediated by variation in experienced response burden at wave 1, which we test through a mediation analysis. On this basis, the analysis had two aims. The first was to compare subjective and objective experienced response burden across respondents completing wave 1 on different devices and using different software (RQ1). The second was to assess whether response device/software used at wave 1 predicts non-participation at wave 2 (RQ2), and if so, whether and to what extent the effect of device on non-participation at wave 2 is mediated by response burden experienced at wave 1 (RQ3).

To address RQ1, we initially conduct pairwise comparisons using t-tests to test differences in means for the subjective and objective measures of response burden described above. Subsequently, we use a regression-based approach as part of the mediation analysis conducted to address the other research questions (described in the next section). We focus on comparing those using a mobile browser with those using a PC browser to provide some control for the software type (notwithstanding possible differences between browser providers), and those using a mobile browser and those using the mobile app to provide some control for the device type. However, for interested readers, we also present the comparison between PC and App respondents (though this confounds device and software). At wave 1, there were only 20 app respondents who used a tablet and 32 mobile browser respondents who used a tablet, and due to small achieved sample sizes overall, we decided to pool tablet and smartphone respondents in all our analyses. We recognise that, for some users, tablets may be used in similar ways to laptop PCs. Our approach emphasises the greater portability of tablets and their typically smaller screen sizes, as well as the fact that the app used was available for tablet, but not PC. Future studies with larger samples responding on tablets may reconsider this classification.

Because our comparisons of interest are confounded by selection effects on the samples responding using different devices/software, we used a propensity score weighting

approach to try to balance the samples using auxiliary data available for all sample members, following general recommendations for addressing questions of causal inference in social research (Harder et al. 2010; Rosenbaum and Rubin 1983). We first computed an inverse probability weight to address selectivity due to nonresponse at the first wave. We then computed separate weights for the pairwise comparisons across devices to control for the differential probability of responding using one software/ device type compared to the other and combined these with the general nonresponse weight for each of the pairwise comparisons. For the remaining analyses, in which we used regression-based methods to test our mediation hypothesis, we use the general nonresponse weight on its own. Details of how the weights were computed are available in the Online Supplemental Material (Computation of Weights).

3.4.1. Testing for Mediation

To address RQ2 and RQ3 and test the hypothesis that the effect of the response device on wave 2 dropout is mediated by experienced response burden at wave 1, we followed procedures for mediation analysis (Hayes 2017) based on those proposed by Baron and Kenny (1986), which are used widely in the social sciences (VanderWeele 2016). Mediation analysis allows the researcher to establish the extent to which an independent variable (e.g., response device) influences a dependent variable (e.g., dropout at wave 2 of a panel) "through one or more *mediator* variables" (e.g., experienced response burden) (Hayes 2017). In other words, it helps to shed light on *how* one variable influences another variable and explicitly tests hypotheses relating to the possible mechanisms involved (Hayes 2017).

At step 1, we regressed the indicator for non-participation in wave 2 (coded 1 if the person did not participate in wave 2 (dropped out) and coded 0 if they did participate) on the wave 1 device indicators (dummy variables indicating those who responded on a PC and those who responded via the app versus those who responded on a mobile browser (reference category). At step 2, we regressed the mediator variables (the two-item subjective burden indicator (questionnaire length and interest) and the measure of completion time) on the device indicators, to assess the relation between device and response burden. As we used multiple indicators of response burden, we first tested the mediation hypothesis separately for the subjective and objective indicators (together with the control variables). Although using the app significantly predicted reduced completion times, in the presence of the other variables in the model, this variable was not significantly associated with non-participation at wave 2. We focused, therefore, on the two-item measure of subjective burden, keeping completion time as a control variable. Both the completion time (in minutes) and subjective burden variables (the mean of two five-point scales) were recoded to range from 0 to 1, where 0 represented the minimum score, and 1 represented the maximum level of experienced burden – that is, the longest completion time and strong disagreement that the questionnaire was interesting and the length was adequate. At step 3, we added the mediator to the model predicting non-participation at wave 2, to assess whether subjective burden significantly predicts dropout, and whether the relation between device and dropout remains significant when controlling for response burden.

The mediation analysis was conducted using the SPSS macro 'PROCESS' version 3.5 (see Hayes 2017). As the main outcome variable was binary, we used logistic regression

analysis at steps 1 and 3, and Ordinary Least Squares (OLS) regression at step 2. The procedure produces estimates of the direct and indirect effects of the dependent variable (device) via the mediator on the logged odds scale. To test whether the indirect effect is statistically significant, the macro uses non-parametric bootstrapping to estimate standard errors and 95% confidence intervals (based on 5,000 bootstrapped samples). To address potential confounding of the assumptions underpinning the mediation hypothesis (see VanderWeele 2016, 19–21), we included a number of control covariates in the logistic regression analyses (shown in Table S2 and discussed in detail in part C of the Online Supplemental Material). The same set of covariates were included in all models (and as independent variables in the OLS regression predicting the mediator). As PROCESS cannot handle weighted data, we present the results of the mediation for the unweighted data only.

4. Results

Before presenting the results of the analyses addressing the three research questions, we first present details of participation rates and breakoffs by response device and software, which provide insight into the extent of self-selection into the response samples and some preliminary differences of interest between groups. These are shown in Table 1. At wave 1, a total of 687 (31.6% – AAPOR RR2) sample members participated in the survey, 366 (33.6%) in the group assigned to the browser condition (group 1) and 321(29.5%) in the group assigned to the app (group 2). The difference in response rates between the treatment groups was statistically significant ($\chi^2(1) = 4.25$; $p < .05$). Of those participating, 298 (43.4%) participated via a PC browser, 152 (22.1%) via a mobile browser, and 237 (34.5%) via the app. Only 358 respondents participated in wave 2 (52.1% of those participating at wave 1), of which 139 (38.8%) used a PC browser, 23 (6.4%) used a mobile browser and 196 (54.7%) used the app. Out of the 184 wave 1 browser respondents from group 1, 89 (48.4%) switched to the app in wave 2, whereas 81 (44.0%) responded via a browser again. Only 14 respondents from group 1 (7.6%) responded on a mobile browser at wave 2. From group 2, 174 (54.2%) respondents participated in wave 2, of which 109 (62.6%) participated via the app and the remainder via a browser (only ten (5.8%) on a mobile). The dropout rate for the 32 tablet users among mobile browser respondents was 56.3% compared with 54.8% for smartphone users; for the 20 tablet users in the app group, the dropout rate was 20% compared with 42.6% for smartphone users.

We define breakoffs based on the modular questionnaire design used in the app in two ways: (1) starting the survey and not completing the 9th and final module; and (2) starting the survey and not completing the last three modules. The final module included the questionnaire evaluation measures used as indicators of subjective burden, so failure to complete this module meant that the respondent was excluded from our analyses. Out of the 687 respondents, 621 had complete data for this module, and form the analytic sample for the subsequent analyses. Not completing the final three modules meant that the participant failed to complete the module of socio-demographic questions (module 7), rendering their preceding answers less usable (module 8 was about motivations for participating, so may have been considered less pertinent to the main survey topic and

Table 1. Participation and break-off rates by treatment group, device and software type.

	Total		PC Browser		Mobile Browser		App	
	n	%	*n*	%	*n*	%	*n*	%
Wave 1								
Group 1 (browser) (N[4]=1,088)	366	33.6	232	63.4	134	36.6	-	-
Group 2 (app) (N=1,087)	321	29.5	66	20.6	18	5.6	237	73.8
Total (N=2,175)	687	31.6	298	43.4	152	22.1	237	34.5
Break-off[1]	65	3.0	22	7.4	16	10.5	27	11.4
Break-off[2]	58	2.7	19	6.4	16	10.5	23	9.7
Prefer not to do W2	153	22.3	68	22.8	42	27.6	43	18.1
Analytic sample[3]	621	28.6	275	44.3	136	21.9	210	33.8
Wave 2								
Group 1 (N=366)	184	50.3	81	44.0	14	7.6	89	48.4
Group 2 (N=321)	174	54.2	55	31.6	10	5.8	109	62.6
W1 PC (N=298)	164	55.0	113	68.9	3	1.8	48	29.3
W1 Mobile (N=152)	62	40.8	6	9.7	14	9.7	42	67.7
W1 App (N=237)	132	55.7	20	8.4	6	4.5	106	80.3
Total (N=687)	358	52.1	139	38.8	23	6.4	196	54.7
Break-off[1]	17	4.7	1	0.7	5	21.7	11	5.6
Break-off[2]	12	3.4	1	0.7	4	17.4	7	3.6

Notes: [1]Break-off = started survey but did not complete final module (questionnaire evaluation). [2]Break-off = started survey but did not complete the last three modules. [3]Analytic sample includes all respondents responding to questionnaire evaluation measures. [4]N= Total sample contacted at waves 1 and 2; *n* = sample responding.

hence, provoked participants to quit). According to these definitions, breakoff rates were higher for respondents on mobile devices (whether using a browser or app) than for PC respondents. Breakoff rate (1) at wave 1 was 7.4% for participants using PCs, 10.5% for participants using a mobile and 11.4% for participants using the app; and rate (2) was 6.4% PC participants, 10.5% for mobile participants and 9.7% for participants using the app. Differences between samples responding on different devices were not significant, however. Respondents who did complete the second-to-last module were asked how they would like to be contacted at wave 2, with the option to say they would prefer not to participate again. A higher proportion of respondents using the mobile browser (27.6%) selected this option than of respondents on a PC (22.8%; difference not statistically significant) or using the app (18.1%; $\chi^2(1) = 4.88$; $p < .05$).

4.1. Device/Software Effects on Experienced Response Burden (RQ1)

Table 2 shows the adjusted means (weighting for nonresponse at wave 1) for the subjective burden indicators by device/ software and the results of independent samples t-tests for each of the pairwise comparisons. Participants' evaluations of the wave 1 questionnaire were generally positive across all response device groups. However, statistical differences were observed between the main comparison groups of interest (mobile browser versus pc and mobile browser versus app) on some measures. App respondents were significantly more likely than mobile respondents to agree the questionnaire was interesting (mean [M] = 2.12, SD = .80 vs. M = 2.36, SD = .89; t-value[t] = 2.70 (Levene's Test for Equality of Variances was statistically significant so equal variances are not assumed), degrees of freedom [d.f.] = 265.1, p-value [p] <.001), and that the length of the questionnaire was adequate (M = 1.76, SD = .82 vs. M = 2.55, SD = 1.10; $t = 7.32$ (equal variances not assumed), d.f. = 230.9, $p < .001$). App respondents also had significantly lower mean scores on the composite measures, indicating overall less experienced burden (see column 5, Table 2). Respondents who used a PC browser at wave 1 were more likely to agree that the length of the questionnaire was adequate compared with respondents who used a mobile browser (M = 2.28, SD = 1.02 vs. M = 2.55, SD = 1.10; $t = 2.61$ (equal variances not assumed), d.f. = 249.6, $p < .05$). These samples also differed significantly on the composite score combining the questionnaire interest and length measures. There were no differences on the other subjective burden measures (see column 4, Table 2). Subjective burden measures were also significantly more positive for app users than for PC users, except for the indicator for difficulty (see column 6, Table 2).

Compared with mobile browser respondents, app respondents had significantly shorter completion times (M = 21.43, SD = 8.75 vs. M = 23.61, SD = 10.95; $t = 2.32$ (equal variances not assumed), d.f. = 238.36, $p < .05$), despite answering more questions in total (M = 106.01, SD = 2.90 vs. M = 98.52, SD = 2.51; $t = -25.68$ (equal variances not assumed), d.f. = 310.93, $p < .001$). There was no significant difference between the two browser groups in terms of completion times. However, the mean number of questions answered varied significantly between respondents using a PC browser and mobile respondents (M = 100.59, SD = 4.45 vs. M = 98.52, SD = 2.51; $t = -4.93$ (equal variances not assumed), d.f. = 390.40., $p < .001$). App respondents also differed from PC

Table 2. Subjective and objective response burden measures – adjusted means by wave 1 device type and pair-wise comparisons.

	Means[1]			Pairwise contrasts[2]		
	(1) PC Browser (n = 275)	(2) Mobile Browser (n = 136)	(3) App (n = 210)	(4) Mobile-PC Browser	(5) Mobile Browser -App	(6) PC Browser - App
	Mean (SD)	Mean (SD)	Mean (SD)	(2)-(1) *p*	(2)-(3) *p*	(3)-(1) *p*
Subjective burden indicators:						
The questionnaire was interesting (1-5[3])	2.27 (0.96)	2.36 (0.89)	2.12 (0.80)	0.09	0.24**	0.15†
The length of the questionnaire was adequate (1-5)	2.28 (1.02)	2.55 (1.10)	1.76 (0.82)	0.27*	0.79***	0.52***
The questions were comprehensible (1-5)	1.43 (0.65)	1.38 (0.53)	1.31 (0.53)	-0.05	0.07	0.12*
Filling in the questionnaire presented no difficulty (1-5)	1.49 (0.76)	1.42 (0.68)	1.45 (0.81)	-0.07	-0.03	0.04
Mean score on all items	1.87 (0.66)	1.93 (0.54)	1.66 (0.55)	0.06	0.27***	0.21***
Mean of interesting and length	2.28 (0.88)	2.46 (0.87)	1.94 (0.71)	0.18*	0.52***	0.34***
Objective burden indicators:						
No. of questions answered	100.59 (4.45)	98.52 (2.51)	106.01 (2.90)	-2.07***	-7.49***	-5.42***
Completion duration in minutes	25.03 (10.36)	23.61 (10.95)	21.43 (8.75)	-1.42	2.18*	3.60**

Notes: [1]Means adjusted for wave 1 nonresponse. [2]Pairwise contrasts based on Independent Samples T-Tests using propensity-score weights to adjust for observed differences in sample composition for each pairwise comparison combined with wave 1 nonresponse weights; [3]Agree-disagree response scales where 1 = completely agree and 5 = do not agree at all. SD = standard deviation; *p* = p-value: † $p < 0.1$, *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

respondents on the objective burden indicators, with faster completion times, despite answering more questions.

To summarise, we find significant variation in experienced response burden as a function of response device and software (RQ1), even when adjusting for observed differences in the device/software comparison groups due to differential selection error at wave 1.

4.2. *Effects of Wave 1 Response Device on Wave 2 Participation (RQ2)*

Parameter coefficients for the logistic regression model predicting drop-out at wave 2, estimated at step 1 of the mediation analysis are shown in Table 3 (Model 1). The overall fit of the model was significant ($\chi^2(21) = 52.836$, $p < .001$; Hosmer and Lemeshow's test was non-significant, also indicating good model fit), and, based on Nagelkerke's pseudo $R^2_{,}$ the model accounted for around 11% of the variation in the probability of wave 2 nonresponse. Controlling for the socio-demographic and the other control variables, both app and PC respondents were significantly less likely to drop out of the study after wave 1 compared to respondents who used a mobile browser (see column 1, Table 3). The odds ratio for responding on an app (exponentiated beta coefficient [Exp(B)] = .516, confidence interval $[CI]_{.95} = [.267, .997]$) indicates that the odds of dropping out of the study after wave 1 were around 48% lower for app respondents than they were for mobile respondents. The odds ratio for responding on a PC (Exp(B) = .630, $CI_{.95} = [.394, 1.01]$) indicates that the odds of dropping out for PC respondents were around 37% lower than they were for mobile respondents. Thus, we indeed find differences in willingness to participate at wave 2 of an online panel study as a function of the response device/software used at wave 1, with mobile browser respondents dropping out at a higher rate than PC or app respondents.

4.3. *Mediating Effects of Experienced Response Burden on Non-Participation at Wave 2 (RQ3)*

To assess whether experienced response burden mediates effects of the response device on the decision not to participate in wave 2, we first estimated the parameter coefficients of the Ordinary Least Squares (OLS) regression equation predicting subjective burden with the device indicators and the control variables from model 1 (see Table 3, Model 2). The overall fit of the model was good ($R^2 = .328$; $F_{(21,599)} = 13.95$; $p < .001$). Responding using the app (compared to a mobile browser) significantly reduced experienced subjective burden ($B = -.101$, SE = .028, $p < .001$), while responding on a PC (compared to a mobile browser) reduced burden, but not significantly ($B = -.035$, SE = .020, $p = .078$). These results support the findings of the t-tests that experienced response burden varies as a function of response device and software (RQ1), even when controlling for other respondent characteristics.

Independent of response device, respondents who were interested in politics also reported significantly less burden ($B = -.056$, SE = .016, $p < .001$), as did respondents who were motivated to participate in wave 1 by the possibility to contribute to science ($B = -.361$, SE = .030, $p < .001$). Respondents who reported using the Internet less than once a day reported higher levels of burden ($B = .040$, SE = .019, $p < .05$), as did

Table 3. Coefficients from logistic regression models predicting non-participation at wave 2(1 and 3) and OLS regression predicting subjective burden (2).

	(1) Non-participation at Wave 2				(2) Subjective Burden			(3) Non-participation at Wave 2 (+ subjective burden)			
	B	SE	*p*	Exp(B)	*B*	SE	*p*	*B*	SE	*p*	Exp(B)
Device (ref. Mobile):											
PC browser	-.463	.239	.053†	.630	-.035	.020	.078†	-.415	.242	.086†	.661
App	-.662	.336	.049*	.516	-.101	.028	.000***	-.516	.342	.132	.597
Subjective burden	-	-		-	-	-		1.493	.500	.003**	4.449
Response duration	.250	.571	.661	1.284	.029	.048	.544	.208	.576	.718	1.231
Respondent sex: female	.126	.187	.502	1.134	.002	.016	.899	.129	.189	.497	1.137
Respondent age (ref. 18–30)											
Aged 31–55	.588	.293	.045*	1.800	.013	.024	.594	.576	.295	.051†	1.779
Aged 56 and over	.590	.378	.119	1.804	-.060	.031	.055†	.687	.383	.073†	1.988
Marital status (ref. never married):											
Married	-.323	.267	.226	.724	.056	.022	.014*	-.408	.271	.132	.665
Divorced, widowed	.005	.309	.987	1.005	.068	.026	.009**	-.096	.314	.761	.909
Household size (ref. 1 member):											
2 members	.501	.284	.078†	1.651	.001	.024	.966	.506	.286	.077†	1.659
3 or more members	.345	.290	.234	1.412	-.002	.024	.935	.355	.292	.224	1.427
Lives in a rural area	-.060	.190	.753	.942	.033	.016	.039*	-.109	.193	.572	.897
Main activity (ref. in F/T paid work):[1]											
In P/T paid work	-.428	.242	.076†	.652	-.023	.020	.245	-.403	.244	.098†	.668
Student / apprentice/ in training	.086	.355	.809	1.090	-.011	.030	.711	.104	.357	.771	1.109
Retired/ unemployed/ home-maker	-.440	.254	.083	.644	.002	.021	.927	-.455	.257	.077†	.634
Education (ref. Secondary level or lower):[2]											
Tertiary level	-.473	.179	.008**	.623	.006	.015	.668	-.494	.181	.006**	.610
Interested in politics	-.141	.184	.446	.869	-.056	.016	.000***	-.057	.188	.761	.944
Motivated to contribute to science	-1.372	.357	.000***	.254	-.361	.030	.000***	-.869	.394	.027*	.419
Uses Internet less than once a day	-.404	.227	.075†	.668	.040	.019	.034*	-.477	.232	.039*	.620

Table 3. Continued

	(1) Non-participation at Wave 2				(2) Subjective Burden			(3) Non-participation at Wave 2 (+ subjective burden)			
	B	SE	*p*	Exp(B)	*B*	SE	*p*	*B*	SE	*p*	Exp(B)
Doesn't use smartphone to access Internet	-.269	.292	.357	.764	-.029	.024	.220	-.224	.295	.447	.799
Does use tablet to access Internet	-.099	.189	.601	.906	.013	.016	.409	-.123	.192	.520	.884
Invited to app at wave 1 (Group 2)	.101	.265	.702	1.107	-.014	.022	.519	.122	.269	.648	1.130
Constant	.949	.473	.045*	2.582	.539	.039	.000***	.171	.540	.752	1.186
N	621				621				621		
Nagelkerke R^2/ R^2	.109				.328				.127		

Notes: [1] F/T = full-time; P/T = part-time. [2]19 cases selecting the 'other education' category were added to the reference category (primary). ref. = reference category; N = cases analysed; B = Unstandardised beta coefficients (on logged-odds scale for logistic regressions [1] and [3]); SE = standard error; Exp(B) = Exponentiated beta coefficient; *p* = p-value: † $p < 0.1$, *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Data weighted to adjust for nonresponse at wave 1.

respondents living in a rural area ($B = .033$, SE $= .016$, $p < .05$). Finally, compared to respondents who were single (never married) married ($B = .056$, SE $= .022$, $p < .05$) and divorced ($B = .068$, SE $= .026$, $p < .05$) respondents reported slightly increased burden. None of the other independent variables were significant predictors of subjective burden.

Given that responding on a PC browser compared with a mobile browser was not significantly predictive of increased burden (only at the 10% level), we focused the mediation analysis on the effect of using the app compared to a mobile browser (keeping responding on a PC in the model as a control covariate). Table 3 (Model 3) shows the log-odds and odds ratios for the full model predicting non-participation at wave 2. Subjective burden has a statistically significant and positive direct effect on the outcome variable ($\text{Exp}(B) = 4.449$, $\text{CI}_{.95} = [1.671, 11.842]$). For each unit increase in reported subjective burden, the odds of not responding at wave 2 of the study increase by around 345%. The direct effect of the app indicator is reduced in the presence of the subjective burden measure and is no longer statistically significant at the 5% level (adjusted $\text{Exp}(B) = .597$, $\text{CI}_{.95} = [.305, 1.167]$; unadjusted $\text{Exp}(B) = .556$, $\text{CI}_{.95} = .278, 1.110]$). The effect of using the app compared to a mobile browser on non-participation at wave 2 is, thus, at least partially mediated by subjective burden. The path model (with the (unadjusted) direct and indirect effects presented in a log-odds metric) is shown in Figure 2 (note that these estimates are based on the full model including all the control variables shown in Table 3). The (unadjusted) indirect effect of the app via burden on drop-out is negative and statistically significant ($\text{Exp}(B) = .855$, $\text{CI}_{.95} = [-.353, -.042]$; effect tested using non-parametric bootstrapping – see Figure 2). In other words, the reduction in experienced subjective burden resulting from using the app instead of a mobile browser results in a reduction in the odds of dropping out of the study after wave one of around 15%. This is illustrated in Figure 3, which shows the probability of dropping out for app and mobile browser respondents across levels of response burden. These findings confirm the hypothesis that experienced response burden at wave 1 mediates device effects on willingness to participate at wave 2 of an online panel study (RQ3).

Among the control variables, it is noteworthy that the size of the coefficient for being motivated to participate by the possibility to contribute to science (highly significant in both models 1 and 2) is also reduced by the inclusion of the mediator, and its effect on the outcome is also partially mediated via subjective burden (Indirect Effect [IE] $\text{Exp}(B) = .571$, $\text{CI}_{.95} = [-1.02, -.187]$.

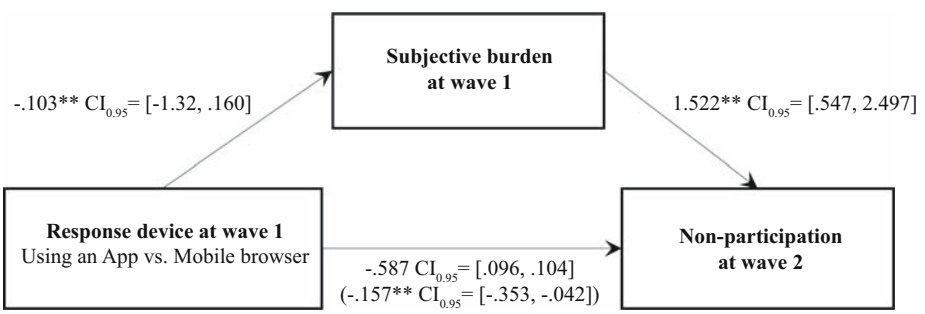


Fig. 2. Regression coefficients and 95% confidence intervals for the relationship between response device at wave 1 and non-participation at wave 2 (direct and indirect effects (in parentheses) shown in a log odds metric).

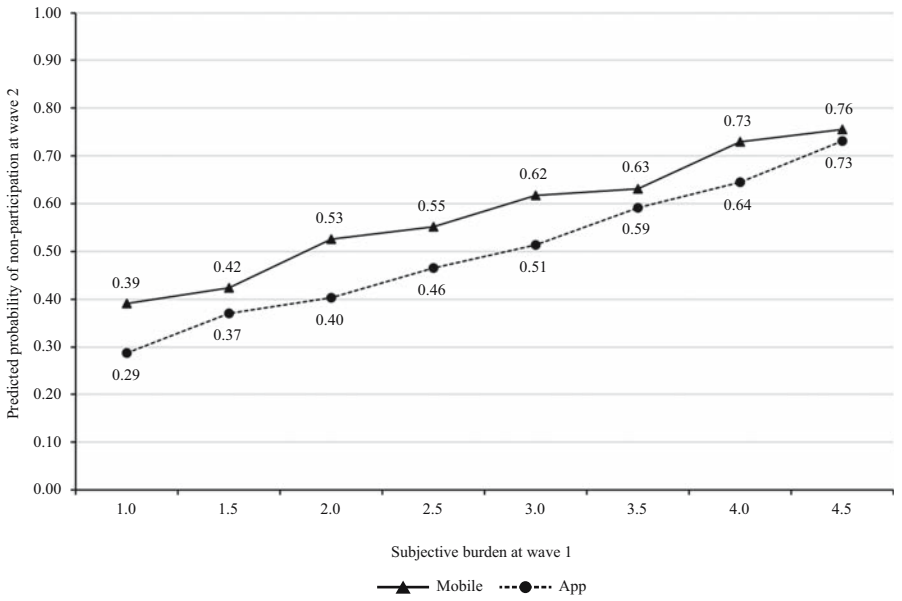


Fig. 3. Predicted probabilities of non-participation at wave 2 for respondents using a mobile browser or an app across levels of experienced subjective burden.

5. Discussion

In this article we reported the results of a study using data from an online panel survey with an embedded experiment comparing an app-based design and a mobile-adapted web browser design. Previous research has found that participants in online surveys using browsers on mobile devices are more likely to break off (see Mavletova and Couper 2015 for a review), due, in part, to greater experienced response burden (Allum et al. 2018) or reduced enjoyment (Bosnjak et al. 2010). In the context of panel surveys, survey experience and increased response burden influence the likelihood of attrition (Gummer and Daikeler 2020; Lugtig 2014), meaning that panel respondents using mobile devices may be at greater risk of dropping out. While studies investigating correlates of hypothetical willingness to participate in mobile data collection have also identified burden-related factors as sources of resistance (e.g., Read 2019; Wenz et al. 2019), few studies have explicitly tested the mediating effect of burden on *actual* participation – especially in the context of a probability-based general population survey. For these reasons, we assessed whether and how respondents' experienced burden using a given device (PC versus smartphone or tablet) or software (browser versus app) in the first wave of a panel affected their response propensity in the second panel wave.

Our first research question concerned the extent to which experienced response burden varied as a function of device and software. In the bivariate analyses, we observed a number of differences in subjective and objective burden between respondents using different devices, with higher levels of burden for mobile browser respondents compared with app and PC respondents. In the mediation analysis, responding using an app (compared to responding on a mobile browser) was associated with significantly reduced subjective burden, even when controlling for completion times, sociodemographic differences in the composition of the response samples, and other control variables known

to correlate with participation in surveys/ attrition, that could plausibly also relate to experienced burden (such as topic interest). Responding on a PC, however, was not significantly associated with burden (or at least, was so to a lesser extent).

Our second research question concerned the extent to which willingness to participate at wave 2 varied as a function of the wave 1 response device. Respondents who used a mobile browser at wave 1 of the study were significantly more likely not to participate in wave 2 than respondents who used the app or a PC browser. Our third research question was whether experienced response burden at wave 1 mediates device effects on willingness to continue participating at wave 2. When the measure of subjective burden was added to the model predicting non-participation at wave 2, it was strongly and positively significant, while the effect of responding using the app was no longer so. The mediation analysis confirmed a significant, negative indirect effect of the app on drop-out in wave 2 via subjective burden. This implies that responding on an app can motivate ongoing panel participation due to the app's capacity to reduce response burden.

Because responding on a PC was not a significant predictor of subjective burden, we did not test the mediation hypothesis for the comparison between PC and mobile browser respondents. However, the effect of responding on a PC on subjective burden was 'significant' at the 10% level. We noted effects of this size because small samples combined with low response rates in this study likely affected the possibility to detect statistical relationships of interest (with a larger sample, we might have observed a similar mitigating effect, as has been concluded by Mavletova and Couper 2015). The focus on the positive effects for respondents of completing the survey using the app, however, is of greater contemporary interest (both theoretical and practical) because of the many opportunities app-based surveys offer researchers in terms of new data collection capabilities (Jäckle et al. 2018; Link et al. 2014).

We addressed potential confounding of the assumptions underpinning the mediation model (VanderWeele 2016) by including several control covariates in the regression analyses. Of these potential confounds, two variables were found to be statistically significant *negative* predictors of non-participation at wave 2 of the study: having completed a tertiary-level qualification and being motivated to participate in the survey by the possibility to contribute to science. Level of education was not significantly associated with experienced burden. However, respondents who were motivated by the possibility to contribute to science reported less burden; they were more likely to evaluate the questionnaire length as adequate and the questionnaire content as interesting, and were subsequently, less likely to drop-out of the panel at wave 2 (in line with findings of Keusch et al. 2019). Another potential confounding variable was topic interest. However, although respondents who were more interested in politics were also more likely to agree the questionnaire was interesting, political interest was not associated with non-participation at wave 2. Sensitivity analyses can be used to assess in more detail whether confounding violates assumptions of mediation analysis (VanderWeele 2015, 2016), but given these findings, we did not investigate this further.

5.1. Limitations

The need to address potential confounds is particularly important given that we did not have strict control over exposure to the treatment of interest – that is, the response devices

selected by respondents. While software type (whether the respondent used the app or a browser) was randomly assigned, the choice of device (PC or mobile) was based on respondent preferences. This means that differences in the characteristics of respondents using different devices at wave 1 could account for both experiences of burden and response propensity at the second wave of the survey. Besides the inclusion of control variables in the regression analyses already described, we also used propensity score weights to control for self-selection into the response samples. It should be noted that the effectiveness of this method depends on which variables are used to estimate response probabilities (Roberts et al. 2020). We were able to make use of administrative data from the sampling frame but found relatively few differences between response samples on the sociodemographic variables analysed (see Online Supplemental Material, Table S2). As a result, we cannot rule out the possibility that unobserved selection errors may be partially responsible for the findings reported here (Lugtig and Toepoel 2016).

Another feature of the study design that may have influenced our findings is the fact that respondents assigned to the browser-based design were invited to switch to the app at wave 2 (and encouraged to do so through the offer of a higher incentive), while the group 2 respondents who used the app at wave 1 could continue using the app. In other words, effects of the response device used at wave 1 are potentially confounded with the effect of the mode switch for group 1 respondents, which has been shown in other studies to lead to drop out (e.g., Sakshaug and Kreuter 2011; Sakshaug et al. 2010; Tourangeau et al. 2013). Nevertheless, respondents in both groups were informed that they could continue using the browser to participate if they wished to do so and presumably those with strong preferences for using a browser (or no alternative) would have continued to use it in subsequent panel waves if motivated to do so.

The difference found in experiences of using the app versus using a mobile browser implies that the underlying causes of burden for mobile respondents may be less about user, device or environmental influences, but rather, the freedom and flexibility (and novelty, perhaps) offered by the software. In the app, it was possible to use a modular questionnaire design where all modules were available at once, giving respondents complete control over when, where, and how to complete them. This had a positive influence on respondents' perceptions of the adequacy of the length of the questionnaire and of how interesting it was, and subsequently, on their probability of participation at wave 2. This is partially consistent with the findings of Johnson et al. (2012), who found positive benefits of a modular design in an app on breakoffs (though in our study, the wave 1 breakoff rate was still higher in the app than for browser respondents). Toepoel and Lugtig (2018) reported greater dropout *with* a modular design, but one in which modules were administered over the course of several days. These findings suggest there may be benefits of making all modules available simultaneously.

As we are not able to disentangle the 'pure' effect of using an app from the effect of the modular questionnaire, we recommend that future research investigate the advantages and potential disadvantages of alternative modular designs, both within and across different response devices, in both app- and browser-based surveys. This is also needed to inform understanding of how modular questionnaire designs are used by respondents and of any possible negative effects they may have (e.g., on measurement quality due to differential context effects linked to question order – see Dillman et al. (2014) for an overview).

6. Conclusion

This study lends further support to a well-established and fairly consistent literature about the negative effects of perceived and experienced burden on willingness to participate in surveys. It also confirms the assumptions of previous research into device differences in survey outcomes that they appear to be mediated by differences in experienced burden. Experienced burden in the study reported here was greatest for respondents who completed the first wave using their mobile browser, and this had a negative impact on willingness to participate at the subsequent panel wave. The novel contribution, however, is the finding that using a mobile app (with a modular questionnaire design) effectively reduces response burden compared to using a mobile browser (with a conventional questionnaire) and positively influences willingness to continue participating in the panel.

App-based data collection platforms offer a range of potential benefits to researchers, and combined with a panel design, a potentially cost-efficient one. The present study suggests there may be considerable advantages for respondents also. If it is indeed the case that apps reduce response burden then our findings would imply a broader utility for them as a data collection mode in general population studies, beyond the more 'niche' uses for which they have been deployed in surveys to date (Link et al. 2014). The use of apps in panel designs offers other advantages also, including the possibility to make use of the new measurement tools they offer and the opportunity to build study identity and loyalty through a purpose-designed interface and infrastructure for panel management (Link et al. 2014). More importantly, however, if apps offer more effective mobile optimisation solutions than browsers, they present a better alternative for reaching the growing number of mobile-dependent Internet users and survey respondents.

Benefits of responding via an app could be emphasised to potentially positive effect in efforts to motivate willingness to participate in app-based studies, which has been shown elsewhere to present an important barrier (so far) to the successful integration of apps in survey research (e.g., Keusch et al. 2019; Mulder and de Bruijne 2019; Revilla et al. 2019; Wenz et al. 2019). Similarly, the finding that being motivated to contribute to science improves both the survey experience and respondent engagement may offer further clues as to how to optimise the framing of invitations to participate in app-based surveys. Given the importance of motivation, combined with perceptions and prior experiences of burden in participation decisions, new research should test ways of making these benefits more salient in survey requests and monitoring and evaluating both in the future development of app-based survey methodology.

7. References

Allum, N., F. Conrad, and A. Wenz. 2018. "Consequences of Mid-Stream Mode-Switching in a Panel Survey." *Survey Research Methods* 12: 43–58. DOI: https://doi.org/10.18148/srm/2018.v12i1.6779.

Antoun, C. 2015. Who are the Internet Users, Mobile Internet Users, and Mobile-Mostly Internet Users?: Demographic Differences Across Internet-Use Subgroups in the U.S." In *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies* by D. Toninelli, R. Pinter, and P. Pedraza: 99–117. Ubiquity Press. DOI: http://dx.doi.org/10.5334/bar.g.

Antoun, C., and A. Cernat. 2020. "Factors Affecting Completion Times: A Comparative Analysis of Smartphone and PC Web Surveys." *Social Science Computer Review* 38: 477–489. DOI: https://doi.org/10.1177/0894439318823703.

Antoun, C., F.G. Conrad, M.P. Couper, and T.B. West. 2019. "Simultaneous Estimation of Multiple Sources of Error in a Smartphone-Based Survey." *Journal of Survey Statistics and Methodology* 7: 93–117. DOI: https://doi.org/10.1093/jssam/smy002.

Antoun, C., M.P. Couper, and F.G. Conrad. 2017. "Effects of Mobile versus PC Web on Survey Response Quality." *Public Opinion Quarterly* 81: 280–306. DOI: https://doi.org/10.1093/poq/nfw088.

Antoun, C., J. Katz, J. Argueta, and L. Wang. 2018. "Design Heuristics for Effective Smartphone Questionnaires." *Social Science Computer Review* 36: 557–574. DOI: https://doi.org/10.1177/0894439317727072.

Baron, R.M. and D.A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51: 1173–1182. DOI: https://doi.org/10.1037/0022-3514.51.6.1173.

Bosnjak, M., G. Metzger, and L. Gräf. 2010. "Understanding the Willingness to Participate in Mobile Surveys: Exploring the Role of Utilitarian, Affective, Hedonic, Social, Self-Expressive, and Trust-Related Factors." *Social Science Computer Review* 28: 350–370. DOI: https://doi.org/10.1177/0894439309353395.

Bradburn, N. 1978. *Respondent burden.* Health Survey Research Methods, DHEW Publication No. 79-3207: 35–40. Washington, DC: U.S. Department of Health, Education, and Welfare. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf (accessed March 2022).

Buskirk, T.D., and C. Andrus. 2012. "Online Surveys Aren't Just for Computers Anymore! Exploring Potential Mode Effects between Smartphone vs. Computer-based Online Surveys". In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association: 5678–5691. Available at: http://www.asasrms.org/Proceedings/y2012f.html (accessed March 2022).

Buskirk, T.D. and C. Andrus. 2014. "Making Mobile Browser Surveys Smarter: Results from a Randomized Experiment Comparing Online Surveys Completed via Computer or Smartphone." *Field Methods* 26: 322–342. DOI: https://doi.org/10.1177/1525822X14526146.

Callegaro, M. 2010. "Do You Know Which Device Your Respondent Has Used to Take Your Online Survey?" *Survey Practice* 3: 1–12. DOI: https://doi.org/10.29115/sp-2010-0028.

Callegaro, M. 2013. "From Mixed-Mode to Multiple Devices: Web Surveys, Smartphone Surveys and Apps: Has the Respondent Gone Ahead of Us in Answering Surveys?", *International Journal of Market Research* 55: 317–320. DOI: https://doi.org/10.2501/IJMR-2013-026.

Callegaro, M., K. Manfreda, and V. Vehovar. 2015. *Web Survey Methodology*. London, UK: SAGE.

Couper, M.P., C. Antoun, and A. Mavletova. 2017. "Mobile Web Surveys." In *Total Survey Error in Practice* edited by P.P. Biemer, E.D. de Leeuw, S. Eckman, B.

Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West: 133–154. Hoboken, NJ: Wiley. DOI: https://doi.org/10.1002/9781119041702.ch7.

Couper, M.P. and G.J. Peterson. 2017. "Why Do Web Surveys Take Longer on Smartphones?" *Social Science Computer Review* 35: 357–377. DOI: https://doi.org/10.1177/0894439316629932.

De Bruijne, M. and A. Wijnant. 2014a. "Mobile Response in Web Panels." *Social Science Computer Review* 32: 728–742. DOI: https://doi.org/10.1177/0894439314525918.

De Bruijne, M. and A. Wijnant, A. 2014b. "Improving Response Rates and Questionnaire Design for Mobile Web Surveys." *Public Opinion Quarterly* 78: 951–962. DOI: https://doi.org/10.1093/poq/nfu046.

De Leeuw, E.D. and P. Lugtig. 2014. *Dropouts in Longitudinal Surveys.* Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd. DOI: 10.1002/9781118445112.stat06661.pub2.

Dillman, D.A., M.D. Sinclair, and J.R. Clark. 1993. "Effects of Questionnaire Length, Respondent-friendly Design, and a Difficult Question on Response Rates for Occupant-addressed Census Mail Surveys." *Public Opinion Quarterly* 57: 289-304. DOI: 10.1086/269376.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Mail, and Mixed-mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons.

Elevelt, A., P. Lugtig, and V. Toepoel. 2019. "Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process?" *Survey Research Methods* 13: 195–213. DOI: https://doi.org/10.18148/srm/2019.v13i2.7385.

Freese, J. and A. Branigan. 2012. "Cognitive Skills and Survey Nonresponse — Evidence from Two Longitudinal Studies in the United States." *EurAmerica* 42: 221–247. Available at: https://www.ea.sinica.edu.tw/eu_file/134009378914.pdf.

Galesic, M. 2006. "Dropouts on the Web: Effects of Interest and Burden Experienced during an Online Survey." *Journal of Official Statistics* 22: 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf

Galesic, M., and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73: 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Geisen, E., and J.R. Bergstrom. 2017. *Usability Testing for Survey Research*. Burlington, MA: Morgan Kaufmann.

Groves, R., S. Presser, and S. Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly* 68: 2–31. DOI: https://doi.org/10.1093/poq/nfh002.

Guidry, K.R. 2012. "Response Quality and Demographic Characteristics of Respondents Using a Mobile Device on a Web-Based Survey?" Presentation at the American Association for Public Opinion Research's 67th Annual Conference. May 18, Orlando, FL, U.S.A. Available at: https://www.aapor.org/AAPOR_Main/media/AnnualMeeting-Proceedings/2012/01_KRGuidry_F1_Mobile-Device-Respondents.pdf (accessed March 2022).

Gummer, T., and J. Daikeler. 2020. "A Note on How Prior Survey Experience with Self-Administered Panel Surveys Affects Attrition in Different Modes." *Social Science Computer Review* 38: 490–498. DOI: https://doi.org/10.1177/0894439318816986.

Haraldsen, G. 2004. "Searching for Response Burdens in Focus Groups with Business Respondents." In Proceedings of the QUEST 2003: Proceedings of the 4th Conference on Questionnaire Evaluation Standards, 21-23 October 2003, edited by P. Prüfer, M. Rexroth, and F.J.J. Fowler: 13–123. Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. Available at: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49200-6 (accessed August 2021).

Harder, V.S., E.A. Stuart, and J.C. Anthony. 2010. "Propensity Score Techniques and the Assessment of Measured Covariate Balance to Test Causal Associations in Psychological Research." *Psychological Methods* 15: 234–249. DOI: https://doi.org/10.1037/a0019623.

Hayes, A.F. 2017. *Introduction to Mediation, Moderation, and Conditional Process Analysis.* New York, NY: The Guilford Press.

Herzing, J.M.E. 2019. *Mobile Web Surveys*. FORS Guide No. 01, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. DOI: https://doi.org/10.24449/FG-2019-00001.

Hill, D.H. and R.J. Willis. 2001. "Reducing Panel Attrition. A Search for Effective Policy Instruments." *Journal of Human Resources* 26: 416–438. DOI: https://doi.org/10.2307/3069625.

Hoogendoorn, A.W., and D. Sikkel. 1998. "Response Burden and Panel Attrition." *Journal of Official Statistics* 14: 189–205. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/response-burden-and-panel-attrition.pdf

Hox, J., E.D. de Leeuw, and T. Klausch. 2017. "Mixed Mode Research: Issues in Design and Analysis." In *Total survey error in practice* edited by P.P. Biemer, E.D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West: 511–530. Hoboken, NJ: Wiley. DOI: https://doi.org/10.1002/9781119041702.ch23.

Jäckle, A., J. Burton, M.P. Couper, and C. Lessof. 2019. "Participation in a Mobile App Survey to Collect Expenditure Data as Part of a Large-Scale Probability Household Panel: Coverage and Participation Rates and Biases." *Survey Research Methods* 13: 23–44. DOI: https://doi.org/10.18148/srm/2019.v1i1.7297

Jäckle, A., A. Gaia, and M. Benzeval. 2018. *The Use of New Technologies to Measure Socioeconomic and Environmental Concepts in Longitudinal Studies*. CLOSER Resource report. London, UK: UCL, Closer, Institute of Education. Available at: https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-The-use-of-new-technology-to-measure-socio-economic-and-environmental-concepts.pdf (accessed March 2022).

Jacobsen, J. and S. Kühne. 2021. Using a Mobile App When Surveying Highly Mobile Populations: Panel Attrition, Consent, and Interviewer Effects in a Survey of Refugees. *Social Science Computer Review* 39: 721–743. DOI: https://doi.org/10.1177/0894439320985250.

Johnson, A., F. Kelly, and S. Stevens. 2012. "Modular Survey Design for Mobile Devices." Presentation at the CASRO Online Conference 2012, March 7–8, Las Vegas, NE, U.S.A. Available at: https://c.ymcdn.com/sites/www.casro.org/resource/collectio-

n/E270CC91-6B72-4C37-BCC0-5503/CBB66C55/Paper_-_Frank_Kelly_and_Alex_-Johnson_-_Lightspeed_Research_and_Kantar_Operations.pdf.

Kalton, G., J. Lepkowski, G.E. Montanari, and D. Maligalig. 1990. "Characteristics of Second Wave Nonrespondents in a Panel Survey." In Proceedings of the Survey Research Methods Section: American Statistical Association: 462–467. In *JSM Proceedings,* Survery Research Methods Section. Alexandria, VA, U.S.A. Available at: http://www.asasrms.org/Proceedings/y1990f.html (accessed August 2021).

Keusch, F., B. Struminskaya, C. Antoun, M.P. Couper, and F. Kreuter. 2019. "Willingness to Participate in Passive Mobile Data Collection." *Public Opinion Quarterly* 83: 210–235. DOI: https://doi.org/10.1093/poq/nfz007.

Keusch, F. and T. Yan. 2017. "Web versus Mobile Web: An Experimental Study of Device Effects and Self-Selection Effects." *Social Science Computer Review* 35: 751–769. DOI: https://doi.org/10.1177/0894439316675566.

Klausch, T., J. Hox, and B. Schouten. 2015. "Selection Error in Single- and Mixed Mode Surveys of the Dutch General Population." *Journal of the Royal Statistical Society Series A* 178: 945–961. DOI: https://doi.org/10.1111/rssa.12102.

Kleinert, C., B. Christoph, and M. Ruland. 2019. "Experimental Evidence on Immediate and Long-term Consequences of Test-induced Respondent Burden for Panel Attrition." *Sociological Methods and Research*: 1–32. DOI: https://doi.org/10.1177/0049124119826145.

Kreuter, F., G.-C. Haas, F. Keusch, S. Bähr, and M. Trappmann. 2018. "Collecting Survey and Smartphone Sensor Data with an App: Opportunities and Challenges Around Privacy and Informed Consent." *Social Science Computer Review* 38: 533–549. DOI: https://doi.org/10.1177/0894439318816389.

Laurie, H. 2008. "Handbook of Longitudinal Research: Design, Measurement, and Analysis." In *Minimizing Panel Attrition* edited by S. Menard: 167–184. Elsevier.

Laurie, H., R. Smith, and L. Scott. 1999. ''Strategies for Reducing Nonresponse in a Longitudinal Panel Survey.'' *Journal of Official Statistics* 15: 269–282.

Lawes, M., C. Hetschko, J.W. Sakshaug, and S. Griessemer. 2021. "Contact Modes and Participation in App-based Smartphone Surveys: Evidence from a Large-scale Experiment." *Social Science Computer Review*. DOI: https://doi.org/10.1177/0894439321993832.

Lee, H., S. Kim, M. Couper, and Y. Woo. 2019. "Experimental Comparison of PC Web, Smartphone Web, and Telephone Surveys in the New Technology Era." *Social Science Computer Review* 37: 234–247. DOI: https://doi.org/10.1177/0894439318756867.

Lemay, M. 2010. ''Understanding the Mechanism of Panel Attrition.'' Unpublished Doctoral thesis, Doctor of Philosophy, University of Maryland, College Park, MD.

Lepkowski, J.M. and M.P. Couper. 2002. "Nonresponse in the Second Wave of Longitudinal Household Surveys." In *Survey Nonresponse* edited by R.M. Groves, D.A. Dillman, J.E. Eltinge, and R.J.A. Little: 259–272. New York: Wiley.

Link, M.W., J. Murphy, M.F. Schober, T.D. Buskirk, J. Hunter Childs, and C. Langer Tesfaye. 2014. "Mobile Technologies for Conducting, Augmenting and Potentially Replacing Surveys." *Public Opinion Quarterly* 78: 779–787. DOI: https://doi.org/10.1093/poq/nfu054.

Lipps, O. 2009. *Attrition of Households and Individuals in Panel Surveys.* SOEP- Papers 164. Berlin: DIW. Available at: http://www.diw.de/documents/publikationen/73/diw_01.c.96125.de/diw_sp0164.pdf (accessed March 2022).

Loosveldt, G. and A. Carton. 2001. "An Empirical Test of a Limited Model for Panel Refusals." *International Journal of Public Opinion Research* 13:173–185. DOI: https://doi.org/10.1093/ijpor/13.2.173

Loosveldt, G., J. Pickery, and J. Billiet. 2002. "Item Nonresponse as a Predictor of Unit Nonresponse in a Panel Survey." *Journal of Official Statistics* 18: 545–557. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/item-nonresponse-as-a-predictor-of-unit-nonresponse-in-a-panel-survey.pdf.

Lugtig, P. 2014. "Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers." *Sociological Methods & Research* 43: 699–723. DOI: https://doi.org/10.1177/0049124113520305

Lugtig, P. 2021. "What Panel Surveys and Smartphone-App Studies can Learn from Each Other." Presentation at the 9th Conference of the European Survey Research Association, Online, July 9. Available at: https://www.europeansurveyresearch.org/conferences/program2021?sess = 29#218 (accessed March 2022).

Lugtig, P. and V. Toepoel. 2016. "The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error." *Social Science Computer Review*, 34: 78–94. DOI: https://doi.org/10.1177/0894439315574248.

Lynn, P. 2008. "The Problem of Nonresponse." In *International Handbook of Survey Methodology* edited by E.D. de Leeuw, J.J. Hox, and D.A. Dillman: 35–55. New York: Psychology Press.

Lynn, P. 2014. "Longer Interviews May Not Affect Subsequent Survey Participation Propensity." *Public Opinion Quarterly* 78: 500–509. DOI: http://dx.doi.org/10.1093/poq/nfu015.

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schütz. 2007. "Compensating for Low Topic Interest and Long Surveys. A Field Experiment on Nonresponse in Web Surveys." *Social Science Computer Review* 25: 372–383. DOI: https://doi.org/10.1177/0894439307297606.

Maslovskaya, O., G.B. Durrant, P.W.F. Smith, T. Hanson, and A. Villar. 2019. "What are the Characteristics of Respondents Using Different Devices in Mixed-Device Online Surveys? Evidence from Six UK Surveys." *International Statistical Review* 87: 326–346. DOI: https://doi.org/10.1111/insr.12311.

Mavletova, A. 2013. "Data Quality in PC and Mobile Web Surveys." *Social Science Computer Review* 31: 725–743. DOI: https://doi.org/10.1177/0894439313485201.

Mavletova, A. and M.P. Couper. 2013. "Sensitive Topics in PC Web and Mobile Web Surveys: Is there a Difference?" *Survey Research Methods* 7: 191–205. DOI: https://doi.org/10.18148/srm/2013.v7i3.5458.

Mavletova, A. and M.P. Couper. 2015. "A Meta-Analysis of Breakoff Rates." In *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies* edited by D. Toninelli, R. Pinter, and P. Pedraza: 81–98. Ubiquity Press. DOI: https://doi.org/10.5334/bar.f.

Mavletova, A., M.P. Couper, and D. Lebedev. 2018. "Grid and Item-by-Item Formats in PC and Mobile Web Surveys." *Social Science Computer Review* 36: 647–668. DOI: https://doi.org/10.1177/0894439317735307.

Mulder, J. and M. de Bruijne. 2019. "Willingness of Online Respondents to Participate in Alternative Modes of Data Collection." *Survey Practice* 12: 1–11. DOI: https://doi.org/10.29115/SP-2019-0001.

Office of Management and Budget. 2006. *Standards and Guidelines for Statistical Surveys*. Executive Office of the President of the United States. Available at: https://unstats.un.org/unsd/dnss/docs-nqaf/USA_standards_stat_surveys.pdf (accessed August 2021).

Olsen, R.J. 2005. "The Problem of Respondent Attrition: Survey Methodology is the Key." *Monthly Labor Review* 128: 63–70. Available at: http://www.jstor.org/stable/23804052 (accessed March 2022).

Peytchev, A. 2009. "Survey Breakoff." *Public Opinion Quarterly* 73: 74–97. DOI: https://doi.org/10.1093/poq/nfp014.

Peytchev, A. and C.A. Hill. 2010. "Experiments in Mobile Web Survey Design: Similarities to Other Modes and Unique Considerations." *Social Science Computer Review* 28: 319–335. DOI: https://doi.org/10.1177/0894439309353037.

Pinter, R. 2015. "Willingness of Online Access Panel Members to Participate in Smartphone Application-Based Research." In *Mobile Research Methods*, edited by D. Toninelli, R. Pinter, and P. de Pedraza: 141–156. London: Ubiquity Press.

Read, B. 2019. "Respondent Burden in a Mobile App: Evidence from a Shopping Receipt Scanning Study." *Survey Research Methods* 13: 45–71. DOI: https://doi.org/10.18148/srm/2019.v1i1.7379.

Revilla, M., M.P. Couper, and C. Ochoa. 2019. "Willingness of Online Panelists to Perform Additional Tasks." *Methods, Data, Analyses* 13: 223–252. DOI: https://doi.org/10.12758/mda.2018.01.

Revilla, M. and C. Ochoa. 2015. "What are the Links in a Web Survey among Response Time, Quality, and Auto-Evaluation of the Efforts Done?" *Social Science Computer Review* 33: 97–114. DOI: https://doi.org/10.1177/0894439314531214.

Revilla, M., C. Ochoa, and G. Loewe. 2017. "Using Passive Data from a Meter to Complement Survey Data in Order to Study Online Behavior." *Social Science Computer Review* 35: 521–536. DOI: https://doi.org/10.1177/0894439316638457.

Revilla, M., D. Toninelli, and C. Ochoa. 2016. "PCs versus Smartphones in Answering Web Surveys: Does the Device Make a Difference?" *Survey Practice* 9: 1–6. DOI: https://doi.org/10.29115/sp-2016-0021.

Roberts, C., Vandenplas, C., and J. Herzing. 2020. "A Validation of R-indicators as a Measure of the Risk of Bias using Data from a Non-response Follow-up Survey." *Journal of Official Statistics* 36: 675-701. DOI: https://doi.org/10.2478/jos-2020-0034.

Rosenbaum, P.R., and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. DOI: https://doi.org/10.1093/biomet/70.1.41.

Sakshaug, J.W., and F. Kreuter. 2011. "Using Paradata and Other Auxiliary Data to Examine Mode Switch Nonresponse in a "Recruit-and-Switch" Telephone Survey." *Journal of Official Statistics* 27: 339–357. Available at: https://www.scb.se/conten-

tassets/ca21efb41fee47d293bbee5bf7be7fb3/using-paradata-and-other-auxiliary-data-to-examine-mode-switch-nonresponse-in-a-34recruit-and-switch34-telephone-survey.pdf.

Sakshaug, J.W., T. Yan, and R. Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items." *Public Opinion Quarterly* 74: 907–933. DOI: https://doi.org/10.1093/poq/nfq057.

Scherpenzeel, A. 2017. "Mixing Online Panel Data Collection with Innovative Methods." In *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* [Methodological Problems of Mixed-Mode Approaches in Survey Research], edited by S. Eifler and F. Faulbaum: 27–49. Wiesbaden: Springer.

Stapleton, C.E. 2013. "The Smart(phone) Way to Collect Survey Data." *Survey Practice* 6: 1–7. DOI: https://doi.org/10.29115/sp-2013-0011.

Struminskaya, B., K. Weyandt, and M. Bosnjak. 2015. "The Effects of Questionnaire Completion using Mobile Devices on Data Quality. Evidence from a Probability-Based General Population Panel." *Methods, Data, Analyses* 9: 261–292. DOI: https://doi.org/10.12758/mda.2015.014.

Struminskaya, B., Toepoel, V., Lutgtig, P., Haan, M., Luiten, A., and B. Schouten. 2021. "Understanding willingness to share smartphone-sensor data." *Public Opinion Quarterly.* DOI: https://doi.org/10.1093/poq/nfaa044.

Toepoel, V. and P. Lugtig. 2015. "Online Surveys are Mixed-Device Surveys. Issues Associated with the Use of Different (Mobile) Devices in Web Surveys." *Methods, Data, Analyses* 9: 155–162. DOI: https://doi.org/10.12758/mda.2015.009.

Toepoel, V. and P. Lugtig. 2018. "Modularization in an Era of Mobile Web: Investigating the Effects of Cutting a Survey into Smaller Pieces on Data Quality." *Social Science Computer Review*. DOI: https://doi.org/10.1177/0894439318784882.

Tourangeau, R., F.G. Conrad, and M.P. Couper. 2013. *The Science of Web Surveys*. Oxford, GB: Oxford University Press.

Tresch, A., L. Lauener, L. Bernhard, and L. Scaperrotta. 2020. *Selects: Panel Survey* (waves 1–3) – 2019 [Dataset]. Distributed by FORS, Lausanne, 2020. www.selects.ch. DOI: https://doi.org/10.23662/FORS-DS-1184-1.

VanderWeele, T.J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford, UK: Oxford University Press.

VanderWeele, T.J. 2016. "Mediation Analysis: A Practitioner's Guide." *Annual Review of Public Health* 37: 17–32. DOI: https://doi.org/0.1146/annurev-publhealth-032315-021402.

Watson, N. and M. Wooden. 2009. ''Identifying Factors Affecting Longitudinal Survey Response.'' In *Methodology of Longitudinal Surveys* edited by P. Lynn: 157–182. Chichester, England: John Wiley.

Wells, T., J.T. Bailey, and M.W Link. 2013. "Filling the Void: Gaining a Better Understanding of Tablet-Based Surveys." *Survey Practice* 6: 1–9. DOI: https://doi.org/10.29115/sp-2013-0002.

Wenz, A. 2021. "Do Distractions During Web Survey Completion Affect Data Quality? Findings from a Laboratory Experiment." *Social Science Computer Review* 3: 148–161. DOI: https://doi.org/10.1177/0894439319851503.

Wenz, A., A. Jäckle, and M.P. Couper. 2019. "Willingness to Use Mobile Technologies for Data Collection in a Probability Household Panel." *Survey Research Methods* 13: 1–22. DOI: https://doi.org/10.18148/srm/2019.v13i1.7298.

Yan, T., S. Fricker, S., and S. Tsai. 2019. "Response Burden: What Predicts It and Who is Burdened Out?" In *Advances in Questionnaire Design, Development, Evaluation and Testing* edited by P. Beatty, D. Collins, L. Kaye, J.L. Padilla, G. Willis, and A. Wilmot, 193-212. Hoboken, NJ: John Wiley & Sons. DOI: https://doi.org/10.1002/9781119263685.ch8.

Received December 2020
Revised August 2021
Accepted March 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1019–1050, http://dx.doi.org/10.2478/JOS-2022-0044

The Effect of Burdensome Survey Questions on Data Quality in an Omnibus Survey

Angelica Phillips[1] *and Rachel Stenger*[2]

In interviewer-administered omnibus surveys, burdensome questions asked early in a survey may result in lower quality responses to questions asked later in a survey. Two examples of these burdensome questions are social network questions, wherein respondents are asked about members of their personal network, and knowledge questions, wherein respondents are asked to provide a factually correct response to a question. In this study, we explore how the presence of potentially burdensome questions are associated with item nonresponse and acquiescence rates on subsequent survey questions, and whether this effect differs by respondent age and education. We use data from the 2010 General Social Survey (AAPOR RR5 = 70.3%, AAPOR 2016), which experimentally varied the location of a social network module and the presence of a knowledge question module. Those who received knowledge questions had higher item nonresponse rates on subsequent questions than those who did not receive knowledge questions, but the quality of responses did not differ by the presence of social network questions. Further, respondents with different characteristics were not differentially burdened by the knowledge questions or the social network questions. We conclude that knowledge questions may be better asked near the end of omnibus surveys to preserve the response quality for subsequent questions.

Key words: Questionnaire design; respondent burden; interviewer-administered surveys; social network questions; knowledge questions.

1. Introduction

Some survey questions are more cognitively burdensome for respondents to answer than other questions. Whether to place highly burdensome questions early or later in the interview is an important questionnaire design decision because burdensome questions may lower the quality of responses to subsequent items (Dillman et al. 2014; Galesic 2006). This research decision is especially important for omnibus surveys. Omnibus surveys contain questions on many different topics, generally with multiple clients purchasing space for their questions to be included in the survey (AmeriSpeak 2020); one example of this type of survey is the US nationally representative General Social Survey (GSS). In the case of an omnibus survey such as the GSS, clients who

[1] University of Nebraska-Lincoln, 734 Oldfather Hall, Lincoln, NE 68588-0342, U.S.A. Email: Angelica.Phillips@huskers.unl.edu
[2] RTI International, 3040 Cornwallis Rd, Durham, NC 27709, U.S.A. Email: rstenger@rti.org
Acknowledgments: We would like to thank Kristen Olson and Jolene Smyth for their thoughtful comments and feedback on earlier drafts of this article. An earlier version of this article was presented at the 2020 Midwest Association for Public Opinion Research and 2021 American Association for Public Opinion Research annual meetings.

purchased space for their questions to be included in the GSS would reasonably expect that the quality of responses to their questions would be unaffected by other questions in the GSS. However, respondents provide lower quality responses as their cognitive burden increases (Galesic 2006; Galesic and Bosnjak 2009), and as such, their responses might be affected by the burden experienced from previous survey question modules.

Burdensome questions come in many forms. One example of burdensome questions is those that ask for details about members of a respondent's personal network, known as social network questions. Previous research has explored how social network questions influence respondent burden (Manfreda et al. 2004) and how respondent burden, in turn, influences response quality (Galesic 2006; Galesic and Bosnjak 2009). Yet to our knowledge, whether social network questions can influence the quality of answers to subsequent non-social network questions in an omnibus survey is unexamined. Another example of highly burdensome survey questions are knowledge questions, which ask respondents to answer factual questions that have a known correct response, such as in a quiz or test (Sturgis et al. 2008). As with social network questions, little to no research has examined how the presence of knowledge questions may influence the quality of answers to subsequent unrelated survey questions. Further, no research has explored whether knowledge questions and social network questions are differentially burdensome for certain groups of respondents. Respondents of differing levels of cognitive capacity (i.e., those who are older and those with lower levels of formal education) might be more burdened by social network questions and knowledge questions compared to respondents with higher levels of cognitive capacity (Holbrook et al. 2007; Krosnick 1991; Olson et al. 2018). As respondents experience more burden by certain survey questions, they are more likely to exhibit satisficing behaviors and provide less than optimal responses (Krosnick 1991). Further, respondents may tend to exhibit satisficing through certain behaviors over others, developing a response style that they use throughout the survey interview. While satisficing may take the form of many respondent behaviors, this article will focus on two indicators of satisficing: (1) item nonresponse, wherein the respondent reports "don't know," another hard-to-code response, or refuses to respond instead of providing a substantive response option, and (2) acquiescence, wherein the respondent tends to agree with opinion questions.

In this study, we examine whether the presence of a series of highly burdensome survey questions, with a particular focus on social network and knowledge questions, is associated with the quality of responses (i.e., item nonresponse and acquiescence) to subsequent survey questions with two research questions: (1) What is the effect of burdensome survey questions (i.e., social network questions and knowledge questions) on the quality of responses to subsequent unrelated survey questions? (2) Are respondents with different characteristics such as cognitive capacity and response styles differentially burdened by these cognitively taxing survey question modules? To address these questions, we use panel data from the 2010 GSS (Smith et al. 2019; AAPOR RR5 = 70.3%, AAPOR 2016), which experimentally varied the question ordering of the social network module and the presence or absence of the knowledge question module.

2. Background

2.1. Burdensome Survey Questions

2.1.1. Respondent Cognitive Burden

Respondent cognitive burden, which is the strain respondents feel during a survey due to difficulty replying to survey questions (Bradburn 1978), can affect the quality of provided responses (Galesic 2006; Galesic and Bosnjak 2009; Krosnick 1991). In response to increased cognitive burden, respondents may choose to reduce their cognitive effort through satisficing, which is the act of providing a merely satisfactory response rather than expending the effort to provide a high-quality optimal response (Krosnick 1991). These cognitive shortcuts can result in lower response accuracy, which can in turn affect the conclusions made from the survey (Hamby and Taylor 2016; Lelkes et al. 2012).

Bradburn (1978) posits four specific aspects of a survey that can increase respondent cognitive burden: the length of the interview, the amount of effort required of the respondent to answer a question, the amount of stress felt by the respondent because of the interview, and the frequency with which the respondent is interviewed. Therefore, if a survey is particularly long or if the amount of effort that is required for the respondent to complete the survey questions is great, respondent satisficing may increase due to increased respondent cognitive burden (Bradburn 1978; Galesic 2006; Krosnick 1991). Certain question types may increase respondent burden at a higher rate compared to other survey questions due to the inherently complex nature of these questions. Furthermore, if cognitive burden can cumulate over the length of a survey, it may be that respondents use satisficing behaviors more often after burdensome survey questions even if subsequent questions are not necessarily burdensome. To examine the first research question, which asks what the effect of burdensome survey questions is on the quality of responses to subsequent unrelated survey questions, we focus on two potentially burdensome types of survey questions – social network questions and knowledge questions.

2.1.2. Social Network Guestions

There are generally two types of social network data (Marsden 2016): whole-network studies and egocentric social network questions. Whole-network studies collect social network data on a bounded group, such as a school or a workplace, while egocentric networks focus on the social connections of a focal individual. In the case of egocentric social network questions, respondents are tasked with both a lengthy and a highly complex series of questions (Manfreda et al. 2004; McCarty et al. 2007; Perry et al. 2018). There are two components to egocentric social network questions: a name generator section and the name interpreter section (Perry et al. 2018). In the name generator section, respondents are asked to list members of their social network, called alters. In the name interpreter section, the respondent is asked details about each of the alters listed in the name generator section. This can include questions about the characteristics of the alters, the types and strength of the relationships with the alters, and the ties between the alters to get a sense of the structure of the social network.

Egocentric social network questions can be time consuming and therefore cognitively burdensome for respondents to answer because the number of follow-up questions asked

about each alter in the name interpreter section of the survey depends on the number of alters listed in the name generator section (Manfreda et al. 2004; McCarty et al. 2007; Perry et al. 2018). The combination of the long length and the complexity of egocentric social network questions can substantially increase respondent cognitive burden, which, in turn, can decrease data quality (Manfreda et al. 2004; Kogovšek 2006; Kogovšek and Ferligoj 2005; Paik and Sanchagrin 2013; Valente et al. 2017).

2.1.3. Knowledge Guestions

Another form of potentially burdensome survey questions are knowledge questions, also known as factual information questions, which require respondents to select or provide a factually correct response to a question that aims to test their knowledge on certain substantive topics (Nadeau and Niemi 1995; Zaller 1992). These questions have appeared in nationally representative surveys such as the American National Election Studies, which has regularly asked political knowledge questions since 1958 (American National Elections Studies 2020). Because knowledge questions ask respondents to provide a factually correct response to a subject-specific question, these questions test the respondent's knowledge about substantive topics such as US politics, resulting in a potentially stressful survey task which may in turn contribute to cognitive burden (Bradburn 1978).

Respondents can experience burden from knowledge questions for two reasons; respondents are aware that there is a correct response that they are evaluated on, which can contribute stress, and if respondents are aware of any poor performance on these questions, they may feel frustrated (Kleinert et al. 2019). However, while there has been extensive research on how to design knowledge questions (e.g., Mondak 1999; Mondak 2001; Sturgis et al. 2008), less attention has been paid to the degree of burden experienced by respondents when encountering these questions. Furthermore, little to no extant research examines how the presence of knowledge questions, as a potential type of highly burdensome series of questions, may influence the quality of responses to subsequent unrelated questions on omnibus surveys.

2.1.4. Burdensome Survey Questions and Response Quality

Most extant literature on the placement of egocentric social network questions within an omnibus survey interview focuses on reducing item nonresponse and measurement error within the social network questions themselves (Manfreda et al. 2004; Kogovšek 2006; Kogovšek and Ferligoj 2005; Paik and Sanchagrin 2013; Valente et al. 2017). Scant attention has been given to how social network questions can influence the quality of responses to unrelated questions that appear after the social network module. Additionally, no research to our knowledge has examined the effect of knowledge questions on the quality of responses to subsequent survey questions. In other words, it has not yet been examined whether the cumulative process of cognitive burden can lead to lower quality responses to subsequent unrelated survey questions following burdensome survey questions.

Social network questions (Manfreda et al. 2004) and knowledge questions (Nadeau and Niemi 1995) can potentially be burdensome for respondents and higher respondent cognitive burden is generally associated with lower response quality (Dillman et al. 2014).

As the length of a survey increases respondent burden, and because question complexity increases cognitive burden (Bradburn 1978), we expect the increased complexity of both social network questions and knowledge questions to yield a particularly increased amount of cognitive burden for respondents throughout the course of a survey. It is then likely that this burden will accumulate over the course of the survey, meaning that the responses to survey questions following particularly burdensome survey question modules will then be more likely to be of poorer quality compared to the responses to survey questions that do *not* follow burdensome survey question modules. Therefore, we hypothesize that (H1) respondents who receive burdensome questions (i.e., social network questions or knowledge questions) will provide poorer quality responses (i.e., higher item nonresponse rates and higher rates of acquiescence) to subsequent questions compared to respondents who do not receive either of these sets of burdensome questions.

2.2. *Differential Effects of the Presence of Burdensome Questions Across Respondent Characteristics*

2.2.1. Respondent Characteristics

An additional area of research that has not yet been examined is how the effect of burdensome survey questions on the quality of responses to subsequent unrelated survey questions may vary depending on respondent characteristics. The second research question of this article examines if respondent characteristics moderate the effect of highly burdensome questions on the quality of responses to questions asked after these burdensome questions (i.e., through the form of item nonresponse rates and rates of acquiescence). It is important for survey methodologists to understand not only the effect that the presence of burdensome survey questions, such as social network questions and knowledge questions, have on the quality of responses to subsequent questions in omnibus surveys, but also to understand for which respondent populations this relationship holds. Certain respondent characteristics may influence the amount of cognitive burden experienced and, subsequently, satisficing behaviors exhibited in a survey. To examine this research question, we focus on two types of respondent characteristics – cognitive capacity and response styles.

2.2.2. Cognitive Capacity

Having a lower cognitive capacity, as frequently measured by older age and fewer years of formal education (Holbrook et al. 2007; Krosnick 1991; Olson et al. 2018), may more easily result in a breakdown of the cognitive response process when the respondent encounters complex survey questions (Tourangeau and Rasinski 1988; Tourangeau et al. 2000). This breakdown of the cognitive response process may, in turn, result in satisficing behaviors such as item nonresponse and acquiescence (Krosnick 1991; Narayan and Krosnick 1996). For instance, older respondents provide "don't know" or otherwise unacceptable responses to cognitively complex survey questions more often than younger respondents (Colsher and Wallace 1989; Knäuper et al. 1997; Olson et al. 2018). Similarly, respondents with fewer years of formal education provide more "don't know" responses and are more likely to agree (i.e., exhibit acquiescence) with survey questions as

compared to those with more years of formal education (Holbrook et al. 2003; Krosnick et al. 1996; Narayan and Krosnick 1996; Tu and Liao 2007).

The placement of highly burdensome survey questions such as egocentric social network questions and knowledge questions may differentially affect the quality of answers by respondents who vary in cognitive capacity. More specifically, respondents with a lower cognitive capacity – as indicated by older age and lower educational attainment – may be more likely to experience cognitive burden when answering social network and knowledge questions, which could then influence the quality of their responses to subsequent questions. We anticipate that those with a lower cognitive capacity – as measured by lower educational attainment and older age – will experience greater amounts of cognitive burden and therefore exhibit more satisficing behaviors in general, and that these respondents will also experience more cognitive burden due to the social network questions and the knowledge questions compared to respondents with a greater cognitive capacity. Therefore, we predict that (H2a) older respondents who receive burdensome survey questions (i.e., either the social network questions or the knowledge questions) will exhibit more satisficing behaviors (i.e., higher item nonresponse rates and higher rates of acquiescence) on subsequent questions than those respondents who did not receive either of these burdensome questions. Similarly, (H2b) respondents with lower rates of educational attainment who receive burdensome survey questions will exhibit more satisficing behaviors on subsequent questions than those who did not receive the burdensome questions.

2.2.3. Response Styles

The second type of respondent characteristic examined in this article is response styles. Response styles are systematic patterned ways that respondents respond to survey questions (Paulhus 1991). As a way to reduce the cognitive burden of survey questions, respondents may use cognitive shortcuts in the form of different satisficing indicators including but not limited to acquiescence and item nonresponse (Krosnick 1991). However, one can conceptualize each of the individual satisficing behaviors as response styles, meaning that respondents who satisfice exhibit one particular satisficing behavior more often than another satisficing behavior (Bolt and Johnson 2009; Leeper 2014; Wetzel et al. 2013). To the extent that a satisficing response style is consistent throughout the questionnaire, then response style can be considered a respondent characteristic. Therefore, those who satisfice through the response style of item nonresponse may tend to continue satisficing with this behavior as opposed to other satisficing behaviors when encountering burdensome survey questions.

We predict that (H2c) respondents who exhibit satisficing behavior in the form of higher item nonresponse rates in previous modules of the survey will have higher item nonresponse rates after receiving the burdensome survey question modules compared to those who did not receive the burdensome survey questions. It is likely that respondents who choose a particular form of satisficing – in this case item nonresponse – will continue to satisfice in this chosen method rather than changing the type of satisficing behaviors. Therefore, we also predict (H2d) that these respondents will have *lower* rates of acquiescence following burdensome survey questions compared to those who did not receive the burdensome survey questions. In other words, we predict that respondents who have an established response pattern of higher nonresponse rates as their form of satisficing will respond to the

increase in cognitive burden from the burdensome survey questions by continuing to exhibit higher item nonresponse rates and *not* increased acquiescence rates.

3. Methods

3.1. Data

The data for this article comes from a subset of the 2010 General Social Survey (GSS) (Smith et al. 2019). The GSS is an omnibus survey of non-institutionalized United States adults which is administered every two years and uses a multi-stage area probability sampling design based on addresses. The 2010 GSS includes a cross-section sample of new respondents as well as a rotating panel design. The rotating panel design includes re-interviews from a random sample of those sampled in 2006 and a random sample of those sampled in 2008. The 2010 GSS included a total of 2,044 interviews, with a response rate of 70.3% (AAPOR RR5, AAPOR 2016). Because only those who were recruited into the rotating panel of the GSS in 2006 received the question order experiment examined here, this study is limited to these panel respondents (n = 1,276; retention rate = 83.1%; Smith and Schapiro 2017).

We use listwise deletion to limit our analytical sample to only respondents with non-missing values on all independent and control variables for all analyses predicting item nonresponse rates (n = 1,112). We further limited our analyses on acquiescence rates to those who responded to all eleven questions which were used to calculate acquiescence (n = 1,087) when predicting acquiescence rates; this was done in order to avoid conflation of acquiescence with item nonresponse rates, as respondents must respond to the survey question in order to exhibit acquiescence. All question wording for the variables used in this study can be found in Supplemental Data Table 1 or in the 2017 GSS Panel Interview Codebook (Smith and Schapiro 2017).

3.2. Measures

3.2.1. Ballot Experiments and Burdensome Survey Modules

GSS panel respondents were randomly assigned to one of three versions, called ballots (*ballot* in GSS documentation), of the survey interview; see Table 1 for a list of all modules included in the GSS, their respective topics, and the module order for each ballot.

Table 1. Organization of question modules on the 2010 general social survey.

Ballot A	**Ballot B**	**Ballot C**
Module A Core	Module A Core	Module A Core
Module D Networks/Groups	Module F Aging	Module E Science
Module F Aging	Module I Misc	Module F Aging
Module I Misc.	Module L Shared capital	Module I Misc.
Module L Shared capital	Module D Networks/Groups	Module L Shared capital
Module P Crime/Jail	Module P Crime/Jail	Module D Networks/Groups
Module S Validation/contact	Module S Validation/contact	Module P Crime/Jail
		Module S Validation/contact

The GSS experimentally varied the placement of Module D (Social network questions; SNQ) within the questionnaire. Module D appears in three different locations across the interview; recipients of Ballot A received the social network second out of seven modules, only preceded by the core GSS questions (Module A), recipients of Ballot B received the social network module fifth out of seven modules, and recipients of Ballot C received the social network module sixth out of eight modules. Only those who received Ballot C were asked Module E (Science), which contains the knowledge questions. In the remainder of this article, the experimental versions will be referred to as Ballot A (SNQ), Ballot B (Core), and Ballot C (Knowledge).

Because the focus of this article is to examine the quality of responses to questions asked following burdensome questions and because the same set of questions must be analyzed for all respondents, the focus of analysis will be on the modules that followed the social network questions and/or knowledge questions across all of the ballots. In particular, we focus on Modules F (Aging), I (Misc.) and L (Shared capital). Although Module P is the last substantive module, due to complications with administering this portion of the 2010 GSS, the data for this section are unavailable (Son 2020).

3.2.2. Focal Dependent Variables: Item Nonresponse Rate and Acquiescence Rate Across Questions That Follow Burdensome Modules

The first dependent variable is the item nonresponse rate across Module F (Aging), Module I (Misc.), and Module L (Shared capital). This item nonresponse rate was calculated as the number of items missing (i.e., where the respondent provided a "don't know" or no answer) across the three modules divided by the number of items asked of the respondent in these three modules, taking into account skip patterns and non-applicable items. As the proportion of items missing within each ballot were low (0.017–0.025; Table 2), "don't know" and no answer responses were not analyzed separately. Further, most of the questions asked in Modules F (Aging), I (Misc.), and L (Shared capital) are factual questions, except for the eleven questions in Module F (Aging) to calculate acquiescence (to be discussed next). While answers of "don't know" are considered a substantive response for attitudinal domains (e.g., Krosnick 1991), they are considered item missing, along with no answer, for these analyses. While respondents assigned to Ballot A (SNQ) and Ballot B (Core) did not significantly differ in the item nonresponse rate across Module F (Aging), Module I (Misc.), and Module L (Shared capital) ($t = 0.07$, $p = 0.943$; Table 2), respondents assigned to Ballot C (Knowledge) had significantly higher item nonresponse rates in these modules compared to those assigned to Ballot A ($t = 2.46$, $p = 0.016$; Table 2) and to Ballot B ($t = 2.29$, $p = 0.025$; Table 2).

The acquiescence rate is the second dependent variable for analysis. Acquiescence is measured as the number of questions wherein a respondent reports "agree" or "strongly agree" to the survey item for items containing an agree/disagree scale as response options. Acquiescence can be calculated for eleven survey items, all contained in Module F (Aging) (items included were *immobile, deciding, findepnd, payhlth, withkids, wokids, kidspars, helpkids, helppars, grankids, and sibspars* in GSS documentation). We limited our analyses on acquiescence rates to those who responded to all eleven survey items ($n = 1{,}087$). Respondents assigned to each ballot did not have significantly different acquiescence rates (t-test p-values all $p > .05$; Table 2). See Online Supplemental Data

Table 3 for focal dependent variables across ballot experiments, among respondents who responded to all eleven survey questions used to calculate acquiescence.

3.2.3. Focal Independent Variables: Respondent Characteristics

The first type of respondent characteristic examined is cognitive capacity, measured by age and education. Age and education are often used as predictors of response behaviors, such as satisficing, within survey literature (e.g., Colsher and Wallace 1989; Holbrook et al. 2007; Knäuper et al. 1997; Krosnick 1991; Olson et al. 2018). Age (*age* in GSS documentation) is measured as a continuous variable in the survey (age in years) but was recoded as a three-category ordinal variable (i.e., those aged 18–44, those aged 45–64, and those aged 65 and older) to follow age categories commonly used by the U.S. Census Bureau (U.S. Census Bureau 2011). Age is categorized into these age groups to account for a possible curvilinear association in cognitive capacity over the life course, in which there may be an incline in cognitive capacity up to a certain point, but then a decline in cognitive capacity in older life (typically 65 years and older) (Park 2000). Respondent age did not significantly differ across the three experimental ballots (Design-based $F(3.78, 260.86) = 0.664$, $p = 0.609$; Table 2).

Education (*educ* in GSS documentation) is measured in the survey as years of education completed but was recoded as a binary variable to indicate whether the respondent has a high school education or less (i.e., 12 years or fewer) or more than a high school education (i.e., 13 years or more) (Holbrook et al. 2003; Narayan and Krosnick 1996; Olson et al. 2018). This binary classification of education follows recommendations by Narayan and Krosnick (1996) who find that those with at least some college education exhibit no more weak satisficing behaviors compared to those with a college degree or more. Respondent educational attainment did not significantly differ across experimental ballots (Design-based $F(1.92, 132.75) = 1.964$, $p = 0.146$; Table 2).

The second type of respondent characteristic examined is response style; response style is operationalized as the item nonresponse rate to the core questions (i.e., the first module asked of all respondents; see Table 1). This item nonresponse rate was calculated as the number of items missing (i.e., where the respondent provided a "don't know" or no answer) divided by the number of items asked of the respondent in this module, taking into account skip patterns and non-applicable items. Similar to the calculation of item nonresponse in Modules F (Aging), I (Misc.), and L (Shared capital), answers of "don't know" and no answer were not analyzed separately due to the low item missing rates within each ballot (0.017–0.018; Table 2). Respondents assigned to each ballot did not significantly differ in item nonresponse rates to the core questions (t-test p values all above $p > .05$; Table 2). The descriptive statistics of respondent characteristics by ballot are presented in Table 2. See Supplemental Data Table 3 for respondent characteristics by ballot experiments, among respondents who responded to all eleven survey questions used to calculate acquiescence.

3.2.4. Control Variables

Because the topic of the modules used to calculate the dependent variables of interest (i.e., aging in Module F, miscellaneous topics in Module I, and shared capital in Module L) could differentially affect item nonresponse and acquiescence rates to the questions in

Table 2. Respondent characteristics and data quality indicators by ballot experiment. N = 1,112.

	Ballot A		Ballot B		Ballot C		t-test p value		
	M/percent	SE	M/percent	SE	M/percent	SE	Ballot A v. ballot B	Ballot B v. ballot C	Ballot A v. ballot C
Respondent characteristics									
Age									
18–44	13.95%		14.37%		14.25%				
45–64	13.74%		11.62%		14.58%				
65+	6.53%		5.04%		5.91%				
Design-based F(3.78, 260.86) = 0.664, p = 0.609									
Education									
12 years or less	14.64%		11.74%		16.03%				
13+ years	19.59%		19.29%		18.71%				
Design-based F(1.92, 132.75) = 1.964, p = 0.146									
INR core									
Proportion	0.018	0.001	0.017	0.001	0.018	0.001	0.769	0.564	0.779
Count	3.456	0.156	3.266	0.257	3.657	0.241	0.605	0.218	0.476
Data quality indicators									
INR in Modules F, I, and L									
Proportion	0.020	0.001	0.017	0.002	0.025	0.003	0.943	0.025	0.016
Count	0.872	0.067	0.739	0.111	1.112	0.144	0.993	0.033	0.021
Acquiescence									
Proportion	0.500	0.007	0.498	0.014	0.507	0.013	0.904	0.624	0.536
Count	5.500	0.079	5.476	0.151	5.573	0.138	0.904	0.624	0.536

those modules, additional variables were included in analyses as controls. Respondent sex (*sex* in GSS documentation) was coded as a binary variable indicating whether the interviewer coded the respondent as male (reference category) or female. Marital status (*marital* in GSS documentation) was coded as a binary variable indicating whether the respondent was married or not married (reference category). The respondent's family income (*incom06* in GSS documentation) was coded as an ordinal variable with 25 categories (1 = under USD 1,000; 25 = USD 150,000 or more), with higher values indicating a higher income. Respondent race and ethnicity has been shown to be related to comprehension difficulties (Holbrook et al. 2006) and therefore, race and ethnicity were controlled for in the analytical models. The respondent's race (*race* in GSS documentation) was coded as a nominal variable indicating whether the respondent identified as White (reference category), Black, or other. The respondent's ethnicity (*hispanic* in GSS documentation) was coded as a binary variable indicating whether the respondent identified as Hispanic or not Hispanic (reference category). Note that the race and ethnicity variables were not mutually exclusive, meaning that individuals of all races can be of either Hispanic or non-Hispanic ethnicities. Last, the respondent's political view (*polview* in GSS documentation) was coded as an ordinal variable with seven categories (1 = extremely liberal; 4 = moderate; 7 = extremely conservative), with higher values indicating identifying as more conservative. Descriptive statistics for all variables are included in Table 3. See Online Supplemental Data Table 2 for descriptive statistics for all variables among respondents who responded to all eleven survey questions used to calculate acquiescence.

3.3. Analysis Methods

To test all hypotheses, we estimated weighted ordinary least squares (OLS) regression models accounting for the complex sample design predicting our two outcomes: the combined item nonresponse rate across the focal modules and the acquiescence rate. As a sensitivity analysis, complex sample design-adjusted negative binomial models predicting the number of items missing and Poisson models predicting the number of acquiescent responses in the eleven survey items are reported in Online Supplemental Tables 4–7. Results across the OLS and the count models were substantively similar, so for ease of interpretation, we discuss results from the OLS models.

Interviewers may potentially influence respondent behaviors including nonresponse behaviors (De Leeuw 1992; Groves et al. 2009). In addition, interviewers can affect data collection of social network questions specifically (Kogovšek 2006; Kogovšek and Ferligoj 2005; Paik and Sanchagrin 2013; Valente et al. 2017). In this data structure, interviewers are nested within PSUs; the variance in our two data quality outcomes that is due to interviewers is accounted for by accounting for the sampling clusters in the analyses. All analyses account for the complex sample design of the survey by including the variable *vstrat* as the strata variable, *vpsu* as the cluster variable, and *wtcombnr* as the weight variable. The analyses accounted for the complex design using the *svy* commands in Stata 15.1, and utilized the *singleunit(centered)* option to center single-unit PSUs at the grand mean rather than the stratum mean.

We estimate six models for each dependent variable. Model 1 includes the control variables in addition to measures of respondent characteristics – age group, educational

Table 3. Descriptive statistics (weighted) of all variables included in analyses. N = 1,112.

Variable	M(SE)/Percent
Dependent variables	
Item nonresponse rate in modules F, I, and L	0.020 (0.001)
Count of items missing in modules F, I, and L	0.872 (0.067)
Acquiescence rate	0.500 (0.007)
Count of acquiescence responses	5.500 (0.079)
Respondent characteristics	
Age:	
18–44	42.57%
45–64	39.95%
65+	17.48%
Education: 13+ years (v. 12 years or less)	0.576 (0.012)
Item nonresponse rate in the core questions	0.018 (0.001)
Count of items missing in the core questions	3.456 (0.156)
Experiment	
Ballot A (SNQ)	34.23%
Ballot B (Core)	31.03%
Ballot C (Knowledge)	34.74%
Controls	
Female (v. male)	0.542 (0.019)
Married (v. not married)	0.562 (0.020)
Race	
White	79.27%
Black	10.76%
Other	9.97%
Hispanic (v. not Hispanic)	0.144 (0.023)
Income (min=1; max=25)	17.814 (0.208)
Political views	
Extremely liberal	3.77%
Liberal	12.57%
Slightly liberal	12.99%
Moderate	35.13%
Slightly conservative	14.41%
Conservative	17.33%
Extremely conservative	3.80%

Note: SNQ = Social network questions. The denominator for the item nonresponse rate varies by respondent due to skip patterns. The denominator for the acquiescence rate is 11 for all respondents.

attainment, and previous item nonresponse patterns in the survey – to establish a baseline model predicting our data quality outcomes. Model 2 includes controls, measures of focal respondent characteristics, and the ballot experiment as predictors in order to test H1. More specifically, it is hypothesized that (H1) those who received Ballot A (SNQ) and Ballot C (Knowledge) will have a higher item nonresponse rate and a higher acquiescence rate than those who received Ballot B (Core).

The second research question asks whether respondents with certain characteristics are differentially burdened by the social network questions and the knowledge questions; Models 3–5 explore the differential effects that the ballot experiment may have on those

with differing respondent characteristics. H2a and H2b hypothesize that respondents with a lower cognitive capacity (i.e., older respondents – H2a; and respondents with lower levels of education – H2b) and who receive burdensome survey questions (i.e., Ballot A (SNQ) or Ballot C (Knowledge)) will exhibit more satisficing behaviors (i.e., higher item nonresponse and acquiescence rates) on subsequent questions than those respondents who did not receive either of these burdensome questions. Model 3 adds an interaction term between the ballot experiment and age and an interaction term between the ballot experiment and education. Because there is not a strong theoretical argument about which measure of cognitive capacity (i.e., age or education) is a better indicator of cognitive capacity, both interactions were included in the same model. To test H2a and H2b, Model 3 includes all main effects and added two interaction terms—the ballot experiment by age group and the ballot experiment by education level.

In addition to cognitive capacity, we also predict that respondents with higher item nonresponse rates to the core questions (Module A) will continue to exhibit higher item nonresponse rates as a response style (H2c) when exposed to the social network questions in Ballot A and the knowledge questions in Ballot C. We further predict that these respondents that have higher item nonresponse rates to the core questions (Module A) will exhibit lower acquiescence rates (H2d) when exposed to the social network questions in Ballot A and the knowledge questions in Ballot C. To test these two hypotheses, Model 4 includes all main effects and added an interaction between the ballot experiment and the mean-centered item nonresponse rate in the core questions (Module A). Model 5 includes all main effects and all interactions. See Table 4 for a depiction of how the models, variables, and hypotheses are related. Note that the analytic model building strategies are identical for the two dependent variables (acquiescence and item nonresponse).

We use $\alpha = 0.05$ as our significance level for each hypothesis test, consistent with a priori hypothesis testing. Although each hypothesis states a directional effect and therefore may be analyzed using a one-sided hypothesis test, our analyses use two-sided hypothesis tests to be conservative. We are primarily concerned with the *risk* of burdensome survey questions affecting data quality on subsequent survey questions. Therefore, we are more concerned with minimizing the risk of a type II error than with that of a type I error. Consequently, we do not

Table 4. Description of analytic models, variables for each model, and which hypothesis the model tests.

Model number	Included variables in the model	Which hypothesis does this model test?
1	Controls, respondent characteristics (age, education, previous item nonresponse behaviors)	–
2	Controls, respondent characteristics, ballot experiment	H1
3	Controls, respondent characteristics, ballot experiment, age*ballot interaction, education*ballot interaction	H2a, H2b
4	Controls, respondent characteristics, ballot experiment, item nonresponse behaviors*ballot interaction	H2c, H2d
5	Controls, respondent characteristics, ballot experiment, age*ballot interaction, education*ballot interaction, item nonresponse behaviors*ballot interaction	H2a, H2b, H2c, H2d

Note. H2c focuses on the item nonresponse rate in Modules F, I, and L as the dependent variable and H2d focuses on the acquiescence rate in 11 survey questions as the dependent variable.

adjust our significance level for multiple hypothesis tests although using the Bonferroni-adjusted $\alpha = 0.025$ would account for the comparisons of the two experimental ballots (i.e., Ballot A (SNQ) and Ballot C (Knowledge)) relative to the control ballot (i.e., Ballot B (Core)).

4. Results

Tables 5 and Table 6 show results from the OLS regression models predicting the item nonresponse rate across Modules F, I, and L. Similarly, Tables 7 and Table 8 show results from the OLS regression models predicting acquiescence rates across the eleven agree/disagree items.

Model 1 in Tables 5 and Table 7 includes all control variables plus the focal respondent characteristics (age, education, and previous item nonresponse rates) as predictors. Unlike previous research (Holbrook et al. 2007; Krosnick 1991; Olson et al. 2018), respondents did not have significantly different item nonresponse rates across education groups ($B = -0.004$, $p = 0.115$) nor did they have significantly different acquiescence rates across age groups ($F(2, 68) = 0.03$, $p = 0.968$) or education groups ($B = -0.030$, $p = 0.069$). However, similar to previous research (Holbrook et al. 2007; Krosnick 1991; Olson et al. 2018), respondents did significantly differ in item nonresponse rates by age groups ($F(2, 68) = 4.77$, $p = 0.012$), with those aged 45–64 having higher item nonresponse rates than those 18–44 ($B = 0.006$, $p = 0.048$) and those aged 65 and older having higher item nonresponse rates than those 18–44 ($B = 0.007$, $p = 0.044$). Those in the middle age category and those in the oldest age category did not significantly differ on item nonresponse rates ($F(1, 69) = 0.15$, $p = 0.703$).

Those with a higher item nonresponse rate in the GSS core questions had significantly higher item nonresponse rates on subsequent questions ($B = 0.866$, $p = 0.007$) and significantly lower acquiescence rates ($B = -1.064$, $p = 0.011$). In other words, these results indicate that respondents that have higher item nonresponse rates in the core questions will tend to have higher item nonresponse rates but lower acquiescence rates throughout the survey interview.

4.1. Research Question 1: Effect of Burdensome Questions on the Quality of Responses

Model 2 in Table 5 and Model 2 in Table 7 show the results of adding the ballot experiment as a predictor. The ballot experiment groups collectively were a significant predictor of the item nonresponse rate ($F(2, 68) = 3.78$, $p = 0.028$) but not of the acquiescence rate ($F(2, 68) = 0.04$, $p = 0.964$). More specifically, those who received Ballot C (Knowledge) had significantly higher item nonresponse rates across subsequent questions than those who received Ballot B (Core) ($B = 0.007$, $p = 0.044$) and those who received Ballot A (SNQ) ($F(1, 69) = 7.61$, $p = 0.007$), meaning that the presence of the knowledge question module was associated with higher item nonresponse in subsequent unrelated survey questions. However, the item nonresponse rate was not significantly different between those who received Ballot A (SNQ) and Ballot B (Core) ($B = -0.001$, $p = 0.623$). In sum, there is partial support for H1 as the ballot experiment produced differential item nonresponse rates but no differences in acquiescence rates.

Table 5. Coefficients, Standard Errors, and Confidence Intervals for Survey-Design Adjusted OLS models predicting item nonresponse rates in Sections F (Aging), I (Misc.), and L (Shared Capital), Models 1-2.

	Model 1			Model 2		
	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI
Respondent Characteristics						
Age Group	F(2,68)=4.77	0.012				
Ref=18–44						
45–4	0.006	0.048	0.000,	0.005	0.06	0.000,
	(0.003)		0.011	(0.003)		0.011
65+	0.007	0.044	0.000,	0.007	0.044	0.000,
	(0.004)		0.015	(0.004)		0.015
Education						
Ref=12 years or less						
13+ years	-0.004	0.115	-0.009,	-0.004	0.152	-0.008,
	(0.002)		0.001	(0.002)		0.001
INR in the core questions (centered)	0.866	0.007	0.247,	0.864	0.006	0.251,
	(0.311)		1.486	(0.308)		1.478
Ballot Experiment				F(2,68)=3.78	0.028	
Ref=Ballot B (Core)						
Ballot A (SNQ)				-0.001	0.623	-0.006,
				(0.003)		0.004
Ballot C (Knowledge)				0.007	0.044	0.000,
				(0.003)		0.013
Controls						
Female	-0.004	0.163	-0.010,	-0.004	0.139	-0.010,
	(0.003)		0.002	(0.003)		0.001
Married	0.002	0.514	-0.003,	0.002	0.462	-0.003,
	(0.002)		0.006	(0.002)		0.006
Income	0.000	0.501	-0.001,	0.000	0.446	-0.001,
	(0.000)		0.000	(0.000)		0.000

Table 5. Continued

	Model 1			Model 2		
	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI
Political views	0.000 (0.001)	0.552	-0.002, 0.001	-0.001 (0.001)	0.463	-0.002, 0.001
Race	F(2,68)=2.99	0.057				
Ref=White						
Black	0.01 (0.004)	0.017	0.002, 0.018	0.01 (0.004)	0.019	0.002, 0.017
Other	0.002 (0.004)	0.661	-0.006, 0.009	0.002 (0.004)	0.673	-0.006, 0.009
Hispanic	0.002 (0.005)	0.741	-0.008, 0.012	0.001 (0.005)	0.855	-0.009, 0.011
Intercept	0.023 (0.005)	0.000	0.014, 0.032	0.022 (0.005)	0.000	0.013, 0.031
Observations	1,112			1,112		
R^2	0.231			0.239		
Design-based F	F(11, 59)=7.25	0.000		F(13,57)=6.60	0.000	

Table 6. Coefficients, Standard Errors, and Confidence Intervals for Survey-Design Adjusted OLS models predicting item nonresponse rates in Sections F (Aging), I (Misc.), and L (Shared Capital), Models 3-5.

	Model 3			Model 4			Model 5		
	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI
Respondent Characteristics									
Age group									
Ref=18–44									
45–64	0.006 (0.005)	0.217	-0.004, 0.016	0.004 (0.003)	0.109	-0.001, 0.009	0.005 (0.005)	0.358	-0.005, 0.015
65+	0.001 (0.006)	0.889	-0.011, 0.013	0.008 (0.003)	0.010	0.002, 0.014	0.003 (0.005)	0.541	-0.007, 0.014
Education									
Ref=12 years or less									
13+ years	-0.011 (0.004)	0.011	-0.019, -0.003	-0.005 (0.002)	0.019	-0.009, -0.001	-0.013 (0.004)	0.001	-0.021, -0.005
INR in the core questions (centered)	0.862 (0.305)	0.006	0.254, 1.47	0.606 (0.152)	0.000	0.302, 0.910	0.597 (0.154)	0.000	0.291, 0.904
Ballot Experiment									
Ref=Ballot B (Core)									
Ballot A (SNQ)	-0.009 (0.005)	0.067	-0.018, 0.001	-0.001 (0.003)	0.649	-0.006, 0.004	-0.009 (0.004)	0.058	-0.018, 0.000
Ballot C (Knowledge)	-0.001 (0.005)	0.807	-0.012, 0.009	0.007 (0.003)	0.051	0.000, 0.013	-0.001 (0.006)	0.824	-0.012, 0.010
Interactions									
Ballot * Age	F(4,66)=0.87	0.489					F(4,66)=0.82	0.515	
Ballot A * 45-64	-0.002 (0.005)	0.664	-0.013, 0.008				-0.002 (0.006)	0.681	-0.013, 0.009
Ballot A * 65+	0.01 (0.006)	0.115	-0.002, 0.022				0.01 (0.006)	0.121	-0.003, 0.022
Ballot C * 45-64	0.000 (0.007)	0.953	-0.013, 0.014				0.001 (0.007)	0.930	-0.014, 0.015

Table 6. Continued

	Model 3			Model 4			Model 5		
	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI
Ballot C * 65+	0.010 (0.008)	0.193	-0.005, 0.025				0.005 (0.009)	0.561	-0.012, 0.022
Ballot * Education	F(2,68)=2.50	0.09					F(2,68)=3.14	0.05	
Ballot A * 13+ years	0.011 (0.005)	0.035	0.001, 0.021				0.011 (0.005)	0.019	0.002, 0.020
Ballot C * 13+ years	0.011 (0.005)	0.058	0.000, 0.021				0.011 (0.006)	0.042	0.000, 0.022
Ballot * INR in core questions				F(2,68)=1.82	0.169		F(2,68)=1.88	0.160	
Ballot A * INR in core				-0.174 (0.166)	0.301	-0.506, 0.158	-0.174 (0.170)	0.311	-0.513, 0.166
Ballot C * INR in core				0.773 (0.542)	0.159	-0.309, 1.854	0.780 (0.555)	0.165	-0.327, 1.887
Controls									
Female	-0.004 (0.003)	0.147	-0.010, 0.002	-0.004 (0.002)	0.116	-0.009, 0.001	-0.004 (0.002)	0.123	-0.009, 0.001
Married	0.002 (0.002)	0.422	-0.003, 0.006	0.002 (0.002)	0.306	-0.002, 0.006	0.002 (0.002)	0.291	-0.002, 0.006
Income	0.000 (0.000)	0.512	-0.001, 0.000	0.000 (0.000)	0.439	-0.001, 0.000	0.000 (0.000)	0.531	-0.001, 0.000
Political views	-0.001 (0.001)	0.348	-0.002, 0.001	-0.001 (0.001)	0.472	-0.002, 0.001	-0.001 (0.001)	0.367	-0.002, 0.001
Race									
Ref=White									
Black	0.009 (0.004)	0.021	0.001, 0.017	0.010 (0.004)	0.007	0.003, 0.017	0.01 (0.004)	0.009	0.003, 0.017
Other	0.002 (0.004)	0.585	-0.005, 0.009	0.003 (0.004)	0.451	-0.005, 0.010	0.003 (0.004)	0.362	-0.004, 0.010

Table 6. Continued

	Model 3			Model 4			Model 5		
	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI
Hispanic	0.001 (0.005)	0.811	-0.009, 0.011	0.000 (0.005)	0.988	-0.010, 0.010	0.000 (0.005)	0.999	-0.010, 0.010
Intercept	0.027 (0.005)	0.000	0.017, 0.038	0.022 (0.004)	0.000	0.013, 0.031	0.028 (0.005)	0.000	0.018, 0.038
Observations	1,112			1,112			1,112		
R^2	0.245			0.29			0.296		
Design-Based F	F(19,51)=5.45	0.000		F(15,55)=7.82	0.000		F(21,49)=6.77	0.000	

Table 7. Coefficients, standard errors, and confidence intervals for survey-design adjusted OLS models predicting rate of acquiescence in eleven survey items, Models 1–2.

	Model 1			Model 2		
	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI
Respondent Characteristics						
Age Group	F(2,68)=0.03	0.968				
Ref=18–44						
45–64	-0.004	0.806	-0.033,	-0.004	0.811	-0.033,
	(0.015)		0.026	(0.015)		0.026
65+	0.000	0.983	-0.045,	0.000	0.992	-0.045,
	(0.022)		0.044	(0.022)		0.044
Education						
Ref=12 years or less						
13+ years	-0.03	0.069	-0.061,	-0.030	0.068	-0.061,
	(0.016)		0.002	(0.016)		0.002
INR in the core questions (centered)	-1.064	0.011	-1.874,	-1.062	0.011	-1.876,
	(0.406)		-0.254	(0.408)		-0.249
Ballot Experiment				F(2,68)=0.04	0.964	
Ref=Ballot B (Core)						
Ballot A (SNQ)				-0.005	0.796	-0.044,
				(0.019)		0.034
Ballot C (Knowledge)				-0.001	0.964	-0.035,
				(0.017)		0.033
Controls						
Female	0.002	0.872	-0.026,	0.002	0.883	-0.026,
	(0.014)		0.030	(0.014)		0.030
Married	-0.020	0.253	-0.054,	-0.019	0.258	-0.054,
	(0.017)		0.014	(0.017)		0.015
Income	-0.005	0.000	-0.008,	-0.005	0.000	-0.008,
	(0.001)		-0.002	(0.001)		-0.002

Table 7. Continued.

	Model 1			Model 2		
	B (Std. Err.)	p	95% CI	B (Std. Err.)	p	95% CI
Political views	0.008	0.123	-0.002,	0.007	0.129	-0.002
	0.005		0.017	(0.005)		0.017
Race	F(2,68)=1.55	0.22				
Ref=White						
Black	0.024	0.249	-0.017,	0.024	0.244	-0.017,
	(0.020)		0.065	(0.020)		0.065
Other	0.039	0.195	-0.021,	0.039	0.196	-0.021,
	(0.030)		0.099	(0.030)		0.099
Hispanic	0.078	0.005	0.024,	0.078	0.006	0.023,
	(0.027)		0.132	(0.027)		0.133
Intercept	0.573	0.000	0.508,	0.575	0.000	0.504,
	(0.032)		0.637	(0.035)		0.645
Observations	1,087			1,087		
R^2	0.074			0.074		
Design-based F	F(11,59)=5.69	0.000		F(13,57)=4.90	0.000	

Table 8. Coefficients, standard errors, and confidence intervals for survey-design adjusted OLS models predicting rate of acquiescence in 11 survey items, models 3–5.

	Model 3			Model 4			Model 5		
	Coef. (Std. err.)	p	95% CI	Coef. (Std.err.	p	95% CI	Coef. (Std. err.)	p	95% CI
Respondent Characteristics									
Age group									
Ref = 18–44									
45–64	-0.014	0.564	-0.062,	-0.003	0.856	-0.032,	-0.016	0.508	-0.065,
	-0.024		0.034	(0.015)		0.027	(0.025)		0.033
65^{+}	-0.013	0.705	-0.080,	0.001	0.969	-0.044,	-0.007	0.85	-0.076,
	(0.034)		0.055	(0.023)		0.046	(0.035)		0.062
Education									
Ref = 12 years or less									
13^{+} years	-0.02	0.440	-0.071,	-0.028	0.075	-0.060,	-0.024	0.358	-0.075,
	(0.026)		0.031	(0.016)		0.003	(0.026)		0.027
INR in the core questions (centered)	-1.082	0.008	-1.867,	-1.819	0.052	-3.656,	-1.837	0.055	-3.713,
	(0.394)		-0.297	(0.921)		0.018	(0.940)		0.038
Ballot Experiment									
Ref=Ballot B (Core)									
Ballot A (SNQ)	0.014	0.669	-0.050,	-0.005	0.788	-0.044,	0.008	0.805	-0.056,
	(0.032)		0.078	(0.019)		0.034	(0.032)		0.072
Ballot C (Knowledge)	-0.022	0.518	-0.090,	-0.001	0.953	-0.035,	-0.026	0.442	-0.093,
	(0.034)		0.046	(0.017)		0.033	(0.034)		0.041
Interactions									
Ballot* Age	F(4,66) = 0.33	0.859					F(4,66) = 0.51	0.731	
Ballot A* 45–64	0.011	0.786	-0.071,				0.018	0.673	-0.065,
	(0.041)		0.093				(0.041)		0.1
Ballot A* 65^{+}	0.000	0.993	-0.085,				-0.012	0.788	-0.097,
	(0.042)		0.084				(0.043)		0.074
Ballot C* 45–64	0.021	0.506	-0.042,				0.024	0.456	-0.040,
	(0.031)		0.084				(0.032)		0.087
Ballot C* 65^{+}	0.04	0.396	-0.054,				0.038	0.437	-0.058,
	(0.047)		0.134				(0.048)		0.133

Table 8. Continued

	Model 3			Model 4			Model 5		
	Coef. (Std. err.)	p	95% CI	Coef. (Std.err.	p	95% CI	Coef. (Std. err.)	p	95% CI
Ballot* Education	F(2,68) = 1.59	0.211					F(2,68) = 1.28	0.284	
Ballot A* 13+ years	-0.039	0.212	-0.100,				-0.03	0.335	-0.093,
	(0.031)		0.023				(0.031)		0.032
Ballot C* 13+ years	0.013	0.706	-0.055,				0.018	0.604	-0.050,
	(0.034)		0.08				(0.034)		0.086
Ballot* INR in core questions				F(2,68)=3.22	0.046		F(2,68)=1.28	0.284	
Ballot A* INR in core				1.751	0.072	-0.164,	1.761	0.083	-0.237,
				(0.960)		3.667	(1.002)		3.759
Ballot C* INR in core				-0.006	0.995	-2.164,	-0.035	0.975	-2.277,
				(1.082)		2.151	(1.124)		2.207
Controls									
Female	0.002	0.904	-0.026,	0.003	0.838	-0.025,	0.003	0.857	-0.025,
	(0.014)		0.030	(0.014)		0.031	(0.014)		0.030
Married	-0.021	0.231	-0.054,	-0.022	0.194	-0.056,	-0.023	0.173	-0.057,
	(0.017)		0.013	(0.017)		0.012	(0.017)		0.01
Income	-0.005	0.000	-0.008,	-0.005	0.000	-0.008,	-0.005	0.000	-0.008,
	(0.001)		-0.002	(0.001)		-0.003	(0.001)		-0.003
Political views	0.008	0.114	-0.002,	0.008	0.115	-0.002,	0.008	0.106	-0.002,
	(0.005)		0.017	(0.005)		0.017	(0.005)		0.017
Race									
Ref=white									
Black	0.026	0.211	-0.015,	0.022	0.283	-0.019,	0.024	0.248	-0.017,
	(0.021)		0.067	(0.021)		0.064	(0.021)		0.066
Other	0.038	0.2	-0.020,	0.039	0.19	-0.020,	0.038	0.197	-0.020,
	(0.029)		0.095	(0.030)		0.099	(0.029)		0.095
Hispanic	0.081	0.004	0.027,	0.081	0.004	0.027,	0.084	0.003	0.030,
	(0.027)		0.135	(0.027)		0.136	(0.027)		0.138
Intercept	0.575	0.000	0.494,	0.576	0.000	0.505,	0.58	0.000	0.499,
	(0.041)		0.656	(0.036)		0.647	(0.041)		0.661
Observations	1,087			1,087			1,087		
R^2	0.078			0.08			0.084		
Design-based F	F(19,51)=3.27	0.000		F(15,55)=5.48	0.000		F(21,49)=3.79	0.000	

4.2. Research Question 2: Interaction Between Respondent Characteristics and Ballot Experiment

Table 6 and Table 8 show the results of including interaction effects between the ballot experiment and the three respondent characteristics of interest: age, education, and satisficing as a response style (i.e., the item nonresponse rate in the core questions).

First, the ballot experiment did not significantly moderate the relationship between age groups and the item nonresponse rates in Modules F, I, and L nor did it moderate the relationship between age groups and the acquiescence rates. Second, the ballot experiment did not significantly moderate the relationship between education groups and acquiescence rates. However, the ballot experiment did have a differential effect on item nonresponse rates across education groups ($F(2, 68) = 3.14$, $p = 0.050$; Model 5, Table 6). More specifically, Ballots A (SNQ; $B = 0.011$, $p = 0.019$) and C (Knowledge; $B = 0.011$, $p = 0.042$) moderated the association between education and subsequent item nonresponse rates. Figure 1 plots this interaction. Among Ballot B (Core) recipients, respondents with lower education levels had higher item nonresponse rates than respondents with higher education levels, consistent with previous research (Holbrook et al. 2007; Krosnick 1991; Olson et al. 2018). However, for those receiving Ballot A (SNQ) or Ballot C (Knowledge), the subsequent item nonresponse rates were similar across education levels. That is, although this interaction effect is statistically significant, it is counter to our hypothesis (H2b). We anticipated that those with lower education who also received a burdensome module (i.e., Ballot A (SNQ) or Ballot C (Knowledge)) would have the highest item nonresponse rates. While respondents who received Ballot C (Knowledge) had among the higher item nonresponse rates in Modules F, I, and L, respondents with lower levels of education were not differentially burdened by the knowledge questions. In summary, there is no support for H2a nor H2b, as respondents with lower levels of cognitive capacity

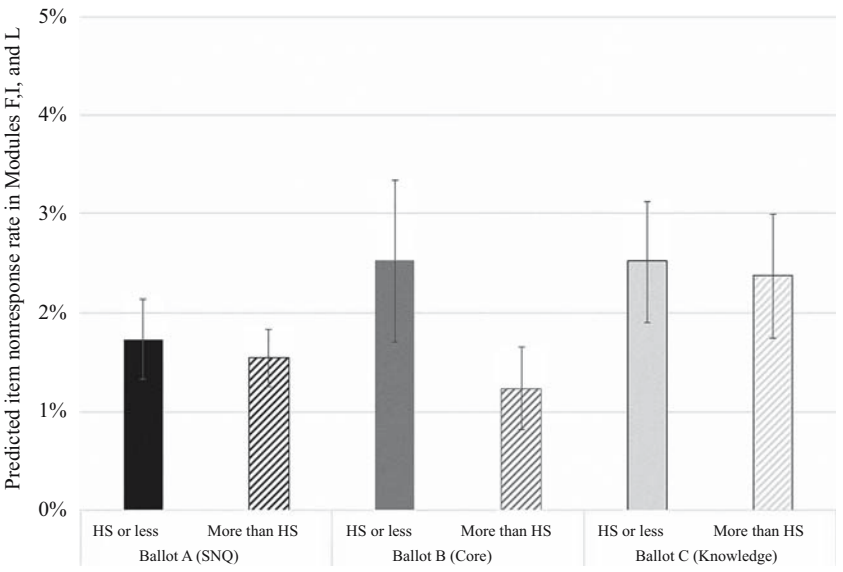


Fig. 1. Predicted item nonresponse rate in Modules F, I, and L by education level and ballot experiment received.

(i.e., older and those with lower education) were not differentially burdened by the social network questions (Ballot A) or the knowledge questions (Ballot C).

Next, the interaction between the item nonresponse rates in the core questions (Module A) and the ballot experiment did not reach the point of statistical significance in the OLS models predicting the item nonresponse rates in Modules F, I, and L ($F(2, 68) = 1.82$, $p = 0.169$, Model 4, Table 6; $F(2, 68) = 1.88$, $p = 0.160$, Model 5, Table 6). However, a higher item nonresponse rate in the core questions (Module A) was associated with a higher item nonresponse rate in Modules F, I, and L ($B = 0.866$, $p = 0.007$, Model 1, Table 5). In other words, although the response style of satisficing through item nonresponse was associated with higher item nonresponse rates throughout the interview, respondents with response styles of higher item nonresponse were not differentially burdened by the social network or the knowledge question modules, providing no support for H2c.

However, the ballot experiment did moderate the relationship between item nonresponse rates in the core questions (Module A) and acquiescence rates ($F(2, 68) = 3.22$, $p = 0.046$; Model 4, Table 8). Figure 2 plots this interaction. More specifically, for respondents that received Ballot B (Core) and Ballot C (Knowledge), a higher item nonresponse rate in the core questions (Module A) was associated with a lower acquiescence rate in the eleven survey questions. For those that received Ballot A (SNQ), the item nonresponse rate in the core questions (Module A) was not associated with the acquiescence rate in the eleven survey questions. However, this interaction was no longer significant in Model 5 ($F(2, 68) = 1.28$, $p = 0.284$), which also included interactions between the ballot experiment and age groups and the ballot experiment and education groups. Further, a higher item nonresponse rate in the core questions (Module A) was associated with a lower acquiescence rate in the eleven survey questions later in the interview ($B = -1.064$,

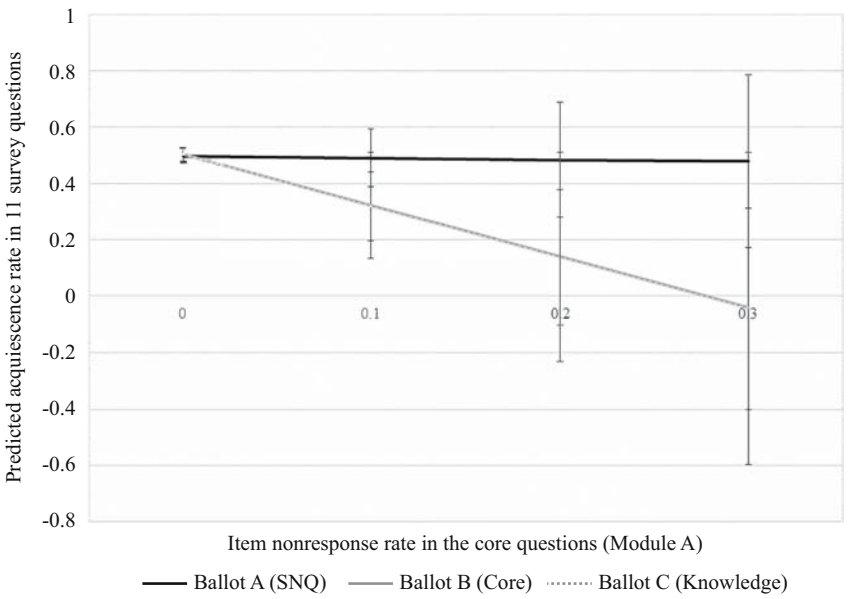


Fig. 2. Predicted acquiescence rate in 11 survey questions by item nonresponse rate in the core questions (Module A) and ballot experiment received.

Note. The lines representing the predicted values for Ballot B (Core) and Ballot C (Knowledge) overlap.

$p = 0.011$; Model 1, Table 7). In summary, there is little to no support for H2d, as H2d hypothesized that those with higher item nonresponse rates in the core questions (Module A) would have lower acquiescence rates in the eleven survey questions.

5. Discussion and Conclusion

In this article, we examined whether the presence of burdensome survey questions, with a particular focus on egocentric social network questions and knowledge questions, influences the quality of responses to subsequent unrelated questions in an omnibus survey. We ask two research questions: (1) What is the effect of burdensome survey questions (i.e., social network questions and knowledge questions) on the quality of responses to subsequent unrelated survey questions? (2) Are respondents with different characteristics such as cognitive capacity and response styles differentially burdened by these cognitively taxing survey question modules? This study has three main findings.

First, our study finds that receiving a burdensome question module is significantly related to the item nonresponse rate for subsequent questions, but that these burdensome survey question modules are not associated with the rate of acquiescent responses, at least for the questions examined here. More specifically, those who received knowledge questions had a higher item nonresponse rate for subsequent survey questions compared to those who did not receive either the knowledge or social network questions, indicating that the knowledge questions may be more detrimental to the quality of responses to subsequent unrelated survey questions as compared to social network questions. However, this association only applies for item nonresponse rates and not rates of acquiescent responses.

Previous research has indicated that egocentric social network questions are burdensome for respondents because they are asked a series of questions about other people and the number of questions that they are asked depends on the number of alters that they list (Manfreda et al. 2004; McCarty et al. 2007; Perry et al. 2018). However, the findings from this study indicate that these egocentric social network questions may not be as burdensome as social network researchers may have anticipated, at least, in interviewer-administered surveys.

Further, no research to our knowledge has explored whether answering knowledge questions in a survey interview affects the quality of subsequent responses. This study adds to the literature on the design of knowledge questions and their inclusion in omnibus surveys, suggesting that there may be need for caution in placing these items early in a questionnaire. As we find that the item nonresponse rate following this knowledge module is higher than the item nonresponse rate for the same questions following a different module, we provide some evidence that knowledge questions may be burdensome for respondents and thus influence the response quality for subsequent questions in an omnibus survey. It may be that the knowledge questions elicited greater cognitive burden compared to the particular social network questions in this survey due to the anxiety that one must choose a "correct" answer in a knowledge question rather than simply provide a response to a survey question. This finding indicates that respondents do not experience the same amount of burden from all types of burdensome survey questions.

Second, our results do not support the hypothesis that respondents of lower cognitive capacity are differentially burdened by social network or knowledge questions. More specifically, there were no significant differences in acquiescence rates by cognitive

capacity and the presence of a burdensome module. Similarly, those who received a burdensome question module and also had a lower cognitive capacity did not have the highest item nonresponse rates. Overall, respondents with a lower cognitive capacity did not exhibit more satisficing behaviors on questions that followed the social network or the knowledge question modules, suggesting they did not experience more cognitive burden from these modules compared to respondents with higher cognitive capacity.

Third, our results suggested that respondents with a response style of satisficing through the form of higher item nonresponse rates at the beginning of the interview would continue to exhibit this form of satisficing throughout the interview, as evidenced through higher item nonresponse rates and lower acquiescence rates later in the interview. However, our results do not support the hypothesis that respondents with this response style would be differentially burdened by social network questions or knowledge questions.

These results have practical implications for survey design. Because knowledge questions were found to be associated with a greater item nonresponse rate for subsequent unrelated survey questions and this finding held for all respondents regardless of cognitive capacity or response style, survey designers should carefully consider how the placement of these questions may impact the quality of responses for subsequent questions. Therefore, it is important for questionnaire designers of omnibus surveys to know that burdensome survey questions carry the risk of reducing response quality on subsequent unrelated survey questions. It may be beneficial to place knowledge questions near the end of interviewer-administered omnibus surveys in order to preserve the quality of responses for subsequent unrelated survey questions.

As with all research, this study has limitations. The conclusions made here about the effect of knowledge questions and social network questions on the quality of responses to subsequent modules are limited to interviewer-administered surveys. Previous research has found that interviewers may help respondents through answering social network questions (Kogovšek 2006; Kogovšek and Ferligoj 2005; Paik and Sanchagrin 2013; Valente et al. 2017) and thus, the association between social network questions and subsequent data quality may differ when social network questions are asked in self-administered modes. Future research should explore whether respondents experience differential amounts of cognitive burden due to knowledge and social network questions in different survey modes.

Further, in the GSS social network module, respondents were only asked three follow-up questions for each alter named in the name generator section – the sex, race, and type of relationship with the person. Because this social network module asked a limited number of follow-up questions that may have been easier to answer, the GSS module may not have been burdensome for respondents. This difference between the difficulty of the social network module and the knowledge question module may explain our finding that respondents who received the ballot with the knowledge questions had greater item nonresponse rates, on average. Future research should explore the relationship between the number and type of follow-up questions asked of respondents in social network questions and the amount of cognitive burden experienced by respondents. Additionally, because only those who received Ballot C (Knowledge) were asked the knowledge questions, we cannot differentiate between the effect of knowledge questions' appearance in the survey from the effect of these questions' location within the survey. Future research should additionally experimentally vary the placement of other burdensome survey question

modules such as the knowledge questions to differentiate the effect of placement and presence of knowledge questions on data quality in subsequent survey questions.

In addition, our measure for acquiescence is limited to questions asking about the same topic (i.e., aging) and our measure is calculated using only eleven survey questions. As the R squared value from the acquiescence models is very low, further research should aim to use more survey questions about a variety of topics to measure acquiescence in order to use a more accurate measure. Furthermore, acquiescence can be difficult to isolate from nondifferentiation, which occurs when respondents select the same or similar responses in multiple consecutive questions that have identical response options (Krosnick 1991). Because our operationalization of acquiescence measures the number of questions to which the respondent reported either strongly agree or somewhat agree, and thus, utilizes questions with the same response options, it is difficult to know whether this is measuring acquiescence or nondifferentiation. Future research should examine how burdensome survey questions are associated with acquiescent behaviors on a wider variety of topics and should examine nondifferentiation in addition to acquiescent behaviors.

Last, this study only examined two data quality indicators. Most questions asked in Module I (Misc.) and Module L (Shared capital) were factual questions, and respondents are less likely to exhibit satisficing behaviors to factual questions than attitudinal questions (Krosnick 1991). Further, the public use data file did not include module-specific interview lengths nor partially completed cases, meaning interview length and breakoffs could not be included in this study's analyses. We additionally were limited in the operationalization of item nonresponse; specifically, we were unable to disentangle no answer and "don't know" responses due to the low prevalence of "no answer" responses. Future research should continue to explore the effect of burdensome survey questions in omnibus surveys using other indicators of data quality and with a more nuanced operationalization of item nonresponse.

This study provides unique insights. The experimental design of the 2010 GSS allowed us to analyze how the presence of a burdensome question module is associated with the quality of responses to subsequent unrelated survey questions. Having data that come from an experimentally designed and nationally representative survey such as the 2010 GSS is greatly beneficial to be able to generalize our findings to all non-institutionalized United States adults. Results indicate that caution should be used when including knowledge questions in omnibus surveys and attention should be paid to their placement within the survey. Future research should continue to investigate the extent to which egocentric social network questions and knowledge questions are burdensome for respondents across different types of surveys (i.e., omnibus or not), across different survey modes (i.e., self-administered versus interviewer-administered), and using different response quality indicators in order to gain more information about best practices for asking these questions and gathering these types of data.

6. References

AAPOR, The American Association for Public Opinion Research. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* AAPOR.

American National Election Studies. 2020. *The ANES Guide to Public Opinion and Electoral Behavior.* Available at: https://electionstudies.org/resources/anes-guide/. (accessed December 2020).

AmeriSpeak. 2020. *AmeriSpeak Omnibus.* Available at: https://amerispeak.norc.org/our-capabilities/Pages/AmeriSpeak-Omnibus.aspx?gclid=CjwKCAjwssD0BRBIEiwA-JP5rIJ_c6ljPjrg0_H0mgV55cRzWceO-qX0X28KAaVkJX3X51Zk8APDvhoCJyo-QAvD_BwE. (accessed April 2020).

Bolt, D.M., and T.R. Johnson. 2009. "Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style." *Applied Psychological Measurement* 33(5): 335–352. DOI: https://doi.org/10.1177/0146621608329891.

Bradburn, N.M. 1978. "Respondent Burden." In Proceedings of the Survey Research Methods Section: American Statistical Association: 35–40. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf. (accessed January 2021).

Colsher, P.L. and R.B. Wallace. 1989. "Data Quality and Age: Health and Psychobehavioral Correlates of Item Nonresponse and Inconsistent Responses." *Journal of Gerontology* 44(2): P45–P52. DOI: https://doi.org/10.1093/geronj/44.2.P45.

De Leeuw, E.D. 1992. *Data Quality in Mail, Telephone and Face to Face Surveys*. Amsterdam: TT-pulikaties.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, And Mixed-Mode Surveys: The Tailored Design Method*. Hoboken: John Wiley & Sons, Inc.

Galesic, M. 2006. "Dropouts on The Web: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics* 22(2): 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf.

Galesic, M. and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73(2): 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Groves, R.M., F.J. Fowler Jr., M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken: John Wiley & Sons, Inc.

Hamby, T. and W. Taylor. 2016. "Survey Satisficing Inflates Reliability and Validity Measures: An Experimental Comparison of College and Amazon Mechanical Turk Samples." *Educational and Psychological Measurement* 76(6): 912–932. DOI: https://doi.org/10.1177/0013164415627349.

Holbrook, A.L., Y.I. Cho, and T. Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70(4): 565–595. DOI: https://doi.org/10.1093/poq/nfl027.

Holbrook, A.L., M.C. Green, and J.A. Krosnick. 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(1): 79–125. DOI: https://doi.org/10.1086/346010.

Holbrook, A.L., J.A. Krosnick, D. Moore, and R. Tourangeau. 2007. "Response Order Effects In Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes." *Public Opinion Quarterly* 71(3): 325–348. DOI: https://doi.org/10.1093/poq/nfm024.

Kleinert, C., B. Christoph, and M. Ruland. 2019. "Experimental Evidence on Intermediate and Long-Term Consequences of Test-Induced Respondent Burden For Panel

Attrition." *Sociological Methods & Research* March 2019:1–32. DOI: https://doi.org/10.1177/0049124119826145.

Knäuper, B., R.F. Belli, D.H. Hill, and A.R. Herzog. 1997. "Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality." *Journal of Official Statistics* 13(2): 181–199. Available At: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/question-difficulty-and-respondents39-cognitive-ability-the-effect-on-data-quality.pdf.

Kogovšek, T. 2006. "Reliability and Validity of Measuring Social Support Networks by Web and Telephone." *Metodološki zvezki* 3(2): 239–252. DOI: https://doi.org/10.51936/kjdq1083.

Kogovšek, T., and A. Ferligoj. 2005. "Effects on Reliability and Validity of Egocentered Network Measurements." *Social Networks* 27(3): 205–229. DOI: https://doi.org/10.1016/j.socnet.2005.01.001.

Krosnick, J.A. 1991. "Response Strategies for Coping with The Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. DOI: https://doi.org/10.1002/acp. 2350050305.

Krosnick, J.A., S. Narayan, and W.R. Smith. 1996. "Satisficing in Surveys: Initial Evidence." *Causal Mechanisms in Program Evaluation* 70: 29–44. DOI: https://doi.org/10.1002/ev.1033.

Leeper, T.J. 2014. "Cognitive Style and the Survey Response." *Public Opinion Quarterly* 78(4): 974–983. DOI: https://doi.org/10.1093/poq/nfu042.

Lelkes, Y., J.A. Krosnick, D.M. Marx, C.M. Judd, and B. Park. 2012. "Complete Anonymity Compromises the Accuracy of Self-Reports." *Journal of Experimental Social Psychology* 48(6): 1291–1299. DOI: https://doi.org/10.1016/j.jesp. 2012.07.002.

Manfreda, K.L., V. Vehovar, and V. Hlebec. 2004. "Collecting Ego-Centred Network Data Via the Web." *Metodološki zvezki* 1(2): 295–321. DOI: https://doi.org/10.51936/sbuu6702.

Marsden, P.V. 2016. "Survey Methods for Network Data." In *The SAGE Handbook of Social Network Analysis*, edited by J. Scott and P.J. Carrington: 370–388. London: SAGE Publications Ltd.

McCarty, C., P.D. Killworth, and J. Rennell. 2007. "Impact of Methods for Reducing Respondent Burden on Personal Network Structural Measures." *Social Networks* 29(2): 300–315. DOI: https://doi.org/10.1016/j.socnet.2006.12.005.

Mondak, J.J. 1999. "Reconsidering the Measurement of Political Knowledge." *Political Analysis* 8(1): 57–82. DOI: https://doi.org/10.1093/oxfordjournals.pan.a029805.

Mondak, J.J. 2001. "Developing Valid Knowledge Scales." *American Journal of Political Science* 45(1): 224–238. DOI: https://doi.org/10.2307/2669369.

Nadeau, R., and R.G. Niemi. 1995. "The Process of Answering Factual Knowledge Questions in Surveys." *Public Opinion Quarterly* 59(3): 323–346. DOI: https://doi.org/10.1086/269480.

Narayan, S., and J.A. Krosnick. 1996. "Education Moderates Some Response Effects in Attitude Measurement." *Public Opinion Quarterly* 60(1): 58–88. DOI: https://doi.org/10.1086/297739.

Olson, K., J.D. Smyth, and A. Ganshert. 2018. "The Effects of Respondent and Question Characteristics on Respondent Answering Behaviors in Telephone Interviews." *Journal of*

Survey Statistics and Methodology 7(2): 275–308. DOI: https://doi.org/10.1093/jssam/smy006.

Paik, A., and K. Sanchagrin. 2013. "Social Isolation in America: An Artifact." *Social Networks* 78(3): 339–360. DOI: https://doi.org/10.1177/0003122413482919.

Park, D.C. 2000. "The Basic Mechanisms Accounting For Age-Related Decline in Cognitive Function." In *Cognitive Aging: A Primer*, edited by D.C. Park and N. Schwarz: 3–19. East Sussex: Psychology Press.

Paulhus, D.L. 1991. "Measurement and control of response bias." In *Measures of Personality and Social Psychological Attitudes*, edited by J.P. Robinson, P.R. Shaver, and L.S. Wrightsman: 19–59. San Diego, CA: Academic Press.

Perry, B.L., B.A. Pescosolido, and S.P. Borgatti. 2018. *Egocentric Network Analysis: Foundations, Methods, and Models*. Cambridge: Cambridge University Press.

Smith, T.W., M. Davern, J. Freese, and S.L. Morgan. 2019. *General Social Surveys, 1972–2018: Cumulative Codebook*. Chicago: NORC University of Chicago. Available at: https://gss.norc.org/documents/codebook/gss_codebook.pdf. (accessed December 2020).

Smith, T.W., and B. Schapiro. 2017. *General Social Surveys – Panel Interviews: 2006–2014 Cumulative Codebook.* Chicago: NORC University of Chicago. Available at: http://gss.norc.org/Documents/codebook/Panel%20Codebook.pdf. (accessed December 2020).

Son, J. 2020. *Email message to author, March 24, 2020*. NORC at the University of Chicago.

Sturgis, P., N. Allum, and P. Smith. 2008. "An Experiment on The Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72(1): 90–102. DOI: https://doi.org/10.1093/poq/nfm032.

The American National Election Studies. 2020. *The ANES Guide to Public Opinion and Electoral Behavior.* Available at: https://electionstudies.org/resources/anes-guide/. (accessed December 2020).

Tourangeau, R. and K A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103(3): 299–314. DOI: https://doi.org/10.1037/0033-2909.103.3.299.

Tourangeau, R., L.J. Rips, and K.A. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tu, S. and P. Liao. 2007. "Social Distance, Respondent Cooperation and Item Nonresponse in Sex Survey." *Quality & Quantity* 41:177-199. DOI: https://doi.org/10.1007/s11135-007-9088-0.

U.S. Census Bureau. 2011. *Age and Sex Composition: 2010*. Available at: https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf. (accessed October 2021).

Valente, T.W., L. Dougherty, and E. Stammer. 2017. "Response Bias Over Time: Interviewer Learning and Missing Data in Egocentric Network Surveys." *Field Methods* 29(4): 303–316. DOI: https://doi.org/10.1177/1525822X17703718.

Wetzel, E., C.H. Carstensen, and J.R. Böhnke. 2013. "Consistency of Extreme Response Style and Non-Extreme Response Style Across Traits." *Journal of Research in Personality* 47(2): 178–189. DOI: https://doi.org/10.1016/j.jrp. 2012.10.010.

Zaller, J.R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: The Press Syndicate of the University of Cambridge.

Received January 2021
Revised November 2021
Accepted March 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1051–1067, http://dx.doi.org/10.2478/JOS-2022-0045

Relationship Between Past Survey Burden and Response Probability to a New Survey in a Probability-Based Online Panel

Haomiao Jin[1] *and Arie Kapteyn*[1]

We conducted an idiographic analysis to examine the effect of survey burden, measured by the length of the most recent questionnaire, or number of survey invitations (survey frequency) in a one-year period preceding a new survey, on the response probability to a new survey in a probability-based Internet panel. The individual response process was modeled by a latent Markov chain with questionnaire length and survey frequency as explanatory variables. The individual estimates were obtained using a Monte Carlo based method and then pooled to derive estimates of the overall relationships and to identify specific subgroups whose responses were more likely to be impacted by questionnaire length or survey frequency. The results show an overall positive relationship between questionnaire length and response probability, and no significant relationship between survey frequency and response probability. Further analysis showed that longer questionnaires were more likely to be associated with decreased response rates among racial/ethnic minorities and introverted participants. Frequent surveys were more likely to be associated with decreased response rates among participants with a large household. We discuss the implications for panel management and advocate targeted interventions for the small subgroups whose response probability may be negatively impacted by longer questionnaires or frequent surveys.

Key words: Survey nonresponse; survey burden; Internet panel; ideographic analysis; latent Markov chain.

1. Introduction

An online panel is a sample of persons who have agreed to complete surveys via the Internet. By tailoring key respondent burden variables like questionnaire length and survey frequency (Bradburn 1978), panel administrators can control the burden of taking surveys among panel participants. Based on common assumptions on the impacts of respondent burden (Bradburn 1978; Sharp and Frankel 1983; Baker et al. 2010), one may surmise that the experiences of long questionnaires and frequent surveys may overburden participants in panel studies and therefore decrease their propensity to complete a future survey. In this article, we aim to investigate the effects of the length of preceding surveys, or the frequency by which respondents are invited to respond to a new survey, on the probability that a respondent in a longitudinal probability-based Internet panel will respond to a new survey invitation.

[1] Center for Economic and Social Research, University of Southern California, 635 Downey Way VPD, Los Angeles, 90089 California, U.S.A. Emails: haomiaoj@usc.edu and kapteyn@usc.edu
Acknowledgments: The research in this article was supported by the U.S. National Institute on Aging under grant 5U01AG054580.

The context for the current article comes from the Understanding America Study (UAS), which is a U.S. nationally representative online panel of approximately 9,000 respondents aged 18 or older. The UAS was launched in 2014 and is supported by grants from several organizations, including the U.S. National Institute on Aging and the Social Security Administration. Panel members were randomly selected through address-based sampling and are compensated based on the estimated time spent on each survey, at a rate of USD 20 per 30 minutes. Interviews never exceed thirty minutes, but often may be shorter with incentive payments proportionately lower. Panel members without prior access to the Internet are provided with a tablet computer and Internet access. Since its launch, the UAS has conducted over 300 surveys of varying length on a broad number of topics including labor market behavior, income and wealth, mental and physical health, financial literacy, consumer behavior, personality, cognitive ability, health, and political views. Frequently, surveys may target specific groups. Hence, different subsamples of the UAS may receive different numbers of invites to do a survey. Details about the UAS can be found on the UAS website (UAS).

Although there exists an extensive literature on the effects of questionnaire length on response probabilities, upon closer inspection it is hard to draw firm conclusions. Most literature on online panels deals with cross-sectional surveys. Surprisingly, the way the length of surveys is defined is often ambiguous. In several cases, respondents are not told *ex ante* how long the survey may be expected to be, or at least it is not clear from the published studies whether respondents were informed about the length of the survey (Guo et al. 2016; Deutskens et al. 2004; Blumenberg et al. 2019). Koitsalu et al. (2018) find that a longer questionnaire leads to higher response rates, but in the announcement to potential respondents, everyone was told the same survey length (20 minutes). In other studies, survey length is defined after the fact, that is, by the realized length of the interview (Nicoletti and Peracchi 2005; Hill and Willis 2001). This makes it difficult to interpret the effect of interview length as a measure of burden. Potentially a longer interview may be an indicator of engagement, which would then predict higher future response rates, which indeed is often found.

The most direct evidence on the effect of survey length on response rates comes from randomized controlled trials that vary the announced length of surveys (Crawford et al. 2001; Galesic and Bosnjak 2009; Marcus et al. 2007). In these studies, there generally appears a clear negative effect of interview length on response rates. However, Crawford et al. (2001) report that although more respondents started the announced short survey, those respondents were also more likely to break off.

One might suspect that the negative effect of survey length on response rates would translate to longitudinal settings where preceding long surveys, may discourage respondents to accept a new survey invitation. Neither Hill and Willis (2001) for the Health and Retirement Study (HRS), nor (Zabel 1998) for the Survey of Income and Program Participation (SIPP) and Panel Study of Income Dynamics (PSID) find such a negative effect. In both studies, survey length is defined by the measured number of minutes the survey took, which once again may not be a measure of burden, but possibly of engagement. Zabel (1998) also takes advantage of a conscious decision by PSID staff in 1973 to reduce the length of the survey. It is found that this does reduce attrition. McCarthy et al. (2006) compare response rates to a number of agricultural surveys. Survey

durations are based on the U.S. Office of Management and Budget's estimates and hence are exogenous to the respondents. The authors report statistically marginally significant and quantitatively small negative effects of the length of preceding surveys on the likelihood that a new survey is returned.

The studies cited here sometimes provided incentives to increase response rates, but generally incentives were not tied to the length of an interview. Deutskens et al. (2004) varied the incentive in that the short questionnaire earned EUR 2 and the long questionnaire earned EUR 5. Whether that is an adequate compensation for the time spent on the survey is debatable, as the short survey already took 15–30 minutes.

The relationship between survey frequency and survey response has been studied much less, which is not surprising as most surveys are cross sectional and most longitudinal studies have a fixed periodicity, so that there is no variation in frequency that can be exploited. Zabel (1998) compares PSID and SIPP and finds that attrition is higher in the latter panel. During the observation period for his study, PSID respondents were interviewed once a year and SIPP respondents three times a year. He finds a higher attrition rate in SIPP. Although conceivably that might reflect the effect of the more frequent interview requests in SIPP, there are of course many other differences between the two panels, which might explain differences in attrition rates. McCarthy et al. (2006) exploit the fact that different agricultural operations have different probabilities of being contacted for a survey. They find only a negligible relation between frequency of survey invitations and response rates. A few studies have evaluated the relationship between survey frequency and survey response in ecological momentary assessments, where respondents may be prompted for brief surveys up to six times a day. The studies mostly suggest no relationship between survey frequency and completion rates (Eisele et al. 2020; Walsh and Brinker 2015; Stone et al. 2003; Ono et al. 2019).

Response rates are not just a function of questionnaire length and number of invites. In particular saliency of a survey topic is found to be important (Marcus et al. 2007). However, what is salient to one respondent may not be salient to another. Leverage-saliency theory proposed by Groves et al. (2000) speaks to this. Essentially, the likelihood that a survey invitation will be accepted depends on many factors, such as survey sponsorship, survey topic, survey length, incentive level, and unobserved personal characteristics. Although the survey topic is of course observable, how salient that topic is to a respondent is generally unknown *ex ante*.

In panel studies, respondents may receive larger incentives from taking longer or more frequent surveys if they are paid based on questionnaire length or time spent on taking surveys (Baker et al. 2010). Depending on how the incentives are evaluated against the other features of a survey, panel participants may exhibit diverse response behaviors to new surveys. Some participants may be motivated by the larger incentives from taking longer or more frequent surveys as a means to earn extra income. Other participants, though, may be less motivated by financial incentives to take more surveys. Such individual differences call for an idiographic (bottom up) analysis focusing on individual-specific relationships between questionnaire length and survey frequency and response propensity to a new survey.

To the best of our knowledge, there have been no published studies looking at the relationships between the length and frequency of past surveys and the response

propensity to a new survey in online panels. The present study examined the relationships between survey burden, measured by the length of the last questionnaire (defined exogenously) and the survey frequency (number of surveys in a one-year period preceding a new survey), and response to a new survey in the UAS. Specifically, the study conducted an idiographic analysis that examines individual-specific relationships between survey burden and probability of response to a new survey. Next, the different response patterns were related to observable demographic and socio-economic characteristics. Given the exploratory nature of the study, no specific hypotheses were postulated about what subgroups would be identified and which characteristics would differ between subgroups. The present study filled the knowledge gap about the impacts of experiencing longer questionnaires and frequent surveys on the response propensity to a new survey in online panels. The study also contributed to the methodology of analyzing individual survey response processes in panel studies by implementing a new dynamic model based on latent Markov Chain and Monte Carlo Markov Chain methods.

2. Methods

2.1. Overview of the Methods

The present study conducted an idiographic analysis for a total of 1,333 individuals randomly selected from all currently enrolled or former (attrited) UAS panel members. The analysis was based on the latent Markov chain method (Bartolucci et al. 2019), which is a flexible and powerful analytic approach to modelling the individual survey response processes, by incorporating unobserved variables like survey commitment, in addition to questionnaire length and survey frequency as covariates. Due to the computational burden of fitting the latent Markov chain model, we did not carry out the analysis for all current and past panel members. A Monte Carlo based method was used to estimate the associations between past survey burden and response to a new survey for each individual. The individual estimates were pooled to derive estimates of the overall relationships and to identify specific groups whose responses were mostly impacted by questionnaire length or survey frequency. Finally, an ordered logistic regression analysis was conducted to examine the relationships between the group memberships and sociodemographic characteristics and personality traits.

2.2. Measures

The survey response of each panel participant was modeled as the outcome of a latent Markov chain model including questionnaire length, survey frequency, and panel commitment as explanatory variables. Questionnaire length was measured by the number of screens viewed by the survey taker in the last survey they responded to. This is not a perfect proxy for questionnaire length. The UAS typically shows one question per screen, but it also uses grids, where respondents have to choose a category on each line. On the other hand, grid use is relatively rare and there is no reason to assume a systematic relation between the number of screens shown and the percent of those that would use grid questions (For example, the UAS does not try to reduce the number of screens by packing

more questions in a grid). Survey frequency was measured by the number of surveys to which a respondent was invited in the year preceding a survey.

Panel commitment was defined as an unobserved continuous variable indicating the feelings of obligation by panel participants to remain with the panel through taking surveys. It was defined as an unobserved variable because it is impossible to measure panel commitment at the time when there was a nonresponse. Survey response was coded as a binary variable with one indicating a survey had been completed and zero indicating the survey had not been completed.

Participants were grouped based on the relationships between questionnaire length, survey frequency and survey response. Group membership was modeled as the outcome of an ordered logistic regression model including socio-demographic characteristics and personality traits as predictors. Selection of the predictors was based on a systematic review of potential factors affecting response behavior in online surveys (Fan and Yan 2010). Socio-demographic variables included gender (coded as female (Yes/No)), age, racial/ethnical minority (not non-Hispanic White (Yes/No)), marital status (married (Yes/No)), education level (high school or less (Yes/No)), household income, household size, and the big five personality traits (McCrae 2004; McCrae and Costa 2008). We included personality traits in addition to commonly examined demographic variables since prior research has suggested certain personality traits like extroversion, conscientiousness, and openness to new experience are related to the probability of response to a subsequent wave in online panels (Cheng et al. 2020; Lugtig 2014). Data on the personality measures were obtained from the first survey every UAS participant takes when joining the panel.

2.3. *Modeling the Individual Survey Response Process*

The individual survey response process was modeled by a latent Markov model. Let panel participants be indexed by $i = 1, 2, ..., I$, and assume that in the past twelve months participant i was invited for a total of J_i surveys indexed by $j = 1, 2, ..., J_i$ in chronological order. Participant $i's$ completion of the jth invited survey, that is, $Y_{i,j}$, was related to survey length, survey frequency, and the unobserved panel commitment level through a logit link function as below.

$$Y_{i,j} \sim Bernoulli\left(\frac{e^{\eta_{i,j}}}{1+e^{\eta_{i,j}}}\right) \tag{1}$$

$$\eta_{i,j} = \theta_i + \lambda_i x_{i,j} + \beta_{i,1} S_{i,j} + \beta_{i,2} S_{i,j}^2 + \delta_{i,1} N_{i,j} + \delta_{i,2} N_{i,j}^2 \tag{2}$$

$\eta_{i,j}$ is the linear predictor. $x_{i,j}$ is the unobserved panel commitment level. $S_{i,j}$ is the number of screens viewed by the survey taker in the last survey. $N_{i,j}$ is the number of survey invitations in the preceding twelve months. Second-order terms $S_{i,j}^2$ and $N_{i,j}^2$ are included to model the potential non-linear relationships between questionnaire length, survey frequency, and survey response. The intercept θ_i, the impact of panel commitment on response λ_i, the impacts of questionnaire length on response $\beta_{i,1}$ and $\beta_{i,2}$, the impacts of survey frequency on response $\delta_{i,1}$ and $\delta_{i,2}$ are unknown coefficients to be estimated from data.

Dynamics of the unobserved panel commitment level $x_{i,j}$ was modeled by a latent Markov chain as below.

$$x_{i,j} = \left(\phi_{i,0} + \phi_{i,1}T_{i,j} + \phi_{i,2}T_{i,j}^2\right)x_{i,j-1} + \varepsilon_i \tag{3}$$

$T_{i,j}$ are the days between the $(j-1)$th and the jth survey invitation. The second-order term $T_{i,j}^2$ are included to model the potential nonlinear impact of the survey interval on the transition of panel commitment level. ε_i are an error term assumed to be normally distributed with mean zero and unknown variance. The impacts of survey interval on the transition of panel commitment $\phi_{i,1}$ and $\phi_{i,2}$ are unknown coefficients.

The unknown coefficients in Equations (2) and (3) were estimated in an idiographic way, that is, for each individual separately. Since the present article focuses on the relationships between questionnaire length, survey frequency, and survey response, we are most interested in the estimates of $\beta_{i,1}$, $\beta_{i,2}$, $\delta_{i,1}$ and $\delta_{i,2}$. The unknown coefficients were estimated by using the Particle Markov Chain Monte Carlo method (Andrieu et al. 2010). This method improves the standard MCMC method by incorporating particle filers to facilitate efficient Bayesian inference on latent Markov chain models (Andrieu et al. 2010). The prior of θ_i was set to be normally distributed with mean 2 and standard deviation 1, reflecting that most UAS participants have a high albeit varying initial response propensity. The prior distribution of λ_i was set to a normal distribution with mean 0.2 and standard deviation 0.1 to reflect the potential positive association between panel commitment level and survey response. The priors of $\beta_{i,1}$, $\beta_{i,2}$, $\delta_{i,1}$ and $\delta_{i,2}$ were set to be normally distributed with mean 0 and standard deviation 0.2. This setting was informed by the prevalence of no-association findings in prior relevant studies (Yetter and Capaccioli 2010; Blumenberg et al. 2019; Eisele et al. 2020; Walsh and Brinker 2015; Stone et al. 2003; Ono et al. 2019). Finally, the priors of $\phi_{i,1}$ and $\phi_{i,2}$ were set to be normally distributed with mean 0 and standard deviation 0.2, reflecting our experience that the survey interval has little impact on the change of panel commitment level. The estimation was implemented using the R package "nimble" (De Valpine et al. 2017).

2.4. *Pooling Individual Estimates, Identifying Subgroups Based on the Direction of Relationship, and Ordered Logistic Regression Analysis*

We used the means of the posterior distributions as individual estimates for $\beta_{i,1}$, $\beta_{i,2}$, $\delta_{i,1}$ and $\delta_{i,2}$ and then pooled those individual estimates to derive the overall estimates of the associations between questionnaire length, survey frequency on survey response.

We used a multi-step procedure to group the individual estimates based on whether the associations are likely to be positive, negative, or indicate no relationship. For the association between questionnaire length and survey response, we first grouped individuals into three subgroups based on the estimates of $\beta_{i,1}$ and $\beta_{i,2}$, respectively. Those subgroups are the positive association group (i.e., the estimate is larger than zero plus two times the standard error of the individual estimates), the negative association group (i.e., the estimate is smaller than zero minus two times the standard deviation), and the no-association group (i.e., the estimate is between the two thresholds). This subgrouping procedure is often used in the area of statistical quality control to identify

subjects that are most impacted by a factor (Oakland and Oakland 2018). After this calculation, each individual had two group memberships for $\beta_{i,1}$ and $\beta_{i,2}$, respectively. Our analysis showed that only two individuals were assigned to groups of opposite directions, that is, the positive association group for $\beta_{i,1}$ and negative association group for $\beta_{i,2}$, or the negative association group for $\beta_{i,1}$ and positive association group for $\beta_{i,2}$. We thereafter removed those cases from the subsequent analysis. For removed cases, the effect of more surveys was in one direction for small number of screens, and the effect of more surveys was in the opposite direction for large number of screens.

The remaining individuals were grouped into the ultimate positive association group if they were in the positive association group for either $\beta_{i,1}$ or $\beta_{i,2}$ or both. They were grouped into the ultimate negative association group if they were in the negative association group for either $\beta_{i,1}$ or $\beta_{i,2}$ or both. Finally, they were grouped into the ultimate no-association group if they were in the no-association groups for both $\beta_{i,1}$ and $\beta_{i,2}$ Similar procedures were applied to $\delta_{i,1}$ and $\delta_{i,2}$ to derive the group membership based on the association between survey frequency and survey response.

Finally, to examine the relationships between participant characteristics and group membership, we conducted an ordered logistic regression analysis with the group membership as an ordinal outcome variable (i.e., negative association $<$ no association $<$ positive association) and sociodemographic characteristics and personality traits as predictors. The personality traits were standardized before entering into the regression analysis. The model was fitted using the "polr" function from the R package "MASS" (Venables and Ripley 2013).

3. Results

A total of 1,333 individuals were randomly selected from all currently enrolled or attrited UAS panel members for the analysis. Individuals in the analysis sample were 58.2% female (n = 774), 35.5% racial/ethnic minority (n = 472), 53.9% married (n = 718), 24.3% with high school or lower education (n = 324), and 21.5% attrited (n = 286). The age at enrollment into the panel was on average 46.8 years old with a standard deviation of 16.2 years. Individuals in the analysis sample were invited for an average of 74.2 surveys with a standard deviation of 45.6. They had been in the UAS panel for an average of 35.5 months with a standard deviation of 21.4 months.

Pooling individual estimates from the latent Markov chain model showed an overall positive relationship between questionnaire length of the last survey, measured by the number of screens, and response to a new survey (estimate for the linear term = 0.043, 95% CI: 0.039 ~ 0.047, $P < 0.001$; estimate for the quadratic term = 0.066, 95% CI: 0.061 ~ 0.070, $P < 0.001$). There was no significant association between survey frequency, measured by the number of survey invitations in the preceding year, and response to a new survey (estimate of the linear term = 0.003, 95% CI: -0.001–0.006, P = 0.22; estimate of the quadratic term = -0.004; 95% CI: -0.009–0.001, P = 0.06).

As shown in Table 1, the multi-step grouping procedure identified a total of 414 individuals for whom seeing more screens in the last survey was associated with increased response propensity to a new survey. In contrast, seeing more screens in the last survey was associated with decreased response propensity to a new survey for 42 individuals. The

Table 1. Descriptive statistics and multiple-group comparisons based on the association between the number of screens in the last survey and responding to a new survey.

	More screens associated with decreased response (N = 42)	No association (N = 871)	More screens associated with increased response (N = 414)	P-value of test for equality across columns
Female	28 (66.7%)	495 (56.9%)	247 (59.8%)	0.32
Age	43.69 (16.88)	46.46 (16.03)	47.85 (16.35)	0.16
Racial/ethnic minority	20 (48.8%)	328 (37.7%)	121 (29.3%)	0.003
Married	22 (52.4%)	470 (54.0%)	222 (53.6%)	0.98
High school or less	10 (23.8%)	209 (24.0%)	105 (25.4%)	0.86
Household income				
• Less than USD 35,000	13 (31.0%)	219 (25.1%)	120 (29.0%)	0.28
• USD 35,000 – USD 59,999	11 (26.2%)	260 (29.9%)	106 (25.6%)	0.27
• USD 60,000 – USD 99,999	9 (21.4%)	184 (21.1%)	92 (22.2%)	0.90
• More than USD 100,000	9 (21.4%)	202 (23.2%)	93 (22.5%)	0.93
Three or more people in household	29 (69.0%)	392 (45.0%)	202 (48.8%)	0.006
Big five personality scores				
• Extroversion	26.53 (6.52)	25.99 (6.22)	27.17 (6.29)	0.007
• Agreeable	37.27 (5.04)	36.14 (5.50)	36.59 (5.65)	0.21
• Conscientiousness	36.17 (5.54)	36.12 (5.69)	36.17 (5.62)	0.99
• Neuroticism	22.41 (6.82)	21.97 (6.58)	21.33 (6.84)	0.23
• Openness	36.05 (6.05)	36.59 (6.06)	36.75 (6.45)	0.76

remaining 871 individuals were classified into the "no-association" group. Multi-group comparisons of sociodemographic variables and personality traits showed that the three groups were significantly different in racial/ethnic minority status, household size, and extroversion. The ordered logistic regression analysis (Table 2) revealed two significant predictors for group membership. For respondents who belong to a racial/ethnic minority, the odds of their response rates being negatively associated with more screens in the last survey was 1.56 times (i.e., 1/0.64) that of respondents who are non-Hispanic Whites, holding constant all other variables. For every one unit increase in respondents' standardized extroversion score, the odds of their response rates being not associated or positively associated with more screens in the last survey increases by 1.19, holding constant all other variables.

As shown in Table 3, having received more survey invitations in the preceding year was associated with increased response propensity to a new survey for a total of 126 individuals. Receiving more survey invitations in the preceding year was associated with decreased response propensity to a new survey for 152 individuals. The remaining 1,052 individuals were classified into the "no-association" group. Multi-group comparisons revealed significant differences by household size and by extroversion. The ordered logistic regression analysis (Table 4) revealed only one significant predictor of group membership. Respondents with three or more people in their household were 1.56 times (i.e., 1/0.64) more likely than respondents with smaller households to lower their response in reaction to being invited to more surveys, holding constant all other variables.

Table 2. Ordered logistic regression results for group membership based on the association between the number of screens in the last survey and responding to a new survey.

	Odds ratio	95% CI – lower bound	95% CI – upper bound	P-value
Female	1.11	0.86	1.43	0.41
Age group				
• 18–29		Reference		
• 30–44	1.04	0.72	1.51	0.84
• 45–64	1.17	0.81	1.70	0.42
• 65 and older	1.21	0.78	1.89	0.40
Racial/ethnic minority	0.64	0.49	0.82	< 0.001
Married	0.99	0.76	1.30	0.95
High school or less	1.12	0.84	1.50	0.43
Household income				
• Less than USD 35,000		Reference		
• USD 35,000 – USD 59,999	0.78	0.56	1.08	0.13
• USD 60,000 – USD 99,999	0.87	0.61	1.25	0.46
• More than USD 100,000	0.81	0.56	1.18	0.27
Three or more people in household	1.13	0.87	1.46	0.35
Big five personality scores				
• Extroversion	1.19	1.05	1.35	0.008
• Agreeable	1.00	0.87	1.15	0.99
• Conscientiousness	0.93	0.81	1.07	0.33
• Neuroticism	0.92	0.79	1.06	0.23
• Openness	1.00	0.89	1.14	0.95

Note: The ordered logistic regressions explain the probability of being in one of the three groups. A significant odds ratio below one means that one is more likely to belong to a group where more screens decrease the response probability; if the odds ratio is significantly above one, the demographic category is more likely to belong to the group for whom more screens imply a higher likelihood of responding.

4. Discussion

This study examined the relationships between past survey burden and response propensity to a new survey in the UAS. The results show an overall positive relationship between questionnaire length of the last survey, measured by the number of screens, and the probability of response to a new survey. This finding may be explained by the relatively generous incentives in UAS, which compensate participants at a rate of USD 20 per thirty minutes for estimated completion time. This rate is about 5.5 times the U.S. minimum wage and about twice the average hourly earnings of employees on private nonfarm payrolls (U.S. Bureau of Labor Statistics n.d.). Further analysis in this article identified subgroups of participants whose response to a new survey was not or negatively related to the length of the last survey. Specifically, the results suggest that racial/ethnical minorities were less likely to complete a new survey following a prior long questionnaire. Our finding adds to the existing evidence showing that minorities are less likely to respond to re-interviews (Broman et al. 1995; Chen and Kandel 1995; Griffin 2002) and less likely to complete surveys in panels (Harris-Kojetin and Tucker 1998). Strategies to identify and address the challenges of conducting surveys in hard-to-survey populations including racial/ethnic

Table 3. Descriptive statistics and multiple-group comparisons based on the association between the number of surveys in the one-year period preceding a new panel survey and responding to a new survey.

	Group membership based on the association between the number of surveys in the one-year period preceding a new survey and responding to a new survey			
	More preceding surveys associated with decreased response (N = 152)	No association (N = 1052)	More preceding surveys associated with increased response (N = 126)	P-value of test for equality across columns
Female	83 (55.0%)	619 (55.0%)	71 (56.3%)	0.60
Age	46.12 (15.29)	46.68 (16.39)	48.51 (15.54)	0.42
Racial/ethnic minority	49 (32.2%)	380 (36.2%)	40 (31.7%)	0.42
Married	87 (57.2%)	561 (53.3%)	68 (54.0%)	0.66
High school or less education	41 (27.0%)	250 (23.8%)	31 (24.6%)	0.68
Household income				
• Less than USD 35,000	38 (25.0%)	282 (26.8%)	32 (25.4%)	0.86
• USD 35,000 – USD 59,999	38 (25.0%	302 (28.7%)	39 (31.0%)	0.52
• USD 60,000 – USD 99,999	31 (20.4%)	227 (21.5%)	28 (22.2%)	0.93
• More than USD 100,000	44 (28.9%)	234 (22.2%)	26 (20.6%)	0.15
Three or more people in household	86 (56.6%)	491 (46.7%)	48 (38.1%)	0.008
Big five personality scores				
• Extroversion	27.59 (6.10)	26.18 (6.26)	26.67 (6.23)	0.03
• Agreeable	36.80 (5.66)	36.19 (5.49)	36.70 (5.72)	0.33
• Conscientiousness	36.87 (5.79)	35.98 (5.65)	36.51 (5.63)	0.15
• Neuroticism	20.87 (6.96)	21.98 (6.55)	21.23 (7.08)	0.10
• Openness	37.01 (6.05)	36.59 (6.25)	36.64 (5.78)	0.74

minorities were discussed in depth by Harkness et al. (2010) and Tourangeau et al. (2014). Specific solutions to address the challenges of surveying racial/ethnic minorities included specially designed sampling frames (Duffy et al. 2006; Stueve et al. 2001; Heckathorn 2007), adaptation and translation (Harkness et al. 2010; Schroepfer et al. 2009; Lee and Schwarz 2014), culture-sensitive survey design (Sun 2012; Pennell et al. 2004; Harkness et al. 2003), and pretesting (Canales et al. 1995). Our analysis found that extroverted participants were more likely to complete a new survey following a prior long questionnaire compared with introverted participants. Prior research has suggested that extraversion was positively associated with participation in probability-based online panels (Saßenroth 2013). The finding may be explained by the relationships between extraversion and reward processing, which suggests that extroverts are more sensitive to rewards, often show greater engagement with rewarding stimuli and situations, and are more motivated to gain a promised reward (Smillie 2013; Lucas et al. 2000; Fu 2013; Furnham et al. 1999).

Table 4. Ordered logistic regression analysis result for the group membership based on the association between the number of surveys in the one-year period preceding a new survey and responding to a new survey.

	Odds ratio	95% CI – lower bound	95% CI – upper bound	P-value
	1.08	0.80	1.44	0.62
Female				
Age group				
• 18–29		Reference		
• 30–44	1.27	0.83	1.95	0.27
• 45–64	1.28	0.83	1.96	0.26
• 65 and older	1.38	0.83	2.32	0.22
Racial/ethnic minority	1.05	0.78	1.42	0.76
Married	0.96	0.70	1.32	0.81
High school or less education	0.87	0.62	1.22	0.42
Household income				
• Less than USD 35,000		Reference		
• USD 35,000 – USD 59,999	1.08	0.74	1.57	0.70
• USD 60,000 – USD 99,999	0.99	0.65	1.51	0.96
• More than USD 100,000	0.80	0.51	1.23	0.30
Three or more people in household	0.64	0.47	0.86	0.004
Big five personality scores				
• Extroversion	0.92	0.79	1.07	0.29
• Agreeable	0.98	0.83	1.14	0.76
• Conscientiousness	0.98	0.83	1.14	0.77
• Neuroticism	1.03	0.87	1.22	0.77
• Openness	0.98	0.85	1.14	0.82

Note: The ordered logistic regressions explain the probability of being in one of the three groups. A significant odds ratio below one means that one is more likely to belong to a group where more survey invitations decrease the response probability; if the odds ratio is significantly above one, the demographic category is more likely to belong to the group for whom more survey invitations imply a higher likelihood of responding.

This study detected no significant relationships between survey frequency in the one-year period preceding a new survey and the response to the new survey. This finding is consistent with evidence from ecological momentary assessment studies (Eisele et al. 2020; Walsh and Brinker 2015; Stone et al. 2003; Ono et al. 2019). Further analysis identified small subgroups of participants whose response was negatively associated with survey frequency. Specifically, the response of participants with three or more members in their households was more likely to be negatively associated with survey frequency. One possible explanation is that participants with large households can experience higher respondent burden especially if surveys contain questions about their household members (Walsh and Fields 2012). As a result, participants with a large household may be more likely to be overburdened when the frequency of surveys is high.

The results from this study have several implications for the management of an Internet panel. First, the results suggest that longer surveys may not be a problem when compensation is high enough and positively related to questionnaire length. However, the results should not be interpreted as supporting the use of extraordinarily long questionnaires. Surveys in the UAS are limited to a maximum of thirty minutes, so that

the results cannot be generalized beyond the thirty-minute time frame. Second, the results suggest that targeted interventions are needed to improve response of participants from minority groups. Panel managers may consider specially designed sampling frames for minorities and introduce culture-sensitive customizations into survey and website design. Third, although our findings suggest that incentives may work well for extroverts, they may work less well for introverts. Research has suggested that while the extroverts are more easily motivated to seek more rewards, the introverts are motivated to avoid punishment (Furnham et al. 1999). Fourth, our analysis suggests that it may not be a problem to engage participants in frequent surveys. In the UAS, participants respond to an average of about two surveys per month. Fifth, our analysis suggests that targeted interventions may be needed to manage the burden of participants with large households. Panel managers may consider higher compensation or breaking down a survey into shorter pieces for participants with a large household when the survey contains a relatively large number of questions about other household members. Lastly, the estimated associations between questionnaire length/survey frequency and response may be informative for identifying panel members at high risk of attrition. Future research may use machine learning techniques to examine whether the estimated impacts are strong predictors of attrition.

It is worth noticing that this study has several limitations. First, results from the present study may not be generalizable to online panels that are not probability-based. Second, the analysis is based on observational data only. The above suggested panel management strategies would need to be tested in randomized controlled trials. Third, although we have tried to choose explanatory variables that can be considered exogenous (e.g., the number of survey invitations, rather than the number of surveys one has answered, or the number of screens shown in previous surveys rather that the total time taken to respond to a survey), these variables can be refined in several ways. For instance, there is a certain seasonality in survey burden in the UAS (the Fall tends to generate the largest number of surveys), so that rather than taking the total number of surveys in the previous year as an explanatory variable, we may also consider shorter time periods to detect these possible peak affects. Another aspect that has not been analyzed yet is the nature of the surveys. We know from the UAS "end-of-year" surveys that respondents vary in the type of surveys they like most (or least). We can in principle group surveys by topic and use these as extra explanatory variables, possibly interacted with respondent characteristics. Fourth, all racial/ethnic minority groups were combined together into one group in the present ordinal regression analysis due to the limited sample sizes of each minority group. This limited our ability to examine the differences between specific minority groups. Finally, we did not carry out the analysis for all current and past panel members due to the computational burden of fitting the latent Markov chain model. Future research can expand the analysis to all current and past panel members.

In conclusion, this study suggests that experiences of longer questionnaires and frequent surveys may not lead to a decreased response propensity to a new survey for the majority of participants in a large probability-based panel. The study advocates targeted interventions for the small subgroups of participants whose response may be negatively impacted by longer questionnaires and frequent surveys.

5. References

Andrieu, C., A. Doucet, and R. Holenstein. 2010. "Particle Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (3): 269–342. DOI: https://doi.org/10.1111/j.1467-9868.2009.00736.x.

Baker, R., S.J. Blumberg, J.M. Brick, M.P. Couper, M. Courtright, J.M. Dennis, D. Dillman, M.R. Frankel, and P.Garland. 2010. "AAPOR Report on Online Panels." *Public Opinion Quarterly* 74 (4): 711–781. DOI: https://doi.org/10.1093/poq/nfq048.

Bartolucci, F., A. Farcomeni, and F. Pennoni. 2019. *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman and Hall/CRC.

Blumenberg, C., A.M.B. Menezes, H. Gonçalves, M.C. Formoso Assunção, F.C. Wehrmeister, F.C. Barros, and A.J.D. Barros. 2019. "The Role of Questionnaire Length and Reminders Frequency on Response Rates to a Web-Based Epidemiologic Study: A Randomised Trial." *International Journal of Social Research Methodology* 22 (6): 625–35. DOI: https://doi.org/10.1080/13645579.2019.1629755.

Bradburn, N. 1978. "Respondent Burden." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 35: 35–40. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y1978f.html (accessed August 2021).

Broman, C.L., V.L. Hamilton, W.S. Hoffman, and R. Mavaddat. 1995. "Race, Gender, and the Response to Stress: Autoworkers' Vulnerability to Long-Term Unemployment." *American Journal of Community Psychology* 23 (6): 813–842. DOI: https://doi.org/10.1007/BF02507017.

Canales, S., P.A. Ganz, and C.A. Coscarelli. 1995. "Translation and Validation of a Quality of Life Instrument for Hispanic American Cancer Patients: Methodological Considerations." *Quality of Life Research* 4 (1): 3–11. DOI: https://doi.org/10.1007/BF02507017.

Chen, K., and D.B. Kandel. 1995. "The Natural History of Drug Use from Adolescence to the Mid-Thirties in a General Population Sample." *American Journal of Public Health* 85 (1): 41–47. DOI: https://doi.org/10.2105/ajph.85.1.41.

Cheng, A., G. Zamarro, and B. Orriens. 2020. "Personality as a Predictor of Unit Nonresponse in an Internet Panel." *Sociological Methods & Research* 49 (3): 672–698. DOI: 10.1177/0049124117747305.

Crawford, S.D., M.P. Couper, and M.J. Lamias. 2001. "Web Surveys: Perceptions of Burden." *Social Science Computer Review* 19 (2): 146–162. DOI: https://doi.org/10.1177/089443930101900202.

Deutskens, E., K. de Ruyter, M. Wetzels, and P. Oosterveld. 2004. "Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study." *Marketing Letters* 15 (1): 21–36. DOI: https://doi.org/10.1023/B:MARK.0000021968.86465.00.

De Valpine, P., Daniel Turek, Christopher J. Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. 2017. "Programming with Models: Writing Statistical Algorithms for General Model Structures with NIMBLE." *Journal of Computational and Graphical Statistics* 26 (2): 403–413. DOI: https://doi.org/10.1080/10618600.2016.1172487.

Duffy, D., J. Goldberg, and D. Buchwald. 2006. "Using Mail to Reach Patients Seen at an Urban Health Care Facility." *Journal of Health Care for the Poor and Underserved* 17 (3): 522–531. DOI: https://doi.org/10.1353/hpu.2006.0104.

Eisele, G., H. Vachon, G. Lafit, P. Kuppens, M. Houben, I. Myin-Germeys, and W. Viechtbauer. 2020. "The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population." *Assessment*. DOI: https://doi.org/10.1177/1073191120957102.

Fan, W., and Z. Yan. 2010. "Factors Affecting Response Rates of the Web Survey: A Systematic Review." *Computers in Human Behavior* 26 (2): 132–139. DOI: https://doi.org/10.1016/j.chb.2009.10.015.

Fu, Y. 2013. "On the Nature of Extraversion: Variation in Conditioned Contextual Activation of Dopamine-Facilitated Affective, Cognitive, and Motor Processes." *Frontiers in Human Neuroscience* 7: 288. DOI: https://doi.org/10.3389/fnhum.2013.00288.

Furnham, A., L. Forde, and K. Ferrari. 1999. "Personality and Work Motivation." *Personality and Individual Differences* 26 (6): 1035–1043. DOI: https://doi.org/10.1016/S0191-8869(98)00202-5.

Galesic, M., and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73 (2): 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Griffin, D.H. 2002. "Measuring Survey Nonresponse by Race and Ethnicity." In Proceedings of the Annual Meetings of the American Statistical Association, August 2002: 11–15. Alexandria, VAL: American Statistical Association. Available at: https://nces.ed.gov/FCSM/pdf/IHSNG_asa02_DG.pdf (assessed October 2021)

Groves, R.M., E. Singer, and A. Corning. 2000. "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." *The Public Opinion Quarterly* 64 (3): 299–308. DOI: https://doi.org/10.1086/317990.

Guo, Y., J.A. Kopec, J. Cibere, L.C. Li, and C.H. Goldsmith. 2016. "Population Survey Features and Response Rates: A Randomized Experiment." *American Journal of Public Health* 106 (8): 1422–1426. DOI: https://doi.org/10.2105/AJPH.2016.303198.

Harkness, J.A., M.E. Braun, B.E. Edwards, T.P. Johnson, L.E. Lyberg, P.Ph. Mohler, B.-E.E. Pennell, and T.W. Smith. 2010. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, NJ: Wiley.

Harkness, J.A., F.J.R. van de Vijver, P.Ph. Mohler, and J. Wiley. 2003. *Cross-Cultural Survey Methods*. 325. Hoboken, NJ: Wiley.

Harris-Kojetin, B.A., and C. Tucker. 1998. "Longitudinal Nonresponse in the Current Population Survey (CPS)." In International Workshop on Household Survey Nonresponse, September 24, 1997: 263–272. Mannheim, Germany. Available at: https://www.ssoar.info/ssoar/bitstream/handle/document/49724/ssoar-1998-harris-kojetin_et_al-Longitudinal_nonresponse_in_the_Current.pdf?sequence = 1 = y = ssoar-1998-harris-kojetin_et_al-Longitudinal_nonresponse_in_the_Current.pdf (assessed October 2021).

Heckathorn, D.D. 2007. "6. Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment." *Sociological Methodology* 37 (1): 151–208. DOI: https://doi.org/10.1111/j.1467-9531.2007.00188.x.

Hill, D.H., and R.J. Willis. 2001. "Reducing Panel Attrition: A Search for Effective Policy Instruments." *Journal of Human Resources*: 416–38. DOI: https://doi.org/10.2307/3069625.

Koitsalu, M., M. Eklund, J. Adolfsson, H. Grönberg, and Y. Brandberg. 2018. "Effects of Pre-Notification, Invitation Length, Questionnaire Length and Reminder on Participation Rate: A Quasi-Randomised Controlled Trial." *BMC Medical Research Methodology* 18 (1): 1–5. DOI: https://doi.org/10.1186/s12874-017-0467-5.

Lee, S, and N. Schwarz. 2014. "Question Context and Priming Meaning of Health: Effect on Differences in Self-Rated Health between Hispanics and Non-Hispanic Whites." *American Journal of Public Health* 104 (1): 179–185. DOI: https://doi.org/10.2105/AJPH.2012.301055.

Lucas, R.E., E. Diener, A. Grob, E.M. Suh, and L. Shao. 2000. "Cross-Cultural Evidence for the Fundamental Features of Extraversion." *Journal of Personality and Social Psychology* 79 (3): 452. DOI: https://doi.org/10.1037//0022-3514.79.3.452.

Lugtig, P. 2014. "Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers." *Sociological Methods & Research* 43 (4): 699–723. DOI: https://doi.org/10.1177/0049124113520305.

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schütz. 2007. "Compensating for Low Topic Interest and Long Surveys: A Field Experiment on Nonresponse in Web Surveys." *Social Science Computer Review* 25 (3): 372–383. DOI: https://doi.org/10.1177/0894439307297606.

McCarthy, J.S., D.G. Beckler, and S.M. Qualey. 2006. "An Analysis of the Relationship between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative?" *Journal of Official Statistics* 22 (1): 97–112. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-analysis-of-the-relationship-between-survey-burden-and-nonresponse-if-we-bother-them-more-are-they-less-cooperative.pdf.

McCrae, R.R. 2004. "Human Nature and Culture: A Trait Perspective." *Journal of Research in Personality* 38 (1): 3–14. DOI: https://doi.org/10.1016/j.jrp. 2003.09.009.

McCrae, R.R., and P.T. Costa Jr. 2008. "The Five-Factor Theory of Personality." Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions: 91-104. Hoboken, NJ: Wiley.

Nicoletti, C., and F. Peracchi. 2005. "Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel." *Journal of the Royal Statistical Society:* 168 (4): 763–781. DOI: https://doi.org/10.1111/j.1467-985X.2005.00369.x.

Oakland, R.J., and J.S. Oakland. 2018. *Statistical Process Control*. London, UK: Routledge.

Ono, M., S. Schneider, D.U. Junghaenel, and A.A. Stone. 2019. "What Affects the Completion of Ecological Momentary Assessments in Chronic Pain Research? An Individual Patient Data Meta-Analysis." *Journal of Medical Internet Research* 21 (2): e11398. DOI: https://doi.org/10.2196/11398.

Pennell, B.-E., A. Bowers, D. Carr, S. Chardoul, G.-Q. Cheung, K. Dinkelmann, N. Gebler, S.E. Hansen, S. Pennell, and M. Torres. 2004. "The Development and Implementation of the National Comorbidity Survey Replication, the National Survey of American Life, and the National Latino and Asian American Survey." *International Journal of Methods in Psychiatric Research* 13 (4): 241–269. DOI: https://doi.org/10.1002/mpr.180.

Saßenroth, D. 2013. "Personality Effects on Participation in the LISS Panel." In *The Impact of Personality on Participation Decisions in Surveys*, 129–160. Berlin, Germany: Springer.

Schroepfer, T.A., J. Matloub, P. Creswell, R. Strickland, and D.M. Anderson. 2009. "A Community-Specific Approach to Cancer Research in Indian Country." *Progress in Community Health Partnerships: Research, Education, and Action* 3 (4): 317. DOI: https://doi.org/10.1353/cpr.0.0096.

Sharp, L.M., and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 47 (1): 36–53. DOI: https://doi.org/10.1086/268765.

Smillie, L.D. 2013. "Extraversion and Reward Processing." *Current Directions in Psychological Science* 22 (3): 167–172. DOI: https://doi.org/10.1177/0963721412470133.

Stone, A.A., J.E. Broderick, J.E. Schwartz, S. Shiffman, L. Litcher-Kelly, and P. Calvanese. 2003. "Intensive Momentary Reporting of Pain with an Electronic Diary: Reactivity, Compliance, and Patient Satisfaction." *Pain* 104 (1–2): 343–351. DOI: https://doi.org/10.1016/s0304-3959(03)00040-x.

Stueve, A., L.N. O'Donnell, R. Duran, A. San Doval, and J. Blome. 2001. "Time-Space Sampling in Minority Communities: Results with Young Latino Men Who Have Sex with Men." *American Journal of Public Health* 91 (6): 922. DOI: https://doi.org/10.2105/ajph.91.6.922.

Sun, H. 2012. *Cross-Cultural Technology Design: Creating Culture-Sensitive Technology for Local Users.* Oxford, UK: Oxford University Press

Tourangeau, R., B. Edwards, T.P. Johnson, K.M. Wolter, and N. Bates. 2014. *Hard-to-Survey Populations*. Cambridge, UK: Cambridge University Press.

U.S. Bureau of Labor Statistics. n.d. *Average Hourly and Weekly Earnings of All Employees on Private Nonfarm Payrolls by Industry Sector, Seasonally Adjusted*. Available at: https://www.bls.gov/news.release/empsit.t19.htm (Assessed May 2021)

UAS (Understanding American Study). Available at: https://uasdata.usc.edu/index.php (accessed October 2022).

Venables, W.N., and B.D. Ripley. 2013. *Modern Applied Statistics with S-PLUS*. Berlin, Germany: Springer Science & Business Media.

Walsh, E., and J.K. Brinker. 2015. "Temporal Considerations for Self-Report Research Using Short Message Service." *Journal of Media Psychology* 28 (4): 200–206. DOI: https://doi.org/10.1027/1864-1105/a000161.

Walsh, R., and J.M. Fields. 2012. "Reducing Respondent Burden: Evaluating the Progress of the SIPP-EHC." In *Federal Committee on Statistical Methodology 2012 Biennial Conference. FCSM, Washington, DC*. Available at: https://nces.ed.gov/FCSM/pdf/-Walsh_2012FCSM_IX-B.pdf (Assessed October 2021).

Yetter, G., and K. Capaccioli. 2010. "Differences in Responses to Web and Paper Surveys among School Professionals." *Behavior Research Methods* 42 (1): 266–272. DOI: https://doi.org/10.3758/BRM.42.1.266.

Zabel, J.E. 1998. "An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior." *Journal of Human Resources*: 479–506. DOI: https://doi.org/10.2307/146438.

Received January 2021
Revised June 2021
Accepted October 2021

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1069–1095, http://dx.doi.org/10.2478/JOS-2022-0046

The Effects of Response Burden – Collecting Life History Data in a Self-Administered Mixed-Device Survey

Johann Carstensen[1], *Sebastian Lang*[1], *and Fine Cordua*[1]

Collecting life history data is highly demanding and therefore prone to error since respondents must retrieve and provide extensive complex information. Research has shown that response burden is an important factor influencing data quality. We examine whether increases in different measures of response burden in a (mixed-device) online survey lead to adverse effects on the data quality and whether these effects vary by the type of device used (mobile versus non-mobile).

We conducted an experimental study in an online mixed-device survey, for which we developed a questionnaire on the educational and occupational trajectories of secondary-school graduates, undergraduates, and university graduates. To address our research question, we randomly assigned different levels of response burden to the participants and compared different measures on the data quality and response.

We found mixed evidence for unfavourable effects of response burden on the examined outcomes. While some of our results were expected, they were not consistent across all subgroups. Most interestingly, the effects of response burden on outcomes seemed to differ based on the device used. Hence, we conclude that further research is needed to optimise the collection of complex data from different groups of participants.

Key words: Respondent burden; data quality; response behavior; experiment; online survey.

1. Introduction

In attempting to reduce bias in survey results, researchers have increasingly turned their attention to response or respondent burden. Response burden can introduce bias through unit (Groves et al. 2011) or item nonresponse (Galesic and Bosnjak 2009), as well as through measurement error (Peytchev and Peytcheva 2017). Thus, survey design to reduce response burden must be carefully balanced with the substantial needs of a data collection endeavour.

Recently, many survey-makers have strived to collect increasingly complex information using web surveys as the most cost-effective survey mode. However, collecting complex life history data – that is, detailed biographical data suited for complex analysis using for instance event history models – in a self administered mode poses new challenges to minimising measurement error as there is no interviewer who can provide help and motivation. One of these challenges concerns reducing response burden.

[1] German Centre for Higher Education Research and Science Studies, Lange Laube 12, 30159 Hannover, Germany. Emails: carstensen@dzhw.eu, lang@dzhw.eu and cordua@dzhw.eu

Acknowledgments: This research was supported by the German Federal Ministry of Education and Research (16P4246). The authors would like to thank the editors and anonymous reviewers as well as the participants of the 8th Conference of the European Survey Research Association and the joint meeting of the Association of Social Science Institutes (ASI) and the DGS section "Methods of Empirical Social Research" for helpful comments on earlier versions of this article.

Multiple factors can influence response burden, such as varying the length or frequency of a survey or the amount and complexity of the requested information. In this experimental study, we vary both the amount and complexity of items regarding life history, to determine how different measures of data quality are influenced by those variations. Using a web survey, we also distinguish survey respondents who use mobile devices from those using desktop computers, to identify heterogeneous effects.

We start by elaborating on our definition of response burden. This serves as the basis to introduce Tourangeau's (2000, 2018) response process approach, which can explain why participants experience various difficulties when answering survey questions. Additionally, satisficing theory (Krosnick 1991) is used to explain participants' reactions on an increasing response burden when engaged in reporting. We then proceed into a review of recent research on the factors that foster response burden. The methods used and the study design are then combined with information from recent research to develop testable hypotheses. We reveal our strategy and describe our data (Section 2) before presenting our results (Section 3). We conclude with a discussion and summary (Section 4).

1.1. Theory

According to Bradburn (1979, 36), response burden' is not an objective characteristic of the task, but is the product of an interaction between the nature of the task and the way in which it is perceived by the respondent.' He identified four key elements of respondent burden (Bradburn 1979, 36): (1) the length of the interview; (2) the amount of effort required of the respondent; (3) the amount of stress on the respondent; and (4) the frequency with which the respondent is interviewed. Three of these elements can be seen as properties of a survey, while the amount of stress on the respondent is a product of the interaction between the survey contents and characteristics of the respondent. While more recent research has often relied on a respondent-centred approach in defining response burden (Hedlin et al. 2008), we focus on properties of a survey, varying them experimentally.

To understand the impact of response burden on the response process, response rates, and data quality, we refer to Tourangeau's (2000, 2018) cognitive response process model. Based on Tourangeau (2000, 2018), participants' response process when answering a survey can be summarised in four steps. First, respondents must **comprehend** the posed question in a way that fits the researchers' conceived concept. Second, they must **retrieve** information from their memory that enables them to answer the question. Third, **judgement and estimation** take place: the respondents can estimate an answer (e.g. the frequency of an event in a specific time period), count, rate, or rely on existing information. Fourth, the respondents **report** the answer. Regarding response burden – as defined above – error may be introduced in this process in different manners.

While **comprehension** error is, by definition, connected to the (subjective) difficulty of a task, it is also more likely to occur due to fatigue in longer interviews and when more detailed information is asked. This is true for all four steps of the cognitive process. Moreover, when increasing the amount of detail in a survey, it becomes necessary to be increasingly specific in formulating the question or instruction. This consequently increases the likelihood that the researchers' intended meaning and the respondents' understanding of a question do not match. Specifically in relation to collecting life history

data, this can happen, for example, in identifying timespan or granularity ("daily", "monthly", "yearly") in reporting events.

Retrieval may be hampered when respondents must recall less salient information, for example when having to report detailed information on long time spans. Fatigue may furthermore increase retrieval errors. Such errors may occur due to forgetting or forward telescoping.

The same general arguments apply to the third step, **judgement and estimation**: the probability of errors is also likely to increase here if more, very detailed, or difficult information is requested, or if this information lies further in the past, in which case respondents are more likely to have to guess or estimate their answers instead of clearly recalling details. Respondents' strategies in this step may vary between low and high effort, depending on the time they have (Burton and Blair 1991). Besides the tendency that available time will decrease in a longer survey, fatigue may lead to choosing low-effort strategies.

Finally, respondents may introduce error in the fourth step, **reporting**. Again, in addition to effects through fatigue, requesting more (and more detailed) information or asking more difficult questions should increase the probability of errors in reporting, as the task of mapping information to scales provided by researchers becomes more complex. In collecting life history data, this could occur when researchers ask, for instance, about changes in the kind of employment contract and require respondents to be more specific about the types of such contracts than they normally would be; this could represent a mismatch between the scale provided and the information respondents would be able to remember.

1.2. State of Research

The *length of the survey* (Bradburn 1979) and the *amount of detail* asked (Warriner 1991) can complicate the process of retrieval because the respondents need to concentrate for a longer time (Crawford et al. 2001; Galesic 2006; Galesic and Bosnjak 2009; Marcus et al. 2007). Relatedly, the *time needed* to participate is found as a factor that increases response burden (Crawford et al. 2001; Galesic 2006; Galesic and Bosnjak 2009; Marcus et al. 2007) and self-reported response burden increases with interview duration (Guin et al. 2012; Yan et al. 2020). The *reference period* has a major influence on participants' response burden. The farther back in the past an event lies, the more poorly it is likely to be remembered (Groves et al. 2011; Tourangeau 2000), and the more uncertainty it can create, for instance when answering questions requires comprehensive processes of recalling memories (Drasch and Matthes 2013). Relatedly, the *difficulty of the item* is another element of the response burden (Galesic and Bosnjak 2009; Hambleton et al. 1991; Reise et al., 2005; Zanon et al. 2016). The *more difficult* it is for the participants to read and understand an item, the *more effort is required to retrieve the information demanded and judge or estimate an answer*, and the lower the participants' *cognitive capacity* is, the higher their response burden is likely to be (Guin et al. 2012; Gummer and Roßmann 2015; Lenzner 2012; Tourangeau 2003; Yan and Tourangeau 2008).

Questions of higher difficulty can lead to difficulties in retrieval (e.g. forgetting events), and can thus produce errors in reporting (Tourangeau 2000). Such errors can also appear due to respondents' tactics to minimise their response burden. Recent research fruitfully explains respondents' reactions to high response burden by applying decision field theory

(e.g. Galesic 2006). 'When confronted with a difficult personal decision, the decision maker tries to anticipate and evaluate all of the possible consequences produced by each course of action.' (Busemeyer and Townsend 1993, 444). Answering questions in a survey can be seen as such a 'difficult personal decision'.

To avoid high effort, respondents may react to a high response burden by *satisficing*. As satisficing theory states, this may lead to incomplete or biased information. Strategies may include wrongly agreeing or disagreeing with statements (e.g. by choosing the first answer regardless of its content), avoiding socially critical statements, or giving no answer (e.g. by choosing the category 'I don't know') (Krosnick 1991). As another means to reduce effort, participants may react with *underreporting frequencies* of events (Kaper and Saris 1999) and thus – more or less intentionally – answering questions in a way that shortens the interview (Peytchev and Peytcheva 2017).

These results seem to be especially relevant when deciding how to go about collecting biographical information. For instance, *looping questions* are questions that are asked several times in the course of the survey and refer to recurring events (e.g. gainful employment), where the frequency of repetition depends on the frequency of the event. This sort of questions is often used in collecting life history data and is also implemented in the questionnaire on which this study is based. Depending on their implementation, looping questions can either lead to inaccurate numbers in event reports but detailed information on the reported events (when asking for subsequent events after each report of the preceding event) or to accurate numbers in event reports but missing information on the details of the reported events (when first asking for the frequency k of an event and then asking for details on each of the reported k events) (Eckman and Kreuter 2018).

Furthermore, respondents can react by not answering parts of the questionnaire (*item-nonresponse*, see Galesic and Bosnjak 2009) or by refusing to answer the interview completely (*unit-nonresponse,* see Groves et al. 2011). Finally, a burdensome survey is discussed to cause *panel attrition*, meaning that it reduces the participant's willingness to participate in future surveys, for instance because the preceding interview experience was frustrating (Hoogendoorn and Sikke 1998; Sharp and Frankel 1983).

Findings on the *device* participants used to answer web surveys have been inconsistent. Some studies have found no evidence for the assumption that mobile, compared to non-mobile devices, could lead to a higher response burden concerning dropout, item nonresponse, the time needed to complete the survey, or the amount of detail an answer contained (Antoun et al. 2017; De Bruijne and Wijnant 2014; Marcus et al. 2007; Toepoel and Lugtig 2014). Others have found no tendencies for respondents to agree regardless of the content (acquiescence), to provide non-substantive answers (congruent with Mavletova 2013), to choose midpoint values or the first response option (primacy effects; congruent with Mavletova 2013), or to choose answers depicted at the same position in the survey (straightlining; Clement et al. 2020), when using mobile devices. Moreover, Daikeler et al. (2020) have not identified more misreporting and Krebs and Höhne (2020) have not identified stronger scale direction effects, when mobile devices have been used. At the same time, researchers have observed that respondents who used mobile phones in a previous survey tended not to use them in the next survey they took (Lugtig and Toepoel 2016). Additionally, they produced more missing data (Keusch and Yan 2017) and showed lower response rates (Lee et al. 2019; Mavletova 2013). This could indicate a tendency for

respondents using mobile devices to experience a higher response burden (Buskirk and Andrus 2014).

Apart from one study that reported lower response time for mobile users (Lee et al. 2019), studies were found to report higher durations in survey completion when respondents used smartphones, compared to stationary computers (Antoun et al. 2017; De Bruijne and Wijnant 2014; Schlosser and Mays 2018). This could be partly ascribed to the need for additional scrolling (Couper and Peterson 2017). The time needed to answer on mobile devices seemed to be less pronounced for respondents who were familiar with them (Antoun and Cernat 2020). Also, participants using mobile devices were more likely to be outside their homes and among other people, restricting the attention they paid to the survey (Antoun et al. 2017; De Bruijne and Oudejans 2015).

Collecting life history data is associated with especially high response burden, as all the above mentioned factors are involved. Posing questions towards several episodes in the respondents' lives seems to inevitably lead to longer surveys than those limited to particular episodes, as the reference period tends to cover events that lie farther back in the respondents' memories. Also, the amount of detail asked tends to be high, as connections between the reported episodes are enquired. Answering questions that encompass life history data can thus turn out to be rated as difficult for the respondents. Such factors likely lead to an increase in the time needed to respond to life history questions.

Starting from education and leading up to the labour market, career paths are highly dependent on preceding episodes in life. Hence, collecting data within a life history framework seems especially promising to gain a comprehensive understanding of how students decide for or against different paths. With this in mind, the present article investigates response burden when collecting life history data from students and employees.

1.3. Hypotheses

In the following, we describe our study in the light of reseach gaps revealed by the literature. Thereafter, we derive our hypotheses.

First, we check whether our treatments have the expected effect on one indirect and one direct measure of response burden: response time and the subjective assessment of the questionnaire's difficulty (see H1a and H1b below). Especially the latter measurement of response burden has not been common in recent research. Many studies have relied on single variables, such as response time, as a measure for response burden only (e.g. Gummer and Roßmann 2015; Yan and Tourangeau 2008).

Second, we use several variables that measure the impact a high response burden should have on participants' behaviour (see H2a to H3c below). Conversely, many studies have focused on one effect response burden has on answering behaviour (e.g. Galesic 2006).

Third, only one study has focused on life history data. This study did not investigate response burden as such, but rather participants' recollections when filling out the survey (Drasch and Matthes 2013). Our study explicitly investigates the impact of response burden on life history data, since such data provides encompassing insights into respondents' processes of decision-making and behaviour. Additionally, we vary relevant variables that should impact response burden (such as the reference period and amount of detail asked; see Subsection 2.1 for more information).

Fourth, research has lacked considerations on participants' behaviour in self-administered web surveys, as many surveys were conducted when the internet was not as established in mainstream use. We offered participants in our survey a choice of which device to use (see H4 below). Altogether, this study provides important insights not only regarding the effects of response burden on life history data, but also regarding the possible moderation effects of devices on response burden.

In this article, we investigate the impact of response burden on several dimensions relating to response behaviour and data quality. We begin by analysing the effect of our experimental treatments on one indirect and one direct measure of response burden. These are response time (H1a) and the subjective evaluation of the difficulty of the questionnaire (H1b). Both should be higher when increasing response burden.

With respect to response behaviour, we use break-offs and panel consent as indicators. Based on our theoretical approach and the state of research presented in the previous section, increasing response burden should move respondents closer to or beyond their specific critical threshold for satisficing. Analogously, this should apply to the refusal of panel consent. Therefore, we assume that an increase in response burden leads to an increase in the probability of the interview being ended before completion (H2a) and a decrease in the probability that the respondent will consent to further interviews (H2b).

The same theoretical argument holds for data quality, with incorrect or insufficient information as other forms of satisficing. We must make some presumptions with respect to this, as we cannot exploit external data to verify the information the respondents provide, such as linked or linkable administrative data. These presumptions are as follows:

1. A higher share of months with no information on activities indicates poorer data quality,
2. A higher number of episodes indicates higher data quality, and
3. Less item non-response (INR) indicates higher data quality.

The simple argument behind all three presumptions is that more information is better. This applies especially for presumption (1), as the alternative would be no information on some periods. Regarding presumptions (2) and (3), we further argue that there is no plausible reaction to response burden, such as satisficing, that would lead to a higher number of episodes or less INR.

In general, we expect a decrease in the data quality with an increase in response burden. Based on these presumptions, we argue that a higher response burden should result in a lower number of episodes (H3a), a higher share of months with no information (H3b), and a higher number of INR (H3c).

Finally, we consider the heterogeneous state of research on mobile devices. Due to the inconsistent results of previous research on this topic, we expect heterogeneous effects depending on the respondents' choice of device. Overall, we think that arguments in favour of generally stronger effects of a higher response burden on mobile devices prevail. Therefore, we expect to find stronger effects of response burden for the participants who decided to answer the survey on a mobile device (H4).

2. Methods

2.1. Study Design and Data

This work is part of a project to prepare the Student Life Cycle Panel (SLC) of the German Centre for Higher Education Research and Science Studies (DZHW). For this project, we developed a questionnaire to collect life history data from secondary-school graduates, undergraduates, and university graduates in a self-administered web survey. As providing life history data is highly demanding for participants, the topic of response burden is of special interest in this context. To investigate the impact of response burden on participation, break-offs, and data quality, we conducted an experiment in which we varied response burden in the survey module for collecting life history data.

Our sample consists of four subsamples with separate sampling frames and sampling procedures: (1) (former) students, (2) school-leavers from vocational schools, (3) international students, and (4) non-traditional students.

Subsample (1) stems from a probability-based online access panel of (former) students maintained by the DZHW. Subsample (2) was drawn from vocational schools in North Rhine-Westphalia to include school-leavers qualified to enter higher education but who took vocational training or started to work immediately after leaving school. These respondents constitute an important comparison group for higher education research. Therefore, we paid special attention to ensuring the adequacy of our questionnaire for this group.

Subsamples (3) and (4) were recruited at a German university and represent two small subgroups with rather low participation rates in regular student samples. Like the members of subsample (2), they are of special interest for higher education research and could have specific requirements for the questionnaire.

For our analyses, we pooled these subsamples and controlled for the origin of the observations.

The participants were invited by e-mail, and received two reminders (each at intervals of three to four weeks). Participation was incentivised with a lottery of one big prize (Apple iPad Pro, 10.5″, 64GB), plus 2 X EUR 100, and 10 X EUR 50-book vouchers. The field phase ended on January 17, 2019. The overall response rates were 8.4% (RR1) and 8.6% (RR2; American Association for Public Opinion Research 2016).

The questionnaire consisted of three main parts. The first and the last part contained about 50 questions on socio-demographics and the respondents' current activities. The middle part was the life history module, where the experimental treatments were implemented. Two experimental variations of the response burden were made in a fully-crossed, 2 X 3 design. (For an overview on the distribution of observations in experimental treatments and types of devices, see Figure 1.)

First, the reference period for which life events should be reported was varied between two and five years. Second, the amount of detail the respondents were asked to provide in the module on dependent employment was varied. We implemented a minimum, medium, and maximum information condition, where the request for additional information in these modules should represent an incremental increase in the response burden, as the medium and maximum condition require more information, each with an increasingly high resolution. (For further details about the information asked in each version, see Table 1.)

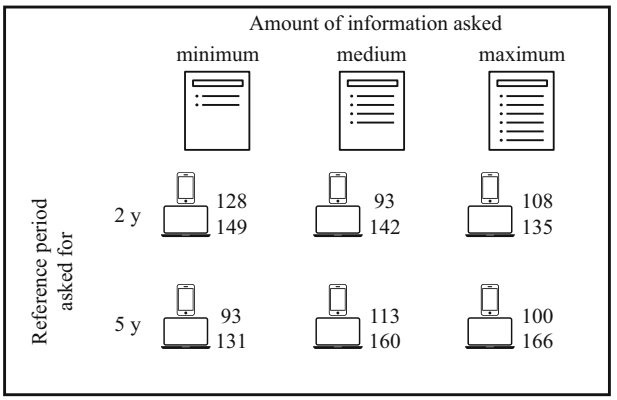


Fig. 1. Number of respondents by experimental treatment and device used to complete the survey.
Note: The experimental treatments (i.e. reference period asked for and the amount of information asked for) were randomly assigned. Device used to complete the survey (i.e., mobile versus non-mobile) was not randomly assigned.

Table 1. Information asked in minimum, medium, and maximum information condition.

	Minimum information condition	Medium information condition	Maximum information condition
Criteria splitting episodes into sub-episodes	- employer - occupation - occupational status - place of work	- employer - occupation - occupational status - place of work	- employer - occupation - occupational status - place of work
Additional information (asked chronologically for existing episodes)	- temporary contract (yes/no) - employment contract - working time (only starting month) - income (only starting month) - economic sector	- working time - income - economic sector	- economic sector
Additional information asked (event-centred, every change tracked)		- temporary contract (yes/no) - employment contract	- temporary contract (yes/no) - employment contract - working time - income

The majority of the life history module consisted of interleafed looping questions (e.g. Eckman and Kreuter 2018) asking, for example, for the first employment in the reference period, then for details on this employment, and then for the next employment. For an independent random sample of the respondents, the looping questions were preceded by a life history calendar. The information from this matrix-like graphical calendar was used for routing the respondents into the relevant looping questions afterwards. This experiment is not part of

the presented article (see Lang and Carstensen 2022 for further details). Beyond employment episodes, structured and reported by the respondents, we asked for changes in the 'criteria splitting episodes into sub-episodes' (see Table 1), initialising further loops. This means, for instance, that an employment episode reported from January 2019 to December 2019 was split into two sub-episodes if the respondent indicated that they changed the work location in June 2019. All further information was then recorded for the resulting two sub-episodes. This and the sort of information asked were identical in the minimum, medium, and maximum information condition. To vary the response burden, the resolution of the requested information increased: we asked the respondents to indicate all changes within the loops for parts of the information (information is listed in the last row of Table 1, instead of the second row).

We decided to make the survey completely responsive (instead of a mobile-first design), meaning that the layout of the survey adapted to the devices of the respondents. For example, the lines were individually wrapped according to the available screen width and matrix questions switched to a carousel layout.

A total of 2,016 people responded to the questionnaire. Since we were also interested in break-offs in the life history module, we excluded only partials that dropped out before the life history module. Excluding such cases and those of respondents who did not consent to the usage of paradata left 1,518 cases to be analysed.

Our analyses encompass seven dependent variables. Since some models include respondents who dropped out of the questionnaire after they began the life history module, case numbers differ between the models. The panel consent could only be asked for a very small subset of the sample: those who were not already part of the access panel and did not terminate the survey earlier. In checking if the randomisation was successful, Tables 2 and 3 show descriptive statistics separately by the groups of both treatments.

Whether the composition of these observed properties differs by treatment group (p-value) was determined by OLS regression for metric and average marginal effects of logistic regressions for categorical variables. Surprisingly, the share differs between the groups of the detail treatment for one of the categories of current activities. We assume that this is a coincidence, since the current activity was asked at the very beginning of the questionnaire and respondents were only excluded when they broke off before the life history module, where the treatments first affected the response process. By controlling for these properties in our models, we compensate for this random error.

Table 4 presents an overview of the same compositional characteristics by type of device. Since this could not be randomised, differences in the composition by device are expected. Overall, there are no counterintuitive differences in this sense. There are two significant differences in the composition of these groups by subsample as well.

2.2. *Measurements*

This section documents how we operationalise each dependent and independent variable. In the independent part, all models include dummy variables for the treatment conditions and for mobile devices (1 = mobile device, 0 = other device), as well as interaction terms between the treatment conditions and the device variable. The treatment conditions were independently and randomly assigned, and the device was measured by coding the user agent string from our paradata. The user agent string was searched for "mobi", resulting in smartphones and tablets

Table 2. Composition by reference period treatment.

Variable	Two years				Five years				Difference mean two years versus five years
	Obs	Min/ Max	Mean	Std. Dev.	Obs	Min/ Max	Mean	Std. Dev.	P-Value
Mobile device	755	0/1	0.44	0.50	763	0/1	0.40	0.49	0.170
Current activity									
Student	755	0/1	0.49	0.50	763	0/1	0.51	0.50	0.607
Vocational training	755	0/1	0.06	0.24	763	0/1	0.04	0.20	0.056
Employed	755	0/1	0.43	0.50	763	0/1	0.44	0.50	0.774
PhD student	755	0/1	0.08	0.28	763	0/1	0.09	0.28	0.904
Other	755	0/1	0.15	0.36	763	0/1	0.19	0.39	0.067
Sample	755				763				
Access panel: (former) students		0/1	0.81	0.40		0/1	0.83	0.37	0.152
Vocational schools: school leavers		0/1	0.07	0.26		0/1	0.07	0.25	0.719
University: international students		0/1	0.07	0.26		0/1	0.07	0.25	0.574
University: non-traditional students		0/1	0.05	0.22		0/1	0.03	0.18	0.115
Gender: female	564	0/1	0.57	0.50	580	0/1	0.56	0.50	0.676
Age	467	17-60	28.34	5.57	438	18-56	28.3	5.72	0.908

Table 3. Composition by level of detail treatment.

Variable	Minimum detail				Medium detail				Difference mean med. to min.	Maximum detail				Difference mean max. to min.
	Obs	Min/ Max	Mean	Std. Dev.	Obs	Min/ Max	Mean	Std. Dev.	P-Value	Obs	Min/ Max	Mean	Std. Dev.	P-Value
Mobile device	501	0/1	0.44	0.50	508	0/1	0.41	0.49	0.252	509	0/1	0.41	0.49	0.296
Current activity														
Student	501	0/1	0.51	0.50	508	0/1	0.47	0.50	0.245	509	0/1	0.52	0.50	0.806
Vocational training	501	0/1	0.07	0.25	508	0/1	0.04	0.20	0.088	509	0/1	0.04	0.20	0.087
Employed	501	0/1	0.40	0.49	508	0/1	0.43	0.49	0.401	509	0/1	0.47	0.50	0.024
PhD student	501	0/1	0.09	0.28	508	0/1	0.1	0.31	0.316	509	0/1	0.07	0.25	0.255
Other	501	0/1	0.17	0.38	508	0/1	0.19	0.39	0.522	509	0/1	0.15	0.36	0.426
Sample	501				508					509				
Aaccess panel: (former) students		0/1	0.81	0.39		0/1	0.81	0.39	0.852		0/1	0.83	0.37	0.347
Vocational schools: school leavers		0/1	0.08	0.27		0/1	0.07	0.26	0.673		0/1	0.06	0.24	0.234
University: international students		0/1	0.08	0.27		0/1	0.06	0.24	0.293		0/1	0.07	0.25	0.58
University: non-traditional students		0/1	0.03	0.18		0/1	0.05	0.22	0.134		0/1	0.04	0.19	0.65
Gender: female	391	0/1	0.56	0.50	375	0/1	0.58	0.49	0.504	378	0/1	0.54	0.50	0.673
Age	320	17-51	28.09	5.43	287	18-60	28.18	5.88	0.838	298	18-51	28.71	5.62	0.172

Table 4. Composition by type of device.

Variable	Non-mobile				Mobile				Difference mean mobile versus non-mobile
	Obs	Min/ Max	Mean	Std. Dev.	Obs	Min/ Max	Mean	Std. Dev.	P-Value
Current activity									
Student	883	0/1	0.50	0.50	635	0/1	0.50	0.50	0.842
Vocational training	883	0/1	0.04	0.18	635	0/1	0.07	0.26	0.001
Employed	883	0/1	0.42	0.49	635	0/1	0.44	0.50	0.462
PhD student	883	0/1	0.11	0.31	635	0/1	0.05	0.22	0.000
Other	883	0/1	0.18	0.38	635	0/1	0.16	0.36	0.261
Sample	883				635				
Access panel: (former) students		0/1	0.85	0.36		0/1	0.78	0.42	0.000
Vocational schools: school leavers		0/1	0.04	0.20		0/1	0.11	0.31	0.000
University: international students		0/1	0.07	0.26		0/1	0.06	0.24	0.422
University: non-traditional students		0/1	0.04	0.18		0/1	0.05	0.22	0.114
Gender: female	669	0/1	0.52	0.50	475	0/1	0.62	0.49	0.001
Age	519	18-60	28.85	5.50	386	17-51	27.61	5.76	0.001

coded as mobile devices. In addition, all models are controlled for another experiment, including the (independent, random) assignment of a life history calendar in advance of the life history module itself, which is used to filter respondents into relevant detail questions. This experiment is not part of the presented article, but we include the treatment conditions as control variable to account for possible confounding through an effect on response burden as well as our dependent variables. (For results on the use of a life history calendar in a self-administered mixed-device online survey, see Lang and Carstensen 2022). Beyond that, all models are controlled for current activity (dummy variables for studying, vocational training, employment, PhD, others); sample (dummy variables for the samples described above); gender (1 = female, 0 = male); and age (in years).

As outlined in Subsection 1.2, the first analyses are devoted to testing hypotheses H1a and H1b and the question of whether the treatments lead to higher values on rather direct measures of response burden. Response time is measured in minutes and calculated from the paradata of our web survey as the sum of minutes spent on each page of the questionnaire. This time was corrected for interruptions, which are defined as a duration of more than five minutes on one page. The corrected response time is calculated as the total time spent in the questionnaire minus the time exceeding five minutes on each page. An exception was made for pages including a life history calendar (LHC). In this case, the threshold was set to ten minutes. For the other dimension, the subjective difficulty of the questionnaire, respondents were asked on a seven-point endpoint labelled "Likert scale" (0 = very easy to 6 = very difficult): "How difficult was it for you to provide the information on your background that we asked you about in this survey?"

For the second group of hypotheses on response behaviour, we investigate the probability that a respondent would break off the survey, and the probability that the participant would provide panel consent for further investigations. "Breaking off" is defined as terminating the questionnaire without resuming it before the end of our field period (0 = completion, 1 = break-off). The panel consent was asked for at the end of the questionnaire, on the same page as the contact details for the lottery (0 = no, 1 = yes). Unfortunately, respondents from subsample (1) could not be asked for their consent, as they were already part of the access panel.

For the third group of hypotheses, we investigate the share of months with no information, the number of episodes, and the number of INR as indicators for data quality. For the share of months without information, we use data from the life history calendar (if applicable) and the life history module. We count every month as being "with no information" if there is no episode that includes that month, regardless of the type of episode. This count serves as the nominator, while the denominator is the maximum number of months for which one respondent could provide information. The number of episodes is the simple count of all episodes a respondent provided, regardless of the type of episode; and the number of INR is a simple count of all items where answers are missing. Besides the variables mentioned above, these models include dummy variables that indicate the point of break-off in the life history module (if applicable), or the logarithmised total number of items each respondent received.

2.3. Analysis Strategy

To test our hypotheses, we run several regression analyses, each adapted to a single hypothesis. We model the reference period and the information treatment separately for

every dependent variable, as we are interested in the single effects of the treatments. Otherwise, we would have to apply the same sample restrictions on the models for the reference period as on the models for the information treatment. Due to the fully-crossed design and independent randomisation, this is not a problem.

We analyse response time, subjective difficulty, dropout, panel consent, number of episodes, share of months without an episode, and the number of INR using poisson, ordinary least squares (OLS), logistic, fractional, double hurdle, and negative binomial regression models. All models are controlled for current activity, sample, gender, and age. Those control variables are necessary because the treatment variables are randomised but the use of mobile versus non-mobile devices to answer the survey could not be allocated at random and former research reported that younger and female participants are more likely to use mobile devices (see De Bruijne and Wijnant 2014; Gummer and Roßmann 2015; Gummer et al. 2019; Sommer et al. 2017). Since we want to check whether the effects of our treatments differ between mobile and non-mobile devices, we thereby try to control for selection bias deriving from device-specific preferences. However, we cannot state with certainty that no selection bias remained after controlling for these variables.

For the models on response burden measured by interview duration and subjective difficulty, we apply poisson regression to implement a log-level-model, and OLS for a linear regression model. To test our hypotheses on response behaviour measured by dropout and panel consent, we run logistic regression models. For the models on data quality, we use more diverse estimators. Moreover, to model the share of months without information, we use a fractional response model with a logistic link function (Papke and Wooldridge 1996). This sort of model is appropriate for dependent variables with values ranging from 0 to 1, such as fractions or ratios.

By inserting dummy variables for the point of break-off within the life history module, we can include cases which answered only part of the module but do not confound our estimate with another mechanism. To model the number of episodes and the number of INR, we adopt Cragg's method of double hurdle regression (Cragg 1971; Burke 2009). Differently from the tobit model, this approach allows one to model the probability of a positive value and actual value (given that it is positive) separately, accounting for the possibility that they are caused by different mechanisms. In the models analysing the level of detail treatment, we only include cases where at least one episode was reported in the employment module; for this reason, we use negative binomial regression here. The models dealing with these two outcomes are controlled for the point of break-off dummies, as well as the logarithmised total number of items each respondent received. The total number of items controls for the sheer possibility of having a higher number of INR with an increasing number of questions being asked, and the logarithmic form allows for modelling a diminishing marginal utility. Besides theoretical reasoning, an empirical comparison with linear and quadratic forms also revealed best model fit with the logarithmic form.

Beyond the case selection mentioned above, we apply *multiple imputation by chained equations* (Van Buuren 1999, 2012; White et al. 2011) and the *multiple imputation then deletion* approach of Von Hippel (2007). We did $m = 50$ imputations with $i = 100$ iterations each, using propensity mean matching with the five nearest neighbours (Morris

et al. 2014; Vink et al. 2014; White et al. 2011) to predict plausible values. Additional auxiliary variables used in the imputation but not included in any of the models below are: day of the week and time of the first and last contact as well as their interactions, height and width of the device used, number of children, indicator for a general higher education entrance qualification and the year of obtaining the higher education entrance qualification.

For most variables, there is not a problem of missing values. Nevertheless, two variables have relevant shares of missing information, and we have to control for them due to the selectivity of mobile versus non-mobile users. These are gender and age, with 24% and 25% of information missing, respectively. Almost all the cases with missing information stem from subsample (1) and go back to the master data of the access panel. This variable was coded incorrectly for one of two samples of origin in the access panel. Therefore, we code the cases with implausibly high values on age (above 65 years) as 'missing', and impute plausible values for those cases, as well. Although the nature of the coding error is rather obvious from the distribution of the cases with an age above 65, we used mutiple imputation instead of simply recoding these cases, as it takes uncertainty into account.

3. Results

We estimate two models for each dependent variable, one for the reference period treatment and one for the detail treatment. Since we use double hurdle regression with two parts to test hypotheses H3c, there are 17 models in total. Table 5 shows the average marginal effects of those 17 different models.

All models contain an interaction term between the respective treatment effect and the dummy, specifying whether a mobile or a non-mobile device was used. Taking this interaction into account, the average marginal effects of the treatments are displayed in Figure 2. Additionally, we test the difference between the treatment effects for mobile and non-mobile devices, and the resulting *p*-values can be found in Table 6.

In what follows, we discuss the results for every dependent variable in detail. We first refer to the effects of the reference period treatment, then to the effects of the detail treatment, and finally move to the more differentiated view by type of device. We start by examining the first two hypotheses regarding the impact of our experiment on two dimensions of response burden (H1a and H1b).

3.1. Response Burden (H1a, H1b)

Models 1 and 2 in Table 5 address the time needed to complete the questionnaire (H1a). We estimate those by applying poisson regressions with robust standard errors. The results show that both treatments lead to a significantly increased duration in processing the questionnaire. The effect of the *reference period treatment* indicates an increase of the response duration in the five-year condition, compared to the two-year condition (model 1). The medium and maximum detail conditions of the *detail treatment* lead to an increase, as well, compared to the minimum detail condition (model 2). Looking at the *interactions between type of device and the treatments* in Figure 2, we see that, somewhat surprisingly, response duration is affected only by the treatments in participants who used non-mobile devices. However, the differences in treatment effects between mobile and non-mobile

Table 5. Regression models

	Response time		Difficulty		Break-off		Consent		Number of episodes			Share of months w/o episode		Item nonresponse			
	Poisson regression		OLS		Logistic regression		Logistic regression		Double hurdle linear regression		Negative binomial regression	Fractional response logistic regression		Double hurdle exponential regression			
	Treatment 1	Treatment 2	Treatment 1	Treatment 2	Treatment 1	Treatment 2	Treatment 1	Treatment 2	Treatment 1 Conditional response	Treatment 1 Pr(Y > 0)	Treatment 2	Treatment 1	Treatment 2	Treatment 1 Conditional response	Treatment 1 Pr(Y > 0)	Treatment 2 Conditional response	Treatment 2 Pr(Y > 0)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
Reference period: five years	0.128***		0.167		0.078***		0.041		1.286***	0.005		0.028*		0.039	-0.017		
(ref.: two years)	(4.65)		(1.71)		(3.34)		(0.58)		(11.59)	(0.36)		(2.00)		(0.27)	(-0.70)		
detail: medium		0.077*		-0.046		0.061*		0.277**			-0.140		-0.015			0.106	0.060
(ref.: minimum)		(2.46)		(-0.34)		(2.47)		(2.66)			(-0.77)		(-1.44)			(0.45)	(1.52)
detail: maximum		0.104**		-0.141		0.054*		-0.040			0.095		0.001			1.047***	0.226***
(ref.: minimum)		(3.05)		(-1.04)		(2.18)		(-0.42)			-0.52		(0.09)			(4.05)	(5.64)
mobile device	0.050	0.130***	0.163	0.071	0.124***	0.047	-0.018	0.183	-0.206	-0.025	-0.190	0.063***	0.016	0.099	0.052*	0.239	0.053
(ref.: non-mobile device)	(1.73)	(4.34)	(1.60)	(0.60)	(5.04)	(1.96)	(-0.23)	(1.53)	(-1.80)	(-1.78)	(-1.20)	(4.30)	(1.59)	(0.67)	(2.04)	(1.13)	(1.58)
Observations	1,518	892	1,053	789	1,518	892	163	82	1,518	1,518	892	1,518	892	1,518	1,518	892	892
additionally controlled for *break-off module*	x	x							x	x	x	x	x	x	x	x	x
ln no. of items seen in total														x	x	x	x
no. of job episodes		x															

Notes: Average Marginal Effects; t statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; all models controlled for: treatment calendar vs. q-list; current activity (studying, in vocational training, employment, PhD-Studies, internship, stay abroad, parental leave, unemployed, other); gender; age

Table 6. Test of interaction effects of independent variable with type of device.

	Response Time	Difficulty	Break-off	Consent	Number of spells	Pr(No. of spells > 0)	Share of Months w/o Spell	Item Nonresponse	Pr(No. of INR > 0)
Reference period: five years (ref.: two years)	0.1614	0.0346	0.0532	0.1689	0.0076	0.3038	0.6111	0.0044	0.0042
detail: medium (ref.: minimum)	0.1603	0.0544	0.0749	0.5223	0.8271	-	0.2833	0.8968	0.0318
detail: maximum (ref.: minimum)	0.0185	0.0639	0.0910	0.0191	0.2382	-	0.5356	0.5517	0.2067

Note: P-Values, Hypothesis test on difference between treatment conditions non-mobile vs. mobile

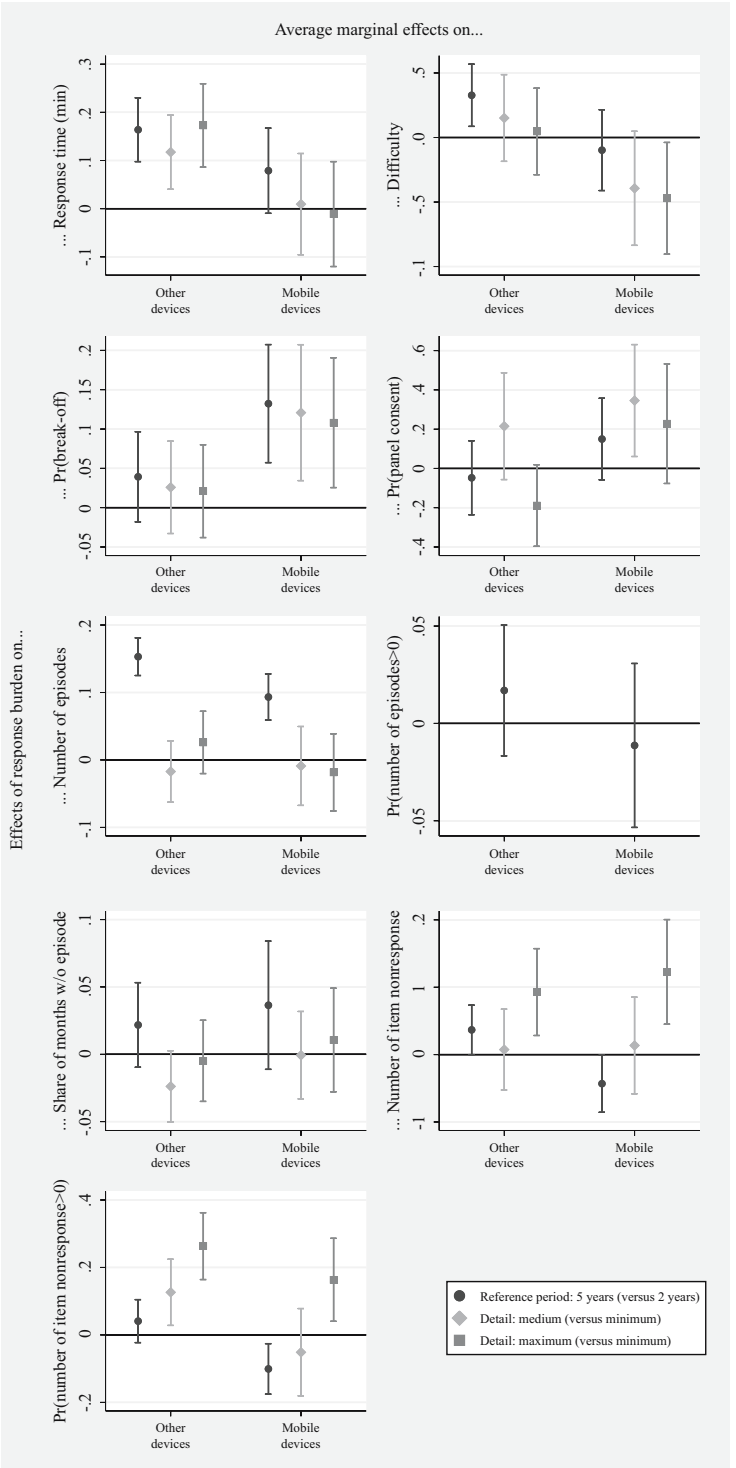


Fig. 2. Average marginal effect plots of interactions with the type of device.

users are significant only for the maximum condition of the detail treatment. One reason for this might be a higher tendency among mobile users to satisfice. But although we controlled for age, gender, and current activity, it is still possible for users of mobile and non-mobile devices to differ from each other in a way that affects their life history, thus leading to different interview durations.

Models 3 and 4 deal with the perceived difficulty of the questionnaire (H1b). The effects of the *reference period* (model 3) and *detail treatment* (model 4) are not significant. Looking at the *interactions*, there is a significant positive effect of the reference period treatment on perceived difficulty only for non-mobile users, and a negative effect of the maximum detail condition for mobile users. The treatment effect of the reference period differs significantly between mobile and non-mobile users, while differences in the effects of the detail treatment by device are only marginally significant.

3.2. Response Behaviour (H2a, H2b)

Here, we discuss the results concerning our hypotheses on response behaviour. In models 5 and 6, we consider the probability that respondents break off the survey (H2a). Increasing the *reference period* (model 5) leads to an increase of almost eight percentage points in this probability. The effects of the *medium and maximum detail conditions* (model 6) are quite similar: each one shows an increase of around five to six percentage points over the minimum detail condition. Looking again at the *interaction* between type of device and treatment effects, it appears that the effect is driven by the mobile users. Referring to the previous results on response burden, the higher likelihood that mobile users break off the survey could explain why there is no increase in response time in this group. However, this does not explain the difference between devices concerning perceived difficulty since this question was asked at the very end of the survey, and it was not seen by respondents who broke off.

Models 7 and 8 address the probability of consenting to future surveys (H2b). The consent question was administered only to respondents who were not already part of our access panel. As a result, case numbers for these two models are rather low. While we do not find a significant effect of the *reference period treatment* on the probability to give consent for future surveys (model 7), we do find a positive effect on consenting to future surveys in the medium condition of the *detail treatment* (model 8), stemming from mobile users. At the same time, looking at the *interactions*, the difference in treatment effects on mobile versus non-mobile users is not significant, given the low case numbers.

3.3. Data Quality (H3a, H3b, H3c)

We now turn to our hypotheses concerning data quality. Models 9–11 address hypothesis H3a, concerning the number of reported episodes. A higher number of episodes is regarded as positively connected to data quality. We estimated the model for the reference period treatment using a double hurdle linear regression, due to excess zeros in the data. Model 9 represents the part of the model conditional on $y_i > 0$, while model 10 represents the probability of $y_i > 0$.

Both parts of the models use the same covariates. It is not surprising that increasing the *reference period* (model 9) leads to an increase in the number of episodes reported by the respondents, since a longer reference period should almost certainly contain more single

episodes than a shorter one. The probability of reporting zero episodes is not affected by the reference period treatment. The *level of detail* (model 10) has no effect on the number of episodes.

Looking at the *interactions with the type of device*, the size of the effect of the reference period treatment differs significantly between mobile and non-mobile users. This may be due to selection, but also to a stronger sensitivity on the part of mobile users to the increased cognitive burden posed by the five-year condition relative to the two-year condition.

We use the share of months in which no episode of any type is reported as another measure of data quality. Hypothesis H3b states that an increase in response burden should lead to an increase in gaps in the reported life history. The results show a positive effect of the *reference period treatment* on the share of months for which no activity was reported (model 12). The *detail treatment*, by contrast, shows no effect (model 13). Turning to the *interactions*, the confidence intervals of the separate effects of the reference period by type of device include zero, due to decreased power. Nevertheless, the point estimate is positive for both groups. For the detail treatment, we do not find significant differences in the treatment effects between devices.

Finally, we test the hypothesis that an increase in response burden leads to an increase in item nonresponse (H3c). Again, due to excess zeros, we use double hurdle regression with the same covariates to estimate both parts of the model. In addition, we include the logarithmised number of items seen in total, as this varies between participants due to heterogeneous life courses.

The *reference period treatment* (models 14 and 15) demonstrates no effect on the number of item nonresponse. Conversely, the *level of detail* has a positive effect on item nonresponse, although only for the maximum detail condition compared to the minimum level of detail. The effect is present both for the conditional part (model 16) and the selection part (model 17) of $\Pr(y_i > 0)$.

Considering the *interactions* of the reference period treatment, significant differences between devices is evident for both parts of the model. A reference period of five years compared to two years leads to a higher amount of item nonresponse (conditional part) for non-mobile users (with a P-value of 0.051, this effect is just out of range of the 95%-confidence level), and a lower one for mobile users (P-value: 0.047). The probability that item nonresponse occurs at all (selection part) is found to be lower in the five-year condition than in the two-year condition, but only for mobile users.

The interactions for the effect of the detail treatment on the amount of item nonresponse (conditional part) show that it is independent of the type of device that was used. For the probability that item nonresponse occurs at all (selection part), there is an additional positive effect of the medium detail condition relative to the minimum detail condition, but it is present only for non-mobile users. This difference between devices is significant. At the same time, the positive effect of the maximum detail condition does not differ significantly between devices.

3.4. Summary of the Findings

We conclude that the *reference period treatment* directly affects response burden, measured as the duration of processing the questionnaire (H1a), and also affects the

subjective evaluation of difficulty (H1b). Surprisingly, the *detail treatment* only affects response time, and not subjective difficulty.

Distinguishing between users of mobile and non-mobile devices yields significant differences for these two hypotheses. While non-mobile users mostly demonstrate the expected effects, the same is not true for mobile users. Furthermore, mobile users show no significant increases in response time or self-assessed difficulty, as opposed to an unexpected negative effect of the maximum detail treatment on the latter. This indicates that either mobile users' reactions to our treatments are different or that the composition of this group differ from that of non-mobile users.

The next two hypotheses pertain to the participants' response behaviour. H2a, which is related to the probability of a break-off, could be confirmed for both treatments. As opposed to H1a and H1b, however, significant effects are present only among mobile users. This might be due to the fact that the users of mobile and non-mobile devices employ different strategies to deal with a higher response burden, with more mobile users breaking off the questionnaire. However, this picture might also be driven by selection into devices, despite our endeavours to account for this.

Despite the very low case numbers, panel consent (H2b) shows some noteworthy effects. Among mobile users and in the average marginal effects without differentiation, there is an increased probability of consent for respondents in the medium detail treatment relative to the minimum detail condition. This is against our expectations in two ways: (1) we assumed an incremental increase in response burden over our detail treatments and (2) we expected a negative effect of a higher response burden. Unfortunately, we do not have a plausible, better explanation for this result than chance. However, the extremely low case numbers in this model suggest that its interpretation should not be overstressed.

The last three hypotheses deal with data quality. Hypothesis 3a is directed only towards the detail treatments, which show no effects. Increasing the reference period should lead to an increased number of episodes even without taking the data quality into account. Nevertheless, the difference in effects for users of mobile and non-mobile devices is remarkable. The number of episodes reported increased significantly less in mobile users, and this could not be accounted for by controlling for age, gender, or current activity. One explanation for this finding might be a higher tendency of underreporting in this group.

The second data-quality hypothesis (H3b) concerns the number of gaps in the reported episodes. There is evidence that increasing the reference period increases the share of months without information, while the detail treatment shows no significant effect on this outcome. Moreover, the analyses of H3c yield partially counterintuitive results. The detail treatment shows expected effects for both mobile and non-mobile users, in both parts of the model; but alternating the reference period has only a small and marginally significant effect on non-mobile users modelling the number of nonresponses conditional on $y_i > 0$, while having a negative effect on mobile users in both parts of the model. One explanation for this might be that respondents using mobile devices show types of satisficing that do not increase item nonresponse but do yield a lower number of reported episodes (underreporting). Since we control for the total number of items seen, this could result in a decrease in item nonresponse. This explanation would also be concurrent with the smaller increase in the number of reported episodes from models 9 and 10.

4. Discussion

This article investigates the question of how response burden variations in terms of the amount and detail of information requested on life history affect different outcomes in a self-administered, mixed-device web survey of a highly educated population. We find several effects which stress the importance of managing response burden in self-administered online surveys.

Our results indicate that, as we expected, our treatments clearly increase response burden measured as response time and break-off. Notably, the differences in results between users of mobile and non-mobile devices point to different experiences of and reactions to increased response burden. Mobile participants show no effect on response time but do show an increase in break-offs. Conversely, non-mobile participants show no effect on break-offs but do show an increase in response time.

The results on subjective difficulty and panel consent are less straightforward. In general, while the effects are in the expected direction among non-mobile users, mobile users show unexpected behaviours. However, it should be kept in mind that the analyses on panel consent are based only on a very small number of cases, and only on subsamples (3) and (4). Participants from subsample (1) could not be asked once again about their panel consent, as they came from a probability-based access panel.

Concerning our data quality measures, we find effects on all three outcomes, but not for every treatment. Results on the number of episodes are consistent with the rest of our findings. The number of reported episodes is expected to be higher when the reference period is increased from two years to five years. However, this increase has a smaller effect on respondents using mobile devices. This might be an indication of specific satisficing strategies used by this group. At the same time, this group exhibits a negative effect of prolonged reference period on item nonresponse.

Both findings could be explained by a higher tendency towards underreporting in the group of mobile users, thus leaving less opportunity for nonresponse. As before, although we aimed to account for selectivity in mobile and non-mobile usage, we cannot rule out that this effect could be traced back to a different composition of these groups' life histories.

The detail level treatment does not affect the number of episodes reported. Regarding item nonresponse, this treatment shows expected results. Increasing the complexity of the survey by raising the level of detail does affect item nonresponse, independently of the device used.

Two key findings emerge from our study. First, varying the amount and complexity of information collected in a mixed-device web survey has different effects on data quality and response behaviour. Increasing the amount of requested information apparently leads to satisficing in mobile users who report a lower number of episodes, while item nonresponse is mainly influenced by the complexity of the questionnaire. Second, mobile, and non-mobile users seem to experience and react differently to response burden. Mobile users tend to break off more easily, while non-mobile users spend more time answering the questionnaire and rating the questionnaire as more difficult. As stated above, mobile users in our sample are also more prone to satisficing.

Although our study relies on experimental variations, there are several limitations to our findings. The possibility of selection bias concerning differences between mobile and non-

mobile users forbids us to interpret differences in the effects as causal. We try to account for all relevant confounders, but we cannot rule out that some of our findings could be driven by selective usage of devices. Furthermore, our experimental treatments operationalising the amount and complexity of information are not necessarily transferable to other surveys, due to the specific nature of our study.

We constructed our questionnaire using an interleafed looping strategy on filter and follow-up questions. Since it is known that this leads to different respondent behaviour than a grouped approach (Eckman and Kreuter 2018), our results might have been different using the latter. Moreover, our sample, comprising mostly participants of an access panel, is on average rather young and highly educated. There is evidence that response burden is perceived to be higher in mobile devices for older and lower-educated participants (Yan and Tourangeau 2008; Couper and Kreuter 2013). Hence, in different samples, the effects we found might be even more pronounced.

It is unclear whether our treatments would have yielded the same effects in a different population, in a survey containing different topics and being conducted by another organisation. Nevertheless, due to these conditions, it is reasonable to assume that we have found somewhat lower effects than would be expected in the general population. Although our study lacks objective measures of data quality, we are confident that we made reasonable presumptions about the expected effects on our outcomes.

For survey practice, our results have major implications. An increase in the scope and complexity of surveys should be considered carefully, given the effects on different outcomes we demonstrated. Furthermore, these effects might lead to a differential and complex bias by the type of device which respondents use to complete the survey, since our results pointed to different experiences and consequences of increased response burden along this divide. We recommend that future studies on this theme create techniques capable of eliminating these differences.

5. References

American Association for Public Opinion Research. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed December 2020).

Antoun, C., and A. Cernat. 2020. "Factors Affecting Completion Times: A Comparative Analysis of Smartphone and PC Web Surveys." *Social Science Computer Review* 38 (4): 477–489. DOI: https://doi.org/10.1177/0894439318823703.

Antoun, C., M.P. Couper, and F.G. Conrad. 2017. "Effects of Mobile Versus PC Web on Survey Response Quality." *Public Opinion Quarterly* 81 (S1): 280–306. DOI: https://doi.org/10.1093/poq/nfw088.

Bradburn, N.M. 1979. "Respondent Burden." In *Health Survey Research Methods: Second Biennial Conference, Williamsburg*, edited by L. Reeder. Washington, D.C. U.S. Government Printing Office.

Burke, W.J. 2009. "Fitting and Interpreting Cragg's Tobit Alternative Using Stata." *Stata Journal* 9 (4): 584–592. Available at: https://www.stata-journal.com/article.html?article = st0179.

Burton, S., and E. Blair. 1991. "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys." *Public Opinion Quarterly* 55 (1): 50. DOI: https://doi.org/10.1086/269241.

Busemeyer, J.R., and J.T. Townsend. 1993. "Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment." *Psychological review* 100 (3): 432–459. DOI: https://doi.org/10.1037/0033-295X.100.3.432.

Buskirk, T.D., and C.H. Andrus. 2014. "Making Mobile Browser Surveys Smarter." *Field Methods* 26 (4): 322–342. DOI: https://doi.org/10.1177/1525822X14526146.

Clement, S.L., M. Kappelgaard Severin-Nielsen, and D. Shamsiri-Petersen. 2020. "Device Effects on Survey Response Quality. A Comparison of Smartphone, Tablet and PC Responses on a Cross Sectional Probability Sample." *Survey Methods: Insights from the Field*. DOI: https://doi.org/10.13094/SMIF-2020-00020.

Couper, M.P., and F. Kreuter. 2013. "Using Paradata to Explore Item Level Response Times in Surveys." *Journal of the Royal Statistical Society:* 176 (1): 271–286. DOI: https://doi.org/10.1111/j.1467-985X.2012.01041.x.

Couper, M.P., and G.J. Peterson. 2017. "Why Do Web Surveys Take Longer on Smartphones?" *Social Science Computer Review* 35 (3): 357–377. DOI: https://doi.org/10.1177/0894439316629932.

Cragg, J.G. 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica* 39 (5): 829. DOI: https://doi.org/10.2307/1909582.

Crawford, S,D., M.P. Couper, and M.J. Lamias. 2001. "Web Surveys." *Social Science Computer Review* 19 (2): 146–162. DOI: https://doi.org/10.1177/089443930101900202.

Daikeler, J., R.L. Bach, H. Silber, and S. Eckman. 2020. "Motivated Misreporting in Smartphone Surveys." *Social Science Computer Review*. DOI: https://doi.org/10.1177/0894439319900936.

De Bruijne, M, and M. Oudejans. 2015. "Online Surveys and the Burden of Mobile Responding." In *Survey Measurements: Techniques, Data Quality and Sources of Error*, edited by U. Engel: 130–145. Frankfurt, New York: Campus.

De Bruijne, M., and A. Wijnant. 2014. "Mobile Response in Web Panels." *Social Science Computer Review* 32 (6): 728–742. DOI: https://doi.org/10.1177/0894439314525918.

Drasch, K., and B. Matthes. 2013. "Improving Retrospective Life Course Data by Combining Modularized Self-Reports and Event History Calendars: Experiences from a Large Scale Survey." *Qual Quant* 47 (2): 817–838. DOI: https://doi.org/10.1007/s11135-011-9568-0.

Eckman, S., and F. Kreuter. 2018. "Misreporting to Looping Questions in Surveys: Recall, Motivation and Burden." *Survey Research Methods* 12 (1): 59–74. DOI: https://doi.org/10.18148/srm/2018.v12i1.7168.

Galesic, M. 2006. "Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics* 22 (2): 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf.

Galesic, M., and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Social Science Computer Review* 73 (2): 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2011. *Survey Methodology*. 2nd edition. Hoboken: John Wiley & Sons. Available at: http://gbv.eblib.com/patron/FullRecord.aspx?p = 819140.

Guin, T.D.-Le., R. Baker, J. Mechling, and E. Ruyle. 2012. "Myths and Realities of Respondent Engagement in Online Surveys." *International Journal of Market Research* 54 (5): 613–633. DOI: https://doi.org/10.2501/IJMR-54-5-613-633.

Gummer, T., F. Quoß, and J. Roßmann. 2019. "Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones?" *Social Science Computer Review* 37 (3): 371–384. DOI: https://doi.org/10.1177/0894439318766836.

Gummer, T., and J. Roßmann. 2015. "Explaining Interview Duration in Web Surveys." *Social Science Computer Review* 33 (2): 217–234. DOI: https://doi.org/10.1177/0894439314533479.

Hambleton, R.K., H. Swaminathan, and H.J. Rogers. 1991. *Fundamentals of Item Response Theroy*. California: Sage.

Hedlin, D., H. Lindkvist, H. Bäckström, and J. Erikson. 2008. "An Experiment on Perceived Survey Response Burden Among Businesses." *Journal of Official Statistics* 24 (2): 301–318. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-experiment-on-perceived-survey-response-burden-among-businesses.pdf.

Hoogendoorn, A.W., and D. Sikke. 1998. "Response Burden and Panel Attrition." *Journal of Official Statistics* 14 (2): 189–205. Available at: https://www.scb.se/contentassets/-ca21efb41fee47d293bbee5bf7be7fb3/response-burden-and-panel-attrition.pdf.

Kaper, E., and W. Saris. 1999. "Effects of Response Burden in a Continuous Survey." *Kwantitative Method* 61: 5–17.

Keusch, F., and T. Yan. 2017. "Web Versus Mobile Web." *Social Science Computer Review* 35 (6): 751–769. DOI: https://doi.org/10.1177/0894439316675566.

Krebs, D., and J.K. Höhne. 2020. "Exploring Scale Direction Effects and Response Behavior Across Pc and Smartphone Surveys." *Journal of Survey Statistics and Methodology* 9: 261. DOI: https://doi.org/10.1093/jssam/smz058.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. DOI: https://doi.org/10.1002/acp.2350050305.

Lang, S., and J. Carstensen. 2022. "A life history calendar in a CAWI? Evidence from higher education research." In *Survey Methoden in der Hochschulforschung*, edited by G. Brandt and S. de Vogel: 121–152. Wiesbaden: Springer VS. DOI: https://doi.org/10.1007/978-3-658-36921-7_6.

Lee, H., S. Kim, M.P. Couper, and Y. Woo. 2019. "Experimental Comparison of PC Web, Smartphone Web, and Telephone Surveys in the New Technology Era." *Social Science Computer Review* 37 (2): 234–247. DOI: https://doi.org/10.1177/0894439318756867.

Lenzner, T. 2012. "Effects of Survey Question Comprehensibility on Response Quality." *Field Methods* 24 (4): 409–428. DOI: https://doi.org/10.1177/1525822X12448166.

Lugtig, P., and V. Toepoel. 2016. "The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey." *Social Science Computer Review* 34 (1): 78–94. DOI: https://doi.org/10.1177/0894439315574248.

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schütz. 2007. "Compensating for Low Topic Interest and Long Surveys." *Social Science Computer Review* 25 (3): 372–383. DOI: https://doi.org/10.1177/0894439307297606.

Mavletova, A. 2013. "Data Quality in PC and Mobile Web Surveys." *Social Science Computer Review* 31 (6): 725–743. DOI: https://doi.org/10.1177/0894439313485201.

Morris, T.P., I.R. White, and P. Royston. 2014. "Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws." *BMC medical research methodology* 14: 75. DOI: https://doi.org/10.1186/1471-2288-14-75.

Papke, L.E., and J.M. Wooldridge. 1996. "Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates." Journal of Applied Econometrics 11 (6): 619–632. DOI: https://doi.org/10.1002/(SICI)1099-1255(199611) 11: 6%3C619::AID-JAE418%3E3.0.CO;2-1.

Peytchev, A., and E. Peytcheva. 2017. "Reduction of Measurement Error Due to Survey Length: Evaluation of the Split Questionnaire Design Approach." *Survey Research Methods* 11 (4): 361–368. DOI: https://doi.org/10.18148/srm/2017.v11i4.7145.

Reise, S.P., A.T. Ainsworth, and M.G. Haviland. 2005. "Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research." *Current directions in psychological science* 14 (2): 95–101. DOI: https://doi.org/10.1111/j.0963-7214.2005.00342.x.

Schlosser, S., and A. Mays. 2018. "Mobile and Dirty." *Social Science Computer Review* 36 (2): 212–230. DOI: https://doi.org/10.1177/0894439317698437.

Sharp, L.M., and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 43. DOI: https://doi.org/10.1086/268765.

Sommer, J., B. Diedenhofen, and J. Musch. 2017. "Not to Be Considered Harmful." *Social Science Computer Review* 35 (3): 378–387. DOI: https://doi.org/10.1177/0894439316633452.

Toepoel, V., and P. Lugtig. 2014. "What Happens If You Offer a Mobile Option to Your Web Panel? Evidence from a Probability-Based Panel of Internet Users." *Social Science Computer Review* 32 (4): 544–560. DOI: https://doi.org/10.1177/0894439313510482.

Tourangeau, R. 2000. "Remembering What Happened: Memory Errors and Survey Reports." In *The Science of Self-Report: Implications for Research and Practice*, edited by A.A. Stone. Mahwah, N.J: Lawrence Erlbaum.

Tourangeau, R. 2003. "Cognitive Aspects of Survey Measurement and Mismeasurement." *International Journal of Public Opinion Research* 15. DOI: https://doi.org/10.1093/ijpor/15.1.3.

Tourangeau, R. 2018. "The Survey Response Process from a Cognitive Viewpoint." *Quality Assurance in Education* 26 (2): 169–181. DOI: https://doi.org/10.1108/QAE-06-2017-0034.

Van Buuren, S. 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine* 18 (6): 681–694. DOI: https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6 < 681:AID-SIM71 > 3.0.CO;2-R.

Van Buuren, S. 2012. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton, Fla. CRC Press.

Vink, G., L.E. Frank, J. Pannekoek, and S. van Buuren. 2014. "Predictive Mean Matching Imputation of Semicontinuous Variables." *Statistica Neerlandica* 68 (1): 61–90. DOI: https://doi.org/10.1111/stan.12023.

Von Hippel, P.T. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology* 37 (1): 83–117. DOI: https://doi.org/10.1111/j.1467-9531.2007.00180.x.

Warriner, G.K. 1991. "Accuracy of Self-Reports to the Burdensome Question: Survey Response and Nonresponse Error Trade-Offs." *Quality & Quantity* 25: 253–269. DOI: https://doi.org/10.1007/BF00167531.

White, I.R., P. Royston, and A.M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–399. DOI: https://doi.org/10.1002/sim.4067.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response Burden: What Is It and What Predicts It?" In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P.C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G.B. Willis, and A. Wilmot: 193–212. Hoboken, NJ, USA: Wiley.

Yan, T., and R. Tourangeau. 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: https://doi.org/10.1002/acp.1331.

Zanon, C., C.S. Hutz, H. Yoo, and R.K. Hambleton. 2016. "An Application of Item Response Theory to Psychological Test Development." Psicologia: Reflexão e Crítica / Psychology: Research and Review 29 (1): 1. DOI: https://doi.org/10.1186/s41155-016-0040-x.

Received January 2021
Revised July 2021
Accepted December 2021

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1097–1123, http://dx.doi.org/10.2478/JOS-2022-0047

Your Best Estimate is Fine. Or is It?

Jerry Timbrook[1], *Kristen Olson*[2], *and Jolene D. Smyth*[2]

Providing an exact answer to open-ended numeric questions can be a burdensome task for respondents. Researchers often assume that adding an invitation to estimate (e.g., "Your best estimate is fine") to these questions reduces cognitive burden, and in turn, reduces rates of undesirable response behaviors like item nonresponse, nonsubstantive answers, and answers that must be processed into a final response (e.g., qualified answers like "about 12" and ranges). Yet there is little research investigating this claim. Additionally, explicitly inviting estimation may lead respondents to round their answers, which may affect survey estimates. In this study, we investigate the effect of adding an invitation to estimate to 22 open-ended numeric questions in a mail survey and three questions in a separate telephone survey. Generally, we find that explicitly inviting estimation does not significantly change rates of item nonresponse, rounding, or qualified/range answers in either mode, though it does slightly reduce nonsubstantive answers for mail respondents. In the telephone survey, an invitation to estimate results in fewer conversational turns and shorter response times. Our results indicate that an invitation to estimate may simplify the interaction between interviewers and respondents in telephone surveys, and neither hurts nor helps data quality in mail surveys.

Key words: Estimation; enumeration; questionnaire design; data quality; respondent burden.

1. Introduction

Survey researchers often use open-ended questions to capture numeric responses for questions that require enumeration of events that occurred over a fixed time period (e.g., "How many cigarettes did you smoke in the last seven days?") or questions asking for financial information (e.g., income). Open-ended numeric questions are often used when a precise number is needed for the survey's analytic goals or when it is difficult to construct meaningful ranges for response options (Dillman et al. 2014). Yet providing an exact, numeric answer can be mentally taxing or impossible for respondents, especially when questions ask about hard-to-enumerate topics (Tourangeau et al. 2000; Conrad et al. 1998). As a result, some respondents may fail to answer a question altogether or give nonsubstantive answers (i.e., responses like "too many cigarettes to count"), which can

[1] RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709, U.S.A. Email: jtimbrook@rti.org
[2] University of Nebraska-Lincoln, 711 Oldfather Hall, Lincoln, NE 68588-0324, U.S.A. Emails: kolson5@unl.edu; jsmyth2@unl.edu

Acknowledgments: This work was supported in part by funds provided to the University of Nebraska-Lincoln under a Cooperative Agreement with the USDA-National Agricultural Statistics Service supported by the National Science Foundation National Center for Science and Engineering Statistics (58-AEU-5-0023 to J.D.S and K.O.). Additional funding was provided by the Office of Research and Economic Development and the Department of Sociology at the University of Nebraska-Lincoln, and the National Science Foundation (Grant Number SES-1132015 to K.O.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

negatively affect data quality (Beatty and Herrmann 2002). Difficulty providing exact answers may also cause some respondents to report responses that must be recoded into a final answer by data processing staff (e.g., ranges like "five to ten cigarettes a week").

To ease the burden of answering numeric response questions and potentially limit undesirable answering behaviors, researchers often invite respondents to estimate their answer by adding phrases like "Your best estimate is fine" to the question stem (Dillman 2007). However, there is no research of which we are aware that establishes whether this phrase is helpful in mail and telephone surveys. Additionally, inviting estimation may also trigger changes in other response behaviors. For example, response heaping (i.e., giving rounded answers) is more prevalent when respondents estimate (Huttenlocher et al. 1990; Burton and Blair 1991; Holbrook et al. 2014), and may increase if respondents are told they can approximate their answer. Estimation can also lead respondents to overreport their behaviors, meaning that an invitation to estimate may inflate survey means (Burton and Blair 1991). Finally, it is possible that these effects vary across self-administered and interviewer-administered surveys because the cognitive and working memory demands on respondents differ across these modes (De Leeuw 2005).

In this article, we experimentally explore the effect of including an invitation to estimate worded as "Your best estimate is fine" on several indicators of data quality in a mail survey and in a separate telephone survey. We compare versions of questions with and without the invitation to estimate to answer the following research questions:

RQ1. Does the item nonresponse rate differ across versions?
RQ2. Does the rate of nonsubstantive answers differ across versions?
RQ3. Does the rate of range and qualified answers differ across versions?
RQ4. Does the heaping or rounding rate differ across versions?
RQ5. Do estimated means differ across versions?

2. Background

Questionnaire designers have several options when asking numeric questions such as how many miles respondents drive in an average week. A closed-ended question format with ranges for response options (e.g., "100 to less than 200" miles) can be simpler for respondents but yields imprecise responses. In contrast, an open-ended numeric item both cues and allows respondents to provide a precise answer (e.g., 150 miles), but respondents may have difficulty recalling such a precise answer (Tourangeau et al. 2000).

2.1. Response Strategies for Open-Ended Numeric Questions

Respondents generally use either enumeration or estimation to answer open-ended numeric questions about events or behaviors. *Enumerating* a precise answer requires respondents to retrieve each episode of the event or behavior from memory, decide if the episode fits the question's requirements, count up the total number of relevant episodes, and report that number (Blair and Burton 1987; Menon 1993; Tourangeau et al. 2000). For example, respondents often use *enumeration* to answer questions that ask about topics like *infrequent events* (e.g., trips to the emergency room) or events over *short and recent reference periods* (e.g., the past week) (Blair and Burton 1987; Burton and Blair 1991;

Conrad et al. 1998). Events that do not happen on a routine schedule (e.g., irregular events like the number of visits to the doctor for an injury) and events whose episodes are distinct from one another (e.g., number of serious illnesses) are often enumerated (Menon 1993; Conrad et al. 1998).

Alternatively, *estimation* involves providing an answer based on an approximate rate of occurrence for a target behavior or event (Blair and Burton 1987; Burton and Blair 1991; Tourangeau et al. 2000). Questions asking about *frequent events* (e.g., steps taken) or events that occur over *long reference periods* (e.g., your adult lifetime) lend themselves to estimation because their exact numbers often do not exist in memory (Blair and Burton 1987; Burton and Blair 1991; Conrad et al. 1998). Similarly, events that happen on a regular schedule (e.g., eating breakfast) and events whose episodes are similar to one another (e.g., minor illnesses) are often estimated (Menon 1993; Conrad et al. 1998). Exact answers to financial questions like one's annual income may also be difficult to remember and often estimated.

A respondent's use of enumeration versus estimation depends on how easily relevant information can be retrieved from memory (i.e., the *cognitive state*) (Beatty and Herrmann 2002); this answer may be modified if the respondent feels the need to edit the answer to a possibly *sensitive question* (i.e., the respondent's communicative intent). We start with discussing cognitive states and then move to communicative intent.

Information that requires minimal effort to retrieve (i.e., *available* information) can easily be enumerated and thus likely pose little burden to respondents to enumerate. Limits in memory, however, can prevent respondents from using enumeration (Huttenlocher et al. 1990). Estimation is commonly used when the exact information requested by a question cannot be retrieved, but a respondent can use other information in memory to approximate their answer (i.e., *generatable* information) (Beatty and Herrmann 2002). Respondents using estimation must decide whether their inexact answer meets the level of precision requested by the question (i.e., an adequacy judgement) (Beatty and Herrmann 2002). If respondents judge their estimated answer to be inadequate, they may skip the question entirely. Alternatively, respondents may indicate that their answer is estimated (i.e., potentially inadequate) by providing nonsubstantive answers (e.g., "too many to count"). Respondents may also indicate that their answer is "generated" by including additional information like ranges (e.g., "10–20") or qualifiers (e.g., "about 20") in their answer. Although these answers can ultimately be recoded into a final response using a set of rules (e.g., ranges like "10–20" coded as the lower bound of ten), they require costly post-survey processing that may introduce processing error.

Accessible information exists in a cognitive state between available and generatable and can be retrieved from memory only if a respondent exerts cognitive effort (Beatty and Herrmann 2002). The enumeration process for accessible information is burdensome; respondents must expend mental resources to retrieve each episode from memory. Respondents lacking sufficient motivation might avoid this burdensome retrieval task by estimating their answer instead, leading to less precise answers. Respondents again must make an adequacy judgment to determine if their estimated answer meets the precision requirements of the question. This judgment may lead respondents to not answer the question at all, resulting in item nonresponse, or to provide nonsubstantive, qualified, or range answers.

Finally, when exact information related to a question is not known and cannot be approximated, it is *inestimable* (e.g., apples eaten in your entire life) (Beatty and Herrmann 2002). In such cases, neither enumeration nor estimation will lead to an answer. The only honest recourse for respondents is to: (1) skip answering the question altogether; (2) select a "don't know" or "refusal" response option when available; or (3) provide a nonsubstantive answer (e.g., "Too many to count"). In general, survey designers are encouraged to avoid questions that ask for inestimable information.

Responses are modified by respondents' communicative intent, or the need to edit answers to sensitive questions (Beatty and Herrmann 2002). Respondents are more likely to skip sensitive questions (i.e., item nonresponse) if they worry their answers are socially undesirable (Beatty and Herrmann 2002; Tourangeau and Yan 2007). Alternatively, they may edit answers to fit within a social norm or expectations. This editing may manifest in behaviors that look like those of estimation – for instance, providing a range or reporting a more socially acceptable (and perhaps rounded) inexact answer.

2.2. *An Invitation to Estimate*

Appending the phrase "Your best estimate is fine" to a question is a common method for indicating that approximate answers are acceptable for numeric response questions (Dillman 2007). This phrase, inviting respondents to estimate their answers, is hypothesized to reduce item nonresponse rates by communicating that: (1) respondents do not have to engage in potentially burdensome enumeration (i.e., for accessible information), and (2) an imprecise answer is preferable to no answer (i.e., reporting generatable or estimated accessible information is acceptable), potentially making respondents' adequacy judgments less burdensome. Explicitly inviting estimation is also hypothesized to reduce nonsubstantive answers through the same mechanisms. By communicating that guesses are acceptable, an invitation to estimate is hypothesized to also reduce range or qualified answers; respondents may feel less inclined to communicate uncertainty over their answer (e.g., ranges, qualified answers) if they are told that estimation is permissible.

Despite the potential to decrease item nonresponse, an invitation to estimate may also have the unintended consequence of encouraging some respondents to estimate when they would have otherwise enumerated their answer. This can occur when the exact information requested by a question is retrievable with cognitive effort (i.e., accessible information), but the respondent approximates their answer when offered the invitation to estimate. An increase in estimation may in turn increase other potentially undesirable response behaviors. Primarily, respondents who estimate are more likely to provide rounded answers (i.e., heaping) (Burton and Blair 1991; Holbrook et al. 2014). An invitation to estimate may encourage this heaping behavior and foster inaccurate responses. Finally, as estimated answers tend to be larger than enumerated answers – that is, respondents round up rather than round down (Burton and Blair 1991) – survey means may be larger with an invitation to estimate. Therefore, when an invitation to estimate succeeds at encouraging estimation behavior, we would expect to see: lower rates of item nonresponse (H1a), nonsubstantive answers (H2a), and qualified or range answers (H3a); higher rates of heaping (H4a); and larger estimated means (H5a; summarized in Table 1).

Table 1. Summary of hypotheses.

Response outcome	Example	With an invitation to estimate (ItE). . .	And will vary by. . .	
			Cognitive state	Question sensitivity
Item nonresponse	Respondent does not answer the question	H1a: Item nonresponse rates will be lower.	H1b: Item nonresponse rates will be lower with an ItE for accessible information than other cognitive states.	H1c: Item nonresponse rates will be lower for sensitive questions with an ItE than for sensitive questions without an ItE.
Nonsubstantive answers	Respondent answers "Too many to count"	H2a: Nonsubstantive answer rates will be lower.	H2b: Nonsubstantive answer rates will be lower with an ItE for accessible information than other cognitive states.	H2c: Nonsubstantive answer rates will be lower for sensitive questions with an ItE than for sensitive questions without an ItE.
Qualified or range answers	Respondent answers "About 20" (qualified) or "10–20" (range)	H3a: Rates of qualified or range answers will be lower.	H3b: Rates of qualified or range answers will be lower with an ItE for accessible information than other cognitive states.	H3c: Rates of qualified or range answers will be lower for sensitive questions with an ItE than for sensitive questions without an ItE.
Heaping	Respondent answers with a rounded number like 20	H4a: Rates of heaped or rounded answers will be higher.	H4b: Rates of heaped or rounded answers will be higher with an ItE for accessible information than other cognitive states.	H4c: Rates of heaped or rounded answers will be higher for sensitive questions with an ItE than for sensitive questions without an ItE.
Survey estimates (means)	–	H5: Responses will be rounded up, yielding higher means.	Not testable in this study	Not testable in this study
Question administration length	–	H6: Fewer conversational turns, and response time will be lower.	Not testable in this study	Not testable in this study

ItE: Invitation to Estimate

We are aware of only one study that investigated using an invitation to estimate: Couper et al. (2011) found that adding this phrase had no effect on the rate of ill-formed answers (i.e., answers that do not conform to the response task like nonsubstantive and qualified/range answers) or response times on three open-ended numeric questions in a web survey. Evaluating more varied question types in different modes may yield different results. Further, nonsubstantive and qualified/range answers have not been evaluated separately despite having different effects on measurement (e.g., nonsubstantive answers are often set to item missing, while qualified/range answers are processed into a final response). Finally, the assumption that an invitation to estimate reduces item nonresponse rates remains uninvestigated.

2.2.1. Invitation to Estimate, Cognitive States, and Question Sensitivity

The effect of inviting estimation may also differ across cognitive states. For example, it may have no effect on requests for information that is available (because enumeration is not burdensome in this case), generatable (because this information could only be estimated in the first place), or inestimable (because this information cannot be retrieved). However, an invitation to estimate may lead respondents to approximate their answers to questions asking about accessible information by giving them permission to skip the cognitively taxing enumeration process and estimate their answer instead. This may lead to the outcomes that we hypothesize accompany a successful invitation to estimate (i.e., H1b-H4b).

The effect of an invitation to estimate may also vary with question sensitivity. An invitation to estimate may change a respondent's adequacy judgement for sensitive questions. In particular, permission to estimate may make respondents more comfortable answering the question by providing an inexact (e.g., rounded/heaped), less sensitive answer. For example, a respondent may be unwilling to report that they have received exactly nine speeding tickets in the past year. Including an invitation to estimate may encourage this respondent to instead provide an answer of five tickets: a rounded (i.e., plausibly estimated), less sensitive answer. Therefore, we hypothesize that adding an invitation to estimate on sensitive questions will lead to less item nonresponse by changing respondents' adequacy judgement to allow for estimated, less sensitive answers (H1c). Accordingly, we also hypothesize that this increase in estimated responses for sensitive questions will lead to the outcomes that we anticipate accompany a successful invitation to estimate (H2c-H4c).

2.2.2. Invitation to Estimate and Data Collection Mode

It is unclear if the effects of an invitation to estimate vary across self-administered and interviewer-administered modes. In self-administered surveys, respondents can see the text of a question and refer to it when considering their response. In a telephone survey, respondents must hold the question in their working memory while also considering their answer. This makes aural telephone surveys more cognitively taxing than visual modes (De Leeuw 2005) and suggests that an invitation to estimate might prove especially useful at reducing item nonresponse in the telephone mode. On the other hand, because telephone interviewers serve as motivating agents, item nonresponse rates are generally already low

in this mode (De Leeuw 2005); there may be little room for further reductions in item nonresponse rates due to an invitation to estimate.

Interviewers also interact with respondents to resolve nonsubstantive and qualified/range answers, for example, by probing these types of responses to obtain a single integer value. Because interviewers do this work during the question-asking process, there may be no detectable effect of an invitation to estimate on rates of nonsubstantive and qualified/range answers in the final recorded responses in telephone surveys (although we still expect increased heaping and larger estimated means). However, an invitation to estimate may decrease instances where respondents give nonsubstantive and qualified/range answers at any point during the question/answer conversation and reduce the need for interviewer intervention. Thus, we expect to see lower rates of respondents ever giving nonsubstantive or qualified/range answers when estimation is explicitly allowed.

Finally, unlike the mail mode, telephone surveys give researchers access to an additional indicator of respondent burden: question administration length. Question administrations lasting longer than a paradigmatic "question asked/question answered" sequence (i.e., around two conversational turns) can indicate that a question is burdensome (Schaeffer and Maynard 1996). Longer response times are also a common sign that survey questions are difficult to answer (Bassili and Scott 1996; Draisma and Dijkstra 2004). If an invitation to estimate reduces burden in the telephone mode, we would expect to see (H6): (1) fewer conversational turns between the interviewer and respondent and (2) shorter response times when estimation is allowed.

In this article, we evaluate the potential benefits and drawbacks of adding an invitation to estimate to open-ended numeric questions. For 22 questions in a mail survey we compare: rates of item nonresponse, nonsubstantive answers, qualified/range answers, and heaping; as well as estimated means across questions with and without an invitation to estimate. In the mail mode, we also explore the effect of inviting estimation across two question characteristics: cognitive state and sensitivity. For three questions in a telephone survey, we compare: rates of item nonresponse, ever giving nonsubstantive answers, ever giving qualified/range answers, and heaping; as well as estimated means, number of conversational turns, and response times across questions with and without an invitation to estimate.

3. Data and Methods

The data for the mail study comes from the National Health, Wellbeing and Perspectives Study (NHWPS) survey. NHWPS was conducted by the University of Nebraska-Lincoln's Bureau of Sociological Research (BOSR) in the spring of 2015. A total of 1,002 respondents completed and returned the survey (AAPOR RR1 = 16.7%). Respondents were randomly selected using the next birthday within-household selection method. The 12-page NHWPS questionnaire contained 77 questions asking about health, mental health, well-being, victimization, current events, and demographics. Sampled households were randomly assigned to one of two versions of the questionnaire (Version 1: $n = 522$, AAPOR RR1 = 17.4%; Version 2: $n = 480$, AAPOR RR1 = 16.0%). In Version 2, an invitation to estimate (i.e., "Your best estimate is fine.") was appended to 22 open-ended

numeric questions asking about the number of times that particular events occurred in their lifetime, number of hours spent on certain behaviors during a typical week, and an income question (see Online supplemental material A for question wording). This invitation was not included in Version 1.

The data for the telephone study come from the Work and Leisure Today 2 (WLT2) survey, a dual-frame random-digit dial telephone survey of U.S. adults. WLT2 was conducted by Abt SRBI in the summer of 2015. For landline numbers, survey respondents were randomly selected among adult household members using the Rizzo et al. (2004) within-household selection method (using the next birthday method for households with 3+ adults). For cell phone numbers, the adult who answered the phone was interviewed. The survey had 902 respondents (AAPOR RR3 = 7.8%), and contained 58 questions asking about leisure time, use of technology, and demographics. We again randomly assigned sampled cases to receive the invitation to estimate (Version 1: n = 451, AAPOR RR3 = 7.4%) for three behavioral questions, or not (Version 2: n = 451, AAPOR RR3 = 8.4%) (see Online supplemental material A for question wording). Three Version 1 cases were removed from analysis due to poor call quality (final n = 899).

3.1. Dependent Variables

3.1.1. Mail Survey

Our first dependent variable in the mail survey is an indicator for item nonresponse coded 1 if a respondent did not provide an answer to a question, and 0 otherwise. Next, we set an indicator for nonsubstantive answers to 1 if a respondent provided a non-numeric answer that could *not* be recoded as an integer (e.g., written notes like "too many to count" or "do the math") and 0 otherwise. Item nonrespondents are excluded from this indicator.

We operationalize qualified/range answers using an indicator variable coded as 1 for non-integer answers that *could* be processed into an integer based on a set of rules (e.g., range answers like "20 to 30" could be coded to 20, the lower limit of the range; qualified answers like "About 20" could be coded to 20; questions like "two hours a day per week" could be coded to 14; decimal answers like 25.2 could be rounded to 25; answers with units like "18 hours" could be coded as 18) and 0 otherwise. We make one exception to these rules for written, negative answers like "no" or "none". For the 21 event/behavior questions in the mail survey, we treat negative answers as a final answer of "0," and do not code them as qualified/range answers. However, these answers may have a different meaning in the context of an income question. Therefore, we code these answers as nonsubstantive for our income question, because negative responses like "no" may indicate a refusal to provide income information rather than a final answer of "0." This indicator of qualified/range answers excludes respondents previously coded as providing either item nonresponse or nonsubstantive answers.

Our main indicator of heaping was coded as a 1 if the response was a multiple of 5, and 0 if it was not. As respondents can also heap answers based on the calendar time in a question's reference period (Huttenlocher et al. 1990), for the five questions with a reference period of "in the last week," we also create a heaping indicator coded as a 1 if the response was a multiple of 7 (i.e., number of days in a week), and 0 if it was not. Both heaping measures exclude cases coded as a 1 for any previous indicator.

Our final dependent variable is the substantive responses to each question (including ranges/qualified answers). To account for outliers, we calculated the 99th percentile for each question, and replaced answers above the 99th percentile with that question's 99th percentile value.

3.1.2. Telephone Survey

In the telephone survey, our indicator of item nonresponse is coded as 1 if a respondent's final answer was "don't know" or a "refusal." All other answers are coded as 0.

Since instances of nonsubstantive and qualified/range answers are often resolved by telephone interviewers and not reflected in final answers, we use behavior coding to identify if these answers ever occurred during the question administration. Behavior coding is a systematic, objective method for identifying deviations from a paradigmatic "question asked/question answered" interviewer/respondent interaction (Fowler and Cannell 1996; Schaeffer and Maynard 1996). For each interview, we transcribe administrations of the three telephone questions at the conversational turn level (i.e., a period of uninterrupted speech by an interviewer or respondent that ends when an actor stops speaking or is interrupted by another actor). Trained undergraduates behavior-coded these turns using Sequence Viewer (Dijkstra 1999). To assess inter-coder reliability, two master coders also coded a 10% random subsample of the transcripts. Kappa values ranged from 0.54 to 0.81, all above the common 0.40 cutoff (Bilgen and Belli 2010).

We create an indicator for whether the respondent *ever* gave a nonsubstantive answer for each question using the behavior codes. For each question administration, we code the indicator as 1 if a respondent *ever* gave a non-integer answer to the question that could not be recoded (e.g., saying "I don't know," "a whole lot," or refusing to answer), regardless of whether they were an item nonrespondent or gave a final answer. All other answers were coded as 0. We similarly use behavior coding to create an indicator for qualified/range answers for each question coded as 1 if a respondent *ever* gave a non-integer answer to the question that could be recoded and 0 otherwise.

For the remaining indicators we exclude all cases coded as 1 for item nonresponse and focus only on those who gave a response. We create two indicators for heaping using final answers: one is coded 1 if a response is a multiple of 5 (to capture common rounding behavior) and 0 otherwise, and the other is coded 1 if a response is a multiple of 7 (because these questions use "week" as a reference period) and 0 otherwise. Means are again examined using the substantive integer responses to each question, accounting for outliers using the same 99th percentile method as the mail survey.

We operationalize the question administration length (i.e., an indicator of the question's burden) in two ways. First, the total number of conversational turns for each question is used as an indicator of administration burden overall. Second, we calculate the number of seconds it takes to reach a final answer after the interviewer has finished reading a question. We do this by summing the length of each conversational turn that occurs after the interviewer's first question-asking turn. To account for skew in this response time measure, we truncate response times below the first percentile and above the 99th percentile for each question (Yan and Olson 2013), and then use a natural log transformation.

3.2. Independent Variables

Our focal independent variable in both surveys is an indicator variable of whether a respondent was randomly assigned to receive an invitation to estimate their answer to a question (= 1) or not (= 0). The randomization in both modes was at the respondent level – for each respondent, all of the numeric questions examined here either included the invitation to estimate or did not include the invitation to estimate.

Our next set of independent variables capture key question characteristics: the likely cognitive state of information for the average person requested by the question and question sensitivity. These question characteristics (as well as those listed in the Controls section) were independently coded by two coders (two of the authors), and were coded relative to their perceptions of how the average person would view these questions (Online supplemental material (B) displays the questions used to rate these characteristics), as one might do when designing a questionnaire. Kappa values for all question characteristics were above 0.70 except for cognitive state and sensitivity. Disagreements between the two coders were resolved by a third coder (a third author) to create the final set of codes used in the analyses (Summarized in Table 2).

In the mail mode only, we use these codes to create three dichotomous indicators for the cognitive state of information requested by the question (available, accessible, and generatable) for the average person. No questions were coded as inestimable. We also create an indicator that describes the sensitivity (sensitive = 1; not sensitive = 0) of the question. In the telephone mode, because there are only three items included in this experiment, we use these coded characteristics for interpretation only (i.e., not as independent variables).

Table 2. Summary of question characteristics.

	NHWPS – mail		WLT2 – telephone	
	n	Percent/mean	n	Percent/mean
Cognitive state				
Available	8	38.10%	2	66.66%
Accessible	4	19.05%	0	0.00%
Generatable	9	42.86%	1	33.33%
Sensitivity				
Not sensitive	12	57.14%	1	33.33%
Sensitive	9	42.86%	2	66.66%
Reference period/similarity				
Short/similar	5	23.81%	3	100.00%
Long/dissimilar	16	76.19%	0	0.00%
Frequency				
Low	17	80.95%	1	33.33%
High	4	19.05%	2	66.66%
Regularity				
Regular	6	28.57%	1	33.33%
Irregular	15	71.43%	2	66.66%
Reading level		8.76		5.03
Number of words in stem		13.19		14.33

3.3. Controls

Question characteristics associated with a respondent's choice of an enumeration or estimation response strategy may affect many of the data quality indicators in this study as well as may influence the average respondent's cognitive state on a question (e.g., Blair and Burton 1987; Burton and Blair 1991; Conrad et al. 1998). Therefore, in the mail mode, we code and control for each question's reference period length (short = 0; long = 1); frequency (low = 0; high = 1), regularity (regular = 0; irregular = 1), and similarity (similar = 0; dissimilar = 1) of episodes of the event asked about; and number of words in the question stem and reading level using the Flesch-Kincaid grade level (both calculated excluding the phrase "Your best estimate is fine," then grand-mean-centered). Each question's reference period and similarity rating were perfectly collinear, so we analyze these two characteristics together (i.e., short/similar = 0; long/dissimilar = 1).

Likewise, decreased cognitive ability may negatively affect a respondent's ability to retrieve information from memory and thus also affect some of the data quality indicators examined here (i.e., skipping a question or estimating instead of enumerating; Krosnick 1991; Knäuper et al. 1997). In both the mail and telephone surveys, we control for respondents' age and level of education (two common proxies for cognitive ability) (Table 3). We multiply imputed missing values for age (12%) and education (6%) in the mail survey; due to the low item nonresponse rates in the telephone survey, we used grand mean imputation (for age, 4%, n = 34) and modal category imputation (for education, 1%, n = 5) in the telephone survey. The models include a grand-mean-centered continuous measure of age; level of education is represented by three dichotomous variables (high school or less, some college, or college graduate or higher).

Table 3. Summary of unimputed respondent characteristics.

	Overall	Without invitation to estimate	With invitation to estimate	t-test/ chi-square
NHWPS – mail				
Number of respondents	1,002	522	480	
Age (in years)	57.35	57.47	57.22	0.218
Education				
High school graduate or less	20.58%	20.25%	20.94%	2.797
Some college	31.98%	29.86%	34.30%	
College graduate or more	47.44%	49.90%	44.37%	
WLT2 – telephone				
Number of respondents	899	451	448	
Age (in years)	54.13	53.73	54.54	0.644
Education				
High school graduate or less	31.32%	31.03%	31.61%	6.022
Some college	26.29%	29.69%	22.87%	
College graduate or more	42.39%	39.29%	45.52%	

Note: There were no significant differences in respondent age or education across questionnaire version in either survey

3.4. Analysis

3.4.1. Mail Survey

We analyze our event/behavior questions (n = 21) together and our income question (n = 1) by itself. For the event/behavior questions, because the invitation to estimate was assigned at the respondent level (i.e., no variation within respondents on the experimental condition; for each respondent, all questions included the invitation to estimate or did not), we analyze each of our dichotomous data quality indicators using a population-averaged model with an exchangeable correlation structure (`xtgee` command in Stata 14.2 with respondents set as the clustering variable) (Agresti 2002; McNeish et al. 2017; Raudenbush and Bryk 2002; Rabe-Hesketh and Skrondal 2012; West et al. 2015). We use a logit link function because we have dichotomous outcomes. Clustering of questions within respondents is accounted for using cluster-robust standard errors, estimated using Huber-White sandwich estimators; the multivariate analysis of the mail survey also accounts for multiple imputation of the age and education variables. Age and education were multiply imputed $D = 5$ times (using adjustment cell random hotdeck imputation via `hotdeckvar` in Stata; Schonlau 2018) and combined for analysis using Rubin's Rules (via Stata's `micombine`).

We start our analyses by exploring the bivariate relationship between the invitation to estimate and each of our dichotomous data quality indicators (i.e., item nonresponse, nonsubstantive answers, qualified/range answers, and heaped answers). We estimate four population-averaged models, each predicting a different dichotomous quality indicator with our invitation to estimate indicator.

We then move to a multivariate framework, adding our remaining independent variables (each question's cognitive state and sensitivity) and controls (reference period/similarity, frequency, regularity, number of words, reading level, respondent age and education) to each of the bivariate models mentioned above. To test our moderation hypotheses, we estimate two new models per data quality indicator; each model includes an interaction term between the invitation to estimate indicator and a different question characteristic (i.e., an interaction with cognitive state in one model, an interaction with sensitivity in another). To interpret significant interactions, we calculate predicted probabilities using the `margins` command in Stata, holding all other variables at their means. Results from our main effects models and models with significant interactions are displayed in-text; results for models with non-significant interactions are displayed in the Online supplemental material

Our final analyses examine unweighted means for each behavior question across our two experimental conditions. For the bivariate tests, we estimate 21 ordinary least squares (OLS) regression models (one per question) predicting each question's average response with the invitation to estimate indicator entered as a predictor. We also tested this relationship using count models (e.g., Poisson and negative binomial regression), and our substantive conclusions do not change. We therefore describe the OLS results for ease of interpretation. For the multivariate models, we add respondent age and education as controls. For these analyses, one model is estimated per question so question characteristics (e.g., cognitive state, sensitivity) are not included as predictors.

For the income question, we conduct bivariate analyses using regression models predicting each of our data quality indicators (logistic regression for dichotomous outcomes, OLS regression for continuous outcomes) with the invitation to estimate as a predictor. We confirm our bivariate findings using multivariate models with our two respondent controls, age and education, as predictors.

3.4.2. Telephone

As the telephone portion of our study contains only three questions, we do not have enough observations within respondents to estimate the population-averaged models. Thus, we examine the data quality indicators separately for each question. Additionally, each respondent is nested within an interviewer, yielding a multilevel data structure (Hox, 1994; Olson and Bilgen, 2011; Olson and Peytchev, 2007). All telephone analyses account for this clustering using Stata's complex survey design procedures (`svy` procedures).

For our bivariate analyses, we estimate design-adjusted regression models predicting each of our data quality indicators (logistic regression for dichotomous outcomes, OLS regression for continuous outcomes) for each question with the invitation to estimate as a predictor. For the multivariate models, we add our two respondent controls, age and education, as predictors.

4. Results

4.1. Mail

Table 4 shows results of the bivariate tests of the relationship between an invitation to estimate and our dichotomous data quality indicators for our 21 event/behavior questions. Rates of item nonresponse, qualified/range answers, and heaped answers (multiples of 5 or 7) did not significantly differ across questions with and without an invitation to estimate ($p > 0.05$). However, question administrations with an invitation to estimate had significantly fewer nonsubstantive answers (0.78%) than those without the invitation (1.58%), although the difference is less than a percentage point (diff = 0.80%; $z = -2.10$, $p < 0.05$).

In the multivariate models, item nonresponse rates are again not significantly different across question administrations with and without an invitation to estimate ($p = 0.39$; Table 5). This indicates that, despite conventional wisdom and our H1a, inviting respondents to estimate does not reduce item nonresponse rates. Further, the interactions between the invitation to estimate indicator and our two independent variables (i.e., cognitive state and question sensitivity) were not significant ($p > 0.05$), indicating that the (non)effect of an invitation to estimate on item nonresponse did not differ across these question characteristics (i.e., hypotheses H1b and H1c were not supported). Table 6 summarizes the test statistics and p-values for each interaction tested in this study. Full results from models with non-significant interactions are displayed in Online supplemental material (Table C).

Consistent with H2a, an invitation to estimate did significantly reduce rates of nonsubstantive answers (e.g., "a lot" or "too many to count") (OR = 0.555, $p < 0.05$; Table 5), though these behaviors were rare overall. Predicted probabilities of nonsubstantive answers reduced by almost half from 0.014 without an invitation to 0.008 with an invitation to estimate. Explicitly accepting estimation, therefore, may communicate to respondents that a guess is preferred to admitting that the exact answer is not known.

Table 4. Percent of question administrations with each data quality indicator overall and by questionnaire version for mail.

Indicator	Overall		Without ItE		With ItE		Difference	
	%	SE	%	SE	%	SE	z/t-value	With ItE – Without ItE
Item nonresponse	5.68%	0.16%	6.19%	0.23%	5.14%	0.22%	-1.03	-1.05%
Nonsubstantive answers	1.20%	0.08%	1.58%	0.12%	0.78%	0.09%	-2.10	-0.80%*
Range or qualified answers	2.66%	0.11%	2.63%	0.16%	2.70%	0.17%	0.16	0.07%
Heaped/rounded answers								
Multiples of 5	60.74%	0.35%	60.36%	0.49%	61.15%	0.51%	0.81	0.80%
Multiples of 7								
Q51A: Working	44.78%	1.69%	44.52%	2.35%	45.06%	2.45%	0.16	0.54%
Q51B: Household work	15.05%	1.21%	15.89%	1.72%	14.15%	1.69%	-0.72	-1.74%
Q51C: Looking after family	56.11%	1.71%	57.24%	2.37%	54.90%	2.47%	-0.68	-2.34%
Q51D: Leisure activities	16.84%	1.27%	16.11%	1.74%	17.62%	1.86%	0.59	1.51%
Q51E: Sleeping	43.45%	1.69%	43.78%	2.34%	43.10%	2.44%	-0.20	-0.68%

SE: standard error; ItE: Invitation to Estimate

*$p < 0.05$

Table 5. Odds ratios and standard errors for models predicting data quality indicators for mail.

	Item nonresponse		Nonsubstantive answers		Nonsubstantive answers ItE x sensitivity		Qualified/range answers		Heaped/rounded answers (multiples of 5)	
	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)
Questionnaire version (ref = Without ItE)										
With ItE	0.813	(0.197)	0.555*	(0.153)	0.664	(0.185)	0.992	(0.186)	1.016	(0.076)
Question characteristics										
Cognitive State (ref = Available)										
Accessible	1.239***	(0.071)	1.874***	(0.207)	1.876***	(0.208)	2.030***	(0.269)	0.505***	(0.022)
Generatable	0.988	(0.063)	1.825***	(0.287)	1.832***	(0.290)	1.775***	(0.283)	0.377***	(0.019)
Sensitivity (ref = not sensitive)										
Sensitive	1.474***	(0.097)	0.851	(0.113)	1.068	(0.171)	0.962	(0.104)	2.363***	(0.143)
Controls										
Frequency (ref = Low)										
High	0.737***	(0.042)	2.289**	(0.631)	2.300**	(0.638)	1.131	(0.129)	1.306***	(0.079)
Reference Period/similarity (ref = short/similar)										
Long/dissimilar	0.316***	(0.062)	1.427	(0.365)	1.428	(0.367)	1.574**	(0.253)	1.202*	(0.096)
Regularity (ref = regular)										
Irregular	1.112	(0.164)	2.670**	(0.868)	2.702**	(0.880)	0.512***	(0.092)	0.422***	(0.043)

Table 5. Continued

	Item nonresponse		Nonsubstantive answers		Nonsubstantive answers ItE x sensitivity		Qualified/range answers		Heaped/rounded answers (multiples of 5)	
	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)
Question stem Reading level (GMC)	1.020*	(0.009)	1.043**	(0.016)	1.043*	(0.017)	1.054**	(0.016)	1.110***	(0.007)
Number of words in question (GMC)	0.986*	(0.006)	0.948***	(0.012)	0.948***	(0.012)	0.963***	(0.010)	0.997	(0.003)
Age (GMC)	1.033***	(0.011)	1.018	(0.012)	1.018	(0.012)	1.000	(0.006)	1.005**	(0.001)
Education (ref = HS or less)										
Some college	1.077	(0.536)	0.732	(0.297)	0.678	(0.269)	0.698	(0.190)	0.889	(0.066)
College +	0.860	(0.361)	0.324**	(0.141)	0.320**	(0.136)	0.688	(0.170)	1.037	(0.069)
Interactions										
ItE x question Sensitivity					0.468*	(0.140)				
N	21042		19846		19846		19608		19086	

OR: Odds Ratio; SE: standard error; ItE: Invitation to Estimate; GMC: grand-mean-centered
* $p < 0.05$; ** $p < 0.01$; ***$p < 0.001$

Table 6. Summary of test statistics and significance values for interactions by data quality indicator for mail.

Data quality indicator	χ^2	p
Item nonresponse		
ItE x Cognitive state	0.65	0.72
ItE x Sensitivity	1.96	0.16
Nonsubstantive answers		
ItE x Cognitive state	0.28	0.87
ItE x Sensitivity	6.44	0.01
Qualified/range answers		
ItE x Cognitive state	0.50	0.78
ItE x Sensitivity	0.49	0.48
Heaped/rounded answers (multiples of 5)		
ItE x Cognitive state	0.35	0.84
ItE x Sensitivity	0.05	0.83

ItE: Invitation to estimate

The effect of inviting estimation on nonsubstantive answers also differed across question sensitivity ($\chi^2 = 6.44$, $p < 0.05$; Table 6; supporting H2c), though the rates of occurrence remain low. For nonsensitive questions, there is no significant difference in the predicted probabilities of nonsubstantive answers between the version with the invitation to estimate (0.010) and the version without (0.014; Figure 1). However, for sensitive questions, the probability of nonsubstantive answers is significantly lower in the version with the invitation to estimate (0.005) than in the version without (0.015). This finding suggests that the invitation to estimate is encouraging some respondents to sensitive questions to provide an edited, usable answer rather than indicating that the question is sensitive by writing a nonsubstantive answer (e.g., "too revealing"). Contrary to H2b, the

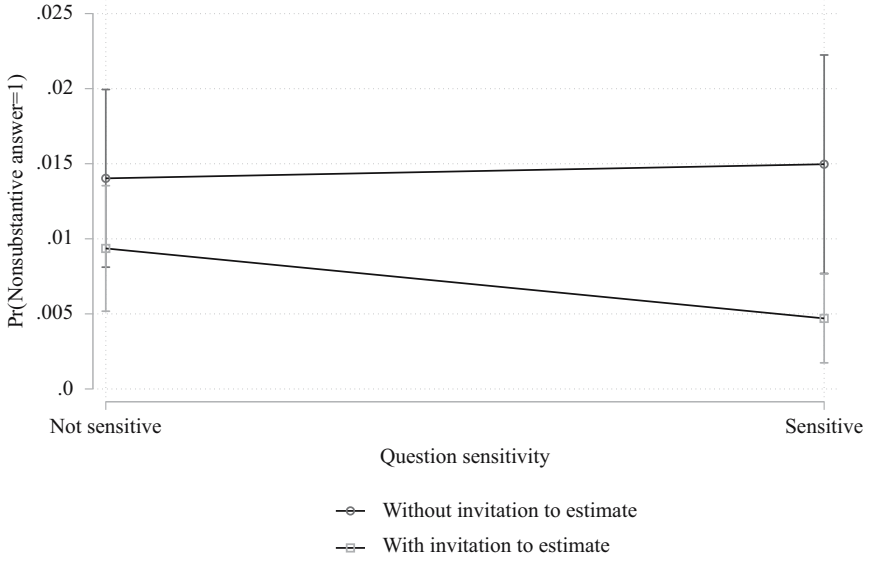


Fig. 1. Predicted probabilities for nonsubstantive answers by questionnaire version and question sensitivity.

interaction between the invitation to estimate indicator and cognitive state was not significant ($p = 0.87$; Table 6; full results in Online supplemental material, Table C).

Inviting estimation did not significantly affect the rate of qualified/range answers in the multivariate models ($p = 0.97$; Table 5). The invitation, therefore, did not help clarify the level of precision requested by the questions in the mail survey. Neither of the interactions between the invitation to estimate indicator and either cognitive state or question sensitivity were significant ($p > 0.05$; Table 6; full results in Online supplemental material, Table D). Thus, none of our H3a-H3c were supported.

Our next set of multivariate models demonstrates that respondents heaped their answers around multiples of 5 slightly more with an invitation to estimate, but the difference is not significant ($p = 0.36$; Table 5). The effect of an invitation to estimate on heaping does not differ across cognitive state or sensitivity ($p > 0.05$; Table 6; full results in Online supplemental material, Table D). Rates of heaping at multiples of 7 also do not significantly differ with or without an invitation to estimate in the multivariate models for the five questions with a reference period of a "week" ($p > 0.05$; Online supplemental material, Table E). Overall, we find no support for H4a-H4c.

Next, we examine substantive responses to the NHWPS questions. Table 7 displays means overall and by questionnaire version for each item. As expected, most event/behavior questions (76%) did have higher means with an invitation to estimate, but differences were generally small and none of the 21 bivariate tests were statistically significant ($p > 0.05$). These results remain unchanged when controlling for respondent age and education in the multivariate models (Online supplemental material, Tables F–H). Therefore, contrary to H5, any tendencies towards over-reporting when estimation was allowed were not large enough to cause significant differences in estimated means for these questions.

For our final set of analyses on the mail survey, we examine the effect of an invitation to estimate on an income question. We interpret and discuss our findings using the bivariate tests because the multivariate and bivariate results are the same for all indicators (multivariate results presented in Online supplemental material, Table I). We find that adding an invitation to estimate to an income question has no effect on rates of item nonresponse, nonsubstantive answers, qualified/range answers, or heaping ($p > 0.05$; Table 8). Further, we find that there are no significant differences in estimated means when an invitation to estimate is added versus when it is not ($p > 0.05$). This collection of results demonstrates that inviting respondents to estimate on an income question does little to affect data quality (i.e., no support for H1a-H4a or H5). It is possible that respondents assume that answers to yearly income questions should be estimated, as exact answers are typically not known; explicitly inviting estimation may not change response behaviors for this question.

4.2. Telephone

In the telephone survey, we explore the effect of an invitation to estimate across three behavioral questions asking about alcohol consumption and cigarette smoking in the past seven days, and number of miles driven in a typical week. We interpret and discuss our findings using only the design-adjusted bivariate tests because the multivariate and bivariate results are identical for all indicators across all questions, with one noted

Table 7. Item means and standard errors overall and by questionnaire version for mail and telephone.

Question	Overall		Without ItE		With ItE		Diff	
	Mean	SE	Mean	SE	Mean	SE	t-value	With ItE – Without ItE
NHWPS – Mail								
Q26A: New job	2.48	0.06	2.43	0.09	2.53	0.09	0.79	0.10
Q26B: Vacation	15.85	0.64	15.69	0.87	16.01	0.95	0.25	0.32
Q26C: Illness or accident	0.56	0.03	0.54	0.04	0.58	0.05	0.67	0.04
Q26D: Victim of crime	0.87	0.04	0.89	0.06	0.84	0.06	-0.61	-0.05
Q26E: Threatened	0.87	0.09	0.83	0.11	0.92	0.13	0.53	0.09
Q26F: Injured by someone else	0.25	0.03	0.23	0.04	0.27	0.05	0.73	0.04
Q26G: Unwanted sexual comments	1.95	0.20	1.92	0.29	1.97	0.28	0.12	0.05
Q26H: Bonus or promotion	4.92	0.22	5.10	0.31	4.73	0.32	-0.84	-0.38
Q27A: Witnessed other, illness or accident	7.46	0.71	6.97	0.91	7.98	1.11	0.71	1.02
Q27B: Witnessed other, victim of crime	1.37	0.08	1.31	0.11	1.44	0.12	0.80	0.13
Q27C: Witnessed other, threatened	1.24	0.09	1.28	0.12	1.19	0.14	-0.46	-0.09
Q27D: Witnessed other, injured by someone else	1.07	0.11	1.00	0.13	1.15	0.17	0.69	0.15
Q27E: Witnessed other, unwanted sexual comments	2.15	0.22	1.94	0.27	2.37	0.35	0.97	0.43
Q28: Told by other, victim of crime	4.57	0.25	4.34	0.34	4.83	0.36	1.00	0.49
Q29: Head injury w/o loss of consciousness	0.95	0.05	0.92	0.07	0.98	0.08	0.62	0.07
Q30: Head injury w/ loss of consciousness	0.41	0.03	0.40	0.04	0.41	0.04	0.20	0.01
Q51A: Working	23.44	0.74	23.87	1.05	22.97	1.06	-0.61	-0.90
Q51B: Household work	10.52	0.35	10.45	0.49	10.60	0.50	0.23	0.16
Q51C: Looking after family	12.23	0.96	12.24	1.33	12.21	1.39	-0.01	-0.03
Q51D: Leisure activities	19.00	0.57	18.98	0.79	19.02	0.81	0.04	0.04
Q51E: Sleeping	41.85	0.61	41.56	0.86	42.18	0.88	0.51	0.62
WLT2 – Telephone								
Q22: Alcohol	2.89	0.24	3.09	0.37	2.69	0.31	-0.96	-0.40
Q23: Cigarettes	12.72	1.18	13.23	1.69	12.21	1.66	-0.42	-1.02
Q31: Miles driven	172.23	7.76	173.18	10.93	171.29	11.04	-0.15	-1.89

SE: standard error; ItE: Invitation to Estimate

Note: There were no significant differences in means across questions with and without an invitation to estimate.

Table 8. Percent of question administrations with each data quality indicator and means overall and by questionnaire version for income question in mail.

Indicator	Overall		Without ItE		With ItE		Difference	
	%	SE	%	SE	%	SE	z/t-value	With ItE – Without ItE
Item nonresponse	17.76%	1.21%	19.16%	1.72%	16.25%	1.69%	-1.07	-2.91%
Nonsubstantive answers	4.98%	0.76%	5.69%	1.13%	4.23%	1.00%	-0.96	-1.46%
Range and qualified answers	5.24%	0.80%	5.53%	1.15%	4.94%	1.11%	-0.29	-0.59%
Heaped/rounded answers (Multiples of 5)	96.23%	0.70%	95.21%	1.10%	97.27%	0.85%	1.47	2.05%
	Mean	SE	Mean	SE	Mean	SE		
Q62: Income	80,045.04	2,486.35	80,519.19	3,644.50	79,554.88	3,377.39	-0.19	-964.31

SE: standard error; ItE: Invitation to Estimate

exception. For completeness, results from the multivariate models are presented in Online supplemental material (Tables J–Q).

Like the mail survey (i.e., contrary to H1a), we find that the relationship between item nonresponse and an invitation to estimate is not significant for any of the three questions in the telephone survey ($p > .05$; Table 9). As expected, interviewers likely motivated respondents to provide an answer across both questionnaire versions, nullifying the effect of an invitation to estimate on item nonresponse.

The percentage of respondents who ever gave a nonsubstantive answer (e.g., "I smoke a lot of cigarettes") was lower for questions with an invitation to estimate in the telephone survey, but this difference was only significant for the question asking about cigarettes ($t = -2.12$, $p < .05$; Table 9). However, this difference becomes not significant when controlling for respondent age and education ($p = .07$). Inviting estimation also did not significantly change the percentage of cases with at least one qualified/range answer (e.g., "between 5 and 10 drinks a week") for questions asking about alcohol or cigarettes ($p = 0.28$; Table 9).

When asking about number of miles driven in a typical week, cases with qualified/range answers were, however, significantly lower when the question included an invitation to estimate (47.33%) compared to when it did not (59.74%; diff = 12.41%; $t = -3.70$, $p < .01$; Table 9; partial support for H3a). This may be because the number of miles driven in a typical week is, by far, the highest-frequency behavior in either the mail or telephone survey (overall mean = 172.23 miles; Table 7) and cannot be reported exactly by most respondents (i.e., generatable information). Without an invitation to estimate, respondents may feel the need to notify the interviewer that they are uncertain of their answer. Including an invitation to estimate, however, may communicate that an imprecise number is acceptable, thus reducing respondent expressions of uncertainty.

Contrary to H4a, an invitation to estimate also does not significantly affect the percentage of respondents that heap their answers around multiples of 5 or 7 ($p > 0.05$; Table 9). For the alcohol and cigarette questions, this may be because these items are sensitive, making respondents more likely to partially conceal their answers via heaping with or without an invitation to estimate. Additionally, since the number of miles driven in a typical week generally cannot be enumerated for many respondents, answers to this question are likely heaped with or without permission to estimate.

As in the mail survey, we find that means to the three telephone questions do not differ with or without an invitation to estimate, contrary to H5. These results again indicate that an invitation to estimate does not trigger any meaningful changes in the magnitude of respondents' answers.

Finally, we examine two related indicators of question administration length that are unique to the telephone mode. We find that adding an invitation to estimate significantly reduces the number of conversational turns required to reach a final answer by about one turn for all three questions ($p < .05$; Table 9). Eliminating this extra turn significantly reduces response time by about one second for the question asking about alcohol ($t = -2.41$, $p < .05$), and by about two seconds for the questions asking about cigarettes ($t = -2.89$, $p < .01$) and miles driven ($t = -5.84$, $p < .001$). While an invitation to estimate may not affect the final answers, it does seem to simplify the interaction between interviewers and respondents required to achieve those final answers, supporting H6.

Table 9. Percent of question administrations with each data quality indicator overall and by questionnaire version for telephone.

	Overall		Without ItE		With ItE		Diff	
	%	SE	%	SE	%	SE	z/t-value	With ItE – Without ItE
Item Nonresponse								
Q22: Alcohol	0.89%	0.42%	0.89%	0.47%	0.89%	0.69%	0.010	0.01%
Q23: Cigarettes	0.56%	0.23%	0.67%	0.33%	0.45%	0.32%	-0.450	-0.22%
Q31: Miles driven	2.96%	0.87%	3.38%	1.48%	2.56%	0.89%	-0.500	-0.82%
Ever gave nonsubstantive answer								
Q22: Alcohol	6.90%	0.80%	7.78%	1.19%	6.01%	1.07%	-1.10	-1.76%
Q23: Cigarettes	4.49%	0.78%	6.08%	1.02%	2.91%	0.90%	-2.12*	-3.17%
Q31: Miles driven	13.24%	1.53%	15.32%	2.01%	11.20%	2.05%	-1.40	-4.13%
Ever gave qualified/range answer								
Q22: Alcohol	16.35%	1.20%	15.78%	1.35%	16.93%	1.97%	0.48	1.15%
Q23: Cigarettes	11.90%	1.18%	13.51%	1.48%	10.29%	1.77%	-1.34	-3.22%
Q31: Miles driven	53.47%	2.12%	59.74%	2.28%	47.33%	2.40%	-3.70**	-12.41%
Heaped final answers								
Multiples of 5								
Q22: Alcohol	66.33%	1.22%	67.11%	1.68%	65.54%	1.71%	-0.66	-1.57%
Q23: Cigarettes	97.20%	0.58%	96.65%	0.71%	97.76%	0.93%	0.86	1.11%
Q31: Miles driven	98.01%	0.47%	98.12%	0.58%	97.90%	0.73%	-0.24	-0.22%
Multiples of 7								
Q22: Alcohol	61.84%	1.38%	63.76%	1.62%	59.91%	2.04%	-1.48	-3.85%
Q23: Cigarettes	87.47%	1.20%	88.39%	1.56%	86.55%	1.69%	-0.80	-1.85%
Q31: Miles driven	9.03%	1.01%	9.14%	1.31%	8.92%	1.53%	-0.11	-0.22%
Number of turns								
Q22: Alcohol	3.69	0.16	4.04	0.22	3.33	0.14	-2.71*	-0.71
Q23: Cigarettes	3.69	0.17	4.17	0.26	3.22	0.08	-3.54**	-0.96
Q31: Miles driven	4.59	0.22	5.02	0.25	4.16	0.28	-2.27*	-0.86
Response time (secs)								
Q22: Alcohol	6.15	0.29	6.73	0.32	5.57	0.36	-2.41*	-1.16
Q23: Cigarettes	5.13	0.37	6.12	0.56	4.15	0.22	-2.89**	-1.97
Q31: Miles driven	13.01	0.39	14.04	0.41	12.00	0.54	-5.84***	-2.03

SE: standard error; ItE: Invitation to Estimate

* $p < 0.05$; ** $p < 0.01$; ***$p < 0.001$

5. Discussion

In this study, we evaluate the common assumption that inviting respondents to estimate when responding to numeric questions about events, behaviors, or income reduces burden. We explore the effect of an invitation to estimate on several indicators of data quality in both a mail and a telephone survey. This is the first study to our knowledge that investigates this questionnaire design choice in either mode. We have three main takeaways from our analyses.

First, we find no support for the idea that an invitation to estimate significantly reduces item nonresponse in either a mail or a telephone survey. Notably, this effect did not differ across the question characteristics in our mail study, meaning that even the most burdensome or sensitive questions did not benefit from inviting estimation. We do find that inviting estimation reduces nonsubstantive answers (i.e., answers that cannot be re-coded and would likely be set to item missing) for the event/behavior questions in our mail survey, especially when they ask about sensitive topics. However, the rate of these answers is quite small, and we do not see the same pattern for our mail income question or for any of the telephone questions. It is possible that the burden of numeric questions is not high enough to warrant item nonresponse even when respondents enumerate, making an invitation to estimate unnecessary. Alternatively, respondents may intuit that an estimated answer is preferable to item nonresponse when an exact answer is difficult or impossible to provide; they may not need encouragement to make a shift from enumeration. Future work could explore this further by, for example, investigating whether the effect of an invitation to estimate differs across respondent characteristics that may be associated with choice of retrieval strategy (e.g., education and age).

Second, if an invitation to estimate is used, researchers have little reason to worry that it will negatively affect data quality. Although the difference was not significant, questions with an invitation to estimate had slightly lower rates of item nonresponse than questions without the invitation. We also find that rates of qualified/range answers are unchanged with and without an invitation to estimate in the mail mode. Any measurement error introduced by re-coding answers like ranges into a single integer is not likely increased by inviting estimation. Further, explicitly accepting estimation does not change substantive responses: we find no effect of an invitation to estimate on rounding or means in either mode. This provides further evidence for the notion that respondents make decisions about when to use enumeration versus estimation on their own, and do not require an invitation to employ one strategy over the other.

Third, we find that an invitation to estimate may have some utility in simplifying the interaction between interviewers and respondents in CATI surveys. For example, inviting estimation reduced the percentage of respondents that gave qualified/range answers to a question asking about the number of miles driven in a typical week (i.e., a high frequency question that would be difficult to enumerate). Further, inviting estimation reduced the total number of turns required to achieve an acceptable final answer and reduced the length of response times for all three CATI questions. Although respondents ultimately provided an acceptable final answer in each of these cases, the invitation to estimate may have reduced the extra work interviewers had to perform (e.g., probing, providing clarifications) to obtain an acceptable answer. Therefore, these results are encouraging for using an invitation to estimate to reduce interactional burden in telephone surveys.

Future research should investigate whether the reduced response times found in the telephone survey can be replicated in a mail survey. Though such a study would likely be restricted to an observational, lab setting, these results would provide more insight into whether and how an invitation to estimate reduces respondent burden. It is also possible that some respondents do not actually read an invitation to estimate when it is included in visual surveys. Eye tracking studies could be used to determine if an invitation to estimate: (1) is seen by respondents, and (2) affects data quality when it is seen.

Although we experimentally examined an invitation to estimate in both a mail and a telephone survey, this was not an experimental study of mode differences. These two surveys had different questions and were fielded at different times. Instead, this study does provide a foundation for future work testing an invitation to estimate across these modes. Such a study would also provide insight into best practices for using an invitation to estimate in mixed-mode (e.g., interviewer- and self-administered) surveys. Further, the CATI survey only included three questions, which limits our ability to make strong conclusions about inviting estimation in the telephone mode. We also did not include a question asking for detailed financial information in the telephone survey. Future work should test the effects of the invitation to estimate on more questions asked in a telephone survey, including questions that vary on the theoretically motivated question characteristics identified here.

This study should also be replicated on the web; previous work in this mode (Couper et al. 2011) did not include all of the data quality indicators we explore here. We would expect that an invitation to estimate would have similar null effects across mail and web surveys, as both are self-administered and use visual channels of communication. Confirming these findings in a web survey that includes mobile respondents would allow questionnaire designers to have insights into whether one can drop the lengthy phrase "your best estimate is fine" and save space when displaying questions on the small screens of mobile devices. While this study investigated more than 20 questions, the characteristics of these questions were somewhat limited. For example, many of our questions asked about events that were unlikely to occur to a person in their lifetime (e.g., head injuries; crime victimization) or about the amount of time spent on activities that may occur with some regularity (e.g., work for pay). Additionally, the reference period and similarity of events of our questions were confounded. Also, none of our questions asked for information in the inestimable cognitive state (a good thing for our respondents, but not as useful for evaluating these questions). Future studies should explore the effect of explicitly accepting estimation on a wider variety of question types.

We also note that these question characteristics – and notably cognitive state – were not rated by the respondents themselves. For example, a respondent who has taken several out-of-state vacations in their lifetime may have more difficulty enumerating their answer than a respondent who has only taken two such vacations. It is possible, therefore, that an invitation to estimate may operate differentially based on a respondent's perception of a question and the utility of the invitation to estimate, rather than the perception of an outside rater. In general, raters disagree about question characteristics (Bais et al. 2019), although the average rating of question characteristics across expert raters has been shown to be related to measurement error (Olson 2010). Although our raters evaluated questions based on the "average respondent," asking a more diverse set of raters to evaluate question

characteristics and having a more diverse set of questions may provide more insight into the conditions under which an invitation to estimate is effective.

Overall, we find that inviting estimation in a mail survey has no significant effect for most of the data quality indicators in this study. Based on this collection of evidence, we see little reason to recommend using an invitation to estimate on questions asking about the frequency of events and behaviors or income in the mail mode, especially if questionnaire space is limited. Though we also note that using an invitation to estimate does not appear to negatively affect data quality for mail surveys either, so researchers do not be concerned about data from past studies where an invitation has been used. We are more optimistic that using an invitation to estimate in telephone surveys simplifies the interaction between interviewers and respondents. We encourage future research to replicate these findings in interviewer-administered modes.

6. References

Agresti A. 2002. *Categorical Data Analysis, Second Edition*, Hoboken, NJ: John Wiley & Sons.

Bais, F., B. Schouten, P. Lugtig, V. Toepoel, J. Arends-Tòth, S. Douhou, N. Kieruj, M. Morren, and C. Vis. 2019. "Can Survey Item Characteristics Relevant to Measurement Error Be Coded Reliably? A Case Study on 11 Dutch General Population Surveys." *Sociological Methods & Research* 48 (2): 263–95. DOI: https://doi.org/10.1177/0049124117729692.

Bassili, J.N., and B.S. Scott. 1996. "Response Latency as a Signal to Question Problems in Survey Research." *Public Opinion Quarterly* 60 (3): 390–99. https://doi.org/10.1086/297760.

Beatty, P., and D. Herrmann. 2002. "To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J. Little: 71–85. New York: Wiley.

Bilgen, I., and R.F. Belli. 2010. "Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires." *Journal of Official Statistics* 26 (3): 481–505. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/comparison-of-verbal-behaviors-between-calendar-and-standardized-conventional-questionnaires.pdf.

Blair, E., and S. Burton. 1987. "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions." *Journal of Consumer Research* 14 (2): 280–288. DOI: https://doi.org/10.1086/209112.

Burton, S., and E. Blair. 1991. "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys." *Public Opinion Quarterly* 55 (1): 50–79. DOI: https://doi.org/10.1086/269241.

Conrad, F.G., N.R. Brown, and E.R. Cashman. 1998. "Strategies for Estimating Behavioural Frequency in Survey Interviews." *Memory* 6 (4): 339–366. DOI: https://doi.org/10.1080/741942603.

Couper, M.P., C. Kennedy, F.G. Conrad, and R. Tourangeau. 2011. "Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys." *Journal of Official Statistics* 27 (1): 65–85. Available at: https://www.scb.se/contentassets/ca21efb41-

fee47d293bbee5bf7be7fb3/designing-input-fields-for-non-narrative-open-ended-responses-in-web-surveys.pdf.

De Leeuw, E.D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21 (2): 233–255. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/to-mix-or-not-to-mix-data-collection-modes-in-surveys.pdf.

Dijkstra, W. 1999. "A New Method for Studying Verbal Interactions in Survey Interviews." *Journal of Official Statistics* 15 (1): 67–85. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-new-method-for-studying-verbal-interactions-in-survey-interviews.pdf.

Dillman, D.A. 2007. *Mail and Internet Surveys: The Tailored Design Method*. Hoboken, N.J: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons.

Draisma, S., and W. Dijkstra. 2004. "Response Latency and (Para) Linguistic Expressions as Indicators of Response Error." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M.P. Couper, J. Lesser, E. Martin, J. Martin, and E. Singer. New York: Wiley.

Fowler Jr, F.J., and C.F. Cannell. 1996. "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman: 15–36. San Francisco: Jossey-Bass/Wiley.

Holbrook, A.L., S. Anand, T.P. Johnson, Y. Ik Cho, S. Shavitt, N. Chávez, and S. Weiner. 2014. "Response Heaping in Interviewer-Administered Surveys: Is It Really a Form of Satisficing?" *Public Opinion Quarterly* 78 (3): 591–633. DOI: https://doi.org/10.1093/poq/nfu017.

Hox, J.J. 1994. "Hierarchical Regression Models for Interviewer and Respondent Effects." *Sociological Methods & Research* 22 (3): 300–318. DOI: https://doi.org/10.1177/0049124194022003002.

Huttenlocher, J., L.V. Hedges, and N.M. Bradburn. 1990. "Reports of Elapsed Time: Bounding and Rounding Processes in Estimation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16 (2): 196. DOI:.https://doi.org/10.1037/0278-7393.16.2.196.

Knäuper, B., R.F. Belli, D.H. Hill, and A.R. Herzog. 1997. "Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality." *Journal of Official Statistics* 13 (2): 181–99. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/question-difficulty-and-respondents39-cognitive-ability-the-effect-on-data-quality.pdf.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (3): 213–36. DOI: https://doi.org/10.1002/acp. 2350050305.

McNeish, D., L.M. Stapleton, and R.D. Silverman. 2017. "On the Unnecessary Ubiquity of Hierarchical Linear Modeling." *Psychological Methods.* 22 (1): 114–140. DOI: http://dx.doi.org/10.1037/met0000078.

Menon, G. 1993. "The Effects of Accessibility of Information in Memory on Judgments of Behavioral Frequencies." *Journal of Consumer Research* 20 (3): 431–440. DOI: https://doi.org/10.1086/209359.

Olson, K., and I. Bilgen. 2011. "The Role of Interviewer Experience on Acquiescence." *Public Opinion Quarterly* 75 (1): 99–114. DOI: https://doi.org/10.1093/poq/nfq067.

Olson, K., and A. Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly* 71 (2): 273–286. DOI: https://doi.org/10.1093/poq/nfm007.

Olson, K. 2010. "An Examination of Questionnaire Evaluation by Expert Reviewers." *Field Methods* 22 (4): 295–318. DOI: https://doi.org/10.1177/1525822X10379795.

Rizzo, L., J.M. Brick, and I. Park. 2004. "A Minimally Intrusive Method for Sampling Persons in Random Digit Dial Surveys." *The Public Opinion Quarterly* 68 (2): 267–274. DOI: https://doi.org/10.1093/poq/nfh014.

Rabe-Hesketh S., and A. Skrondal A. 2012. *Multilevel and Longitudinal Modeling Using Stata, Third Edition, Volume II: Categorical Responses, Counts, and Survival*, College Station, TX: Stata Press.

Raudenbush S.W., and A.S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Newbury Park, CA: Sage.

Schaeffer, N.C., and D.W. Maynard. 1996. "From Paradigm to Prototype and Back Again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman: 65–88. San Francisco: Jossey-Bass.

Schonlau, M. 2018. *HOTDECKVAR: Stata module for hotdeck imputation*, Statistical Software Components S458527, Boston College Department of Economics, revised 19 April 2022. Available at: https://EconPapers.repec.org/RePEc:boc:bocode:s458527.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. New York, NY: Cambridge University Press.

Tourangeau, R., and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83. DOI: https://doi.org/10.1037/0033-2909.133.5.859.

West B.T., K.B. Welch, and A.T. Galecki. 2015. *Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition*, Boca Raton, FL: CRC Press.

Yan, T., and K. Olson. 2013. "Analyzing Paradata to Investigate Measurement Error." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter: 73–96. New Jersey: John Wiley & Sons.

Received January 2021
Revised September 2021
Accepted May 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1125–1144, http://dx.doi.org/10.2478/JOS-2022-0048

Analyzing the Association of Objective Burden Measures to Perceived Burden with Regression Trees

Daniel K. Yang[1] *and Daniell S. Toth*[1]

Higher levels of perceived burden by respondents can lead to ambiguous responses to a questionnaire, item nonresponse, or refusals to continue participation in the survey which can introduce bias and downgrade the quality of the data. Therefore, it is important to understand what might influence the perception of burden in respondents. In this article, we demonstrate, using U.S. Consumer Expenditure Survey data, how regression tree models can be used to analyze the associations between perceived burden and objective burden measures conditioning on household demographics and other explanatory variables. The structure of the tree models allows these associations to easily be explored.

Our analysis shows a relationship between perceived burden and some of the objective measures after conditioning on different demographic and household variables and that these relationships are quite affected by different respondent characteristics and the mode of the survey. Since the tree models were constructed using an algorithm that accounts for the sample design, inferences from the analysis can be made about the population. Therefore, any insights could be used to help guide future decisions about survey design and data collection to help reduce respondent burden.

Key Words: nonparametric; nonresponse; respondent burden; sample design; survey data.

1. Introduction

The Consumer Expenditure Survey (CE) is a national survey conducted by the U.S. Bureau of Labor Statistics (BLS) to collect data on how American households spend their money. The collected data are used to estimate consumer expenditures, which are published twice a year, as well as to annually produce public-use microdata files to allow researchers to do their own analyses. This is the only federal survey that provides information on U.S. consumer expenditures as well as household income and demographic characteristics, making the data it collects critically important to government and private agencies examining the association of consumer expenditures and income to household characteristics. This type of analysis is used by economic policy makers to understand the effects of policy changes on households among diverse socioeconomic groups.

Importantly, CE data are inputs for producing the Consumer Price Index (CPI), a Principal Federal Economic Indicator, used by The Federal Reserve to help set U.S. monetary policy. The data are used to construct new "market baskets" of goods and services, determine the relative importance of components, and to derive cost weights for

[1] U.S. Bureau of Labor Statistics Office of Survey Methods Research, 2 Massachusetts Avenue Suite 5930, NE Washington D.C. 20212, U.S.A. Emails: yang.daniel@bls.gov and toth.daniell@bls.gov

the market baskets used in the calculation of the CPI. CE data are also used by the Department of Commerce for calculating the Supplemental Poverty Measure, by the Department of Agriculture for estimating the cost of raising a child, by the Internal Revenue Service for calculating alternate sales tax standard deductions and by the Department of Defense for determining cost-of-living allowances for military personnel.

Because of the essential role that CE data play in setting policy and in the managing of the U.S. economy, it is imperative that the quality of the data be maintained at the highest level possible. Lower response rate is one way that the quality of survey data can be degraded. A low response rate can potentially introduce response bias as well as increase the variability of statistics obtained from the data (Groves 2006). In addition, low survey response rates erode user confidence in the data.

For these reasons, the BLS has put a great deal of effort into maintaining a high response rate for its surveys. These efforts include the introduction of computer-assisted personal interviewing and an instrument to track interviewer contacts, as well as a redesign of the survey (Edgar et al. 2013a,b). Despite these changes the CE has observed a decay of the response rate to its surveys over the last two decades.

Figure 1 shows the response rates of the CE interview survey falling from a rate of around 80% to a rate of less than 50% over a 21 year period. Falling response rates is hardly unique to CE. Czajka and Beyler (2016) found that response rates were declining at a similar rate for all the U.S. federal surveys that they studied including the National Health and Nutrition Examination Survey and the Current Population Survey. The large dip in the CE response rate that occurred late in 2013 was due to the disruption of data collection and nonresponse follow-up efforts caused by the government shutdown.

One factor that may have an effect on response rates of a survey is the amount of burden a survey puts on respondents. Burden is something that is difficult to bear, worrisome, stressful, or oppressive. In survey research, burden is often thought of as the collection of all costs that the survey respondent incurs for responding to the survey, including loss of

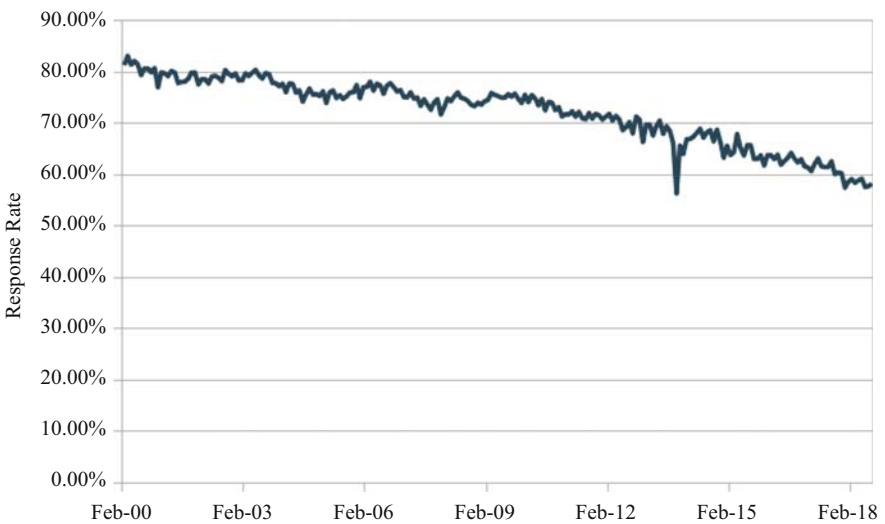


Fig. 1. Response rate of CE interview survey from january 2000 to june 2020.

time, exertion of effort, and stress associated with providing the requested information (Ashmead et al. 2017).

While how much stress or worry is felt by a respondent will depend on the individual, there are objective measures of burden that can be collected for each survey interview. These include length of the interview, number of questions asked, and whether information from records were required to answer the questions. Because of its breadth, the CE Quarterly Interview Survey (CEQ) is likely burdensome based on these measures, taking almost an hour to complete with many questions requiring the respondent to look through receipts or other records to answer.

Though these objective measures do not measure the actual burden felt by a respondent (perceived burden), one would expect that they are related and therefore have an effect on the response rate. However, evidence of the effect objective measures of burden have on the response rate of a survey have been mixed in the literature. Indeed, Bogen (1996) reviewed several observational and experimental studies of the relationship between questionnaire length and the response rate and found evidence both supporting and refuting a relationship. For instance, Lynn (2014) found no evidence that the initial interview length affected the participation rate for subsequent interviews while Galesic and Bosnjak (2009) did find evidence that the length of the questionnaire affected response rates and data quality. Nevertheless, time required to take a survey is definitely a cost that potential respondents weigh when deciding whether to participate in a survey and must be accounted for when measuring burden. Interview length was a significant input variable in the respondent burden model of Fricker et al. (2014).

These mixed results may be attributed, in part, to the fact that it is the perceived burden that affects a potential respondent's decision whether or not to participate and how a respondent reacts to the objective measures of burden vary for different respondents (Sharp and Frankel 1983). The amount of burden felt (or perceived) by a respondent is likely not perfectly correlated with the objective measures of burden like length and difficulty of the survey, but rather an interaction between these measures and characteristics of the respondent. For example, though Fekete et al. (2017) do not find a strong correlation between objective and subjective burden, they find that high subjective burden was linked to poorer general health. Though not tested in their report, this result could indicate that respondents with health issues are more sensitive to the objective measures of burden than other respondents.

Though a short survey is likely to have less perceived burden than a long survey in general, which surveys are judged to be short, and which are judged to be long will depend on the individual doing the judging, on their current circumstances, on their interests and on the topic of the survey, among other things. Likewise, which questions are difficult to answer can vary considerably among people depending on their household and personal situation at the time they are participating in the survey. Therefore, a measure of the amount of perceived burden a survey is likely to impose along with the objective measures of burden is needed to determine how the response rate of a survey is likely to be affected by the objective measures of burden.

Another factor found to influences the relationship between perceived burden and quantitative measures of burden is the participant's impression of the importance of the survey or it's salience. Salience is defined as "the quality of being particularly noticeable

or important" (Cannell et al. 1981). Survey researchers have long known that how salient the topic of a survey or particular questions are to a respondent can affect the response rates. For instance, Bradburn (1978) found that a boring questionnaire may drive respondents away, while an interesting one may motivate them.

Bradburn (1978) explained that survey interviews are social interactions and researchers must first understand the respondent's motivation to participate in a survey interview. He recommended using a sense of civic duty and knowledge about the importance of the survey as motivators for participation, putting emphasis on how the questionnaire design can contribute or mask burden of a survey; noting that burden appeared to become more tolerable for respondents who are persuaded that the collected data are crucial.

Groves et al. (2000) found evidence that respondents maybe more willing to participate in a burdensome cognitive activity for a salient topic than for a non-salient one. Groves et al. (2001) used measures of opportunity cost and salience as components of social interaction. Likewise, Connelly et al. (2003) studied mail survey response rates and found the salience of the study topic was among the five most significant explanatory variables in their model for response rate. They found a 25% increase in response rates for the highly salient survey compared to the non-salient survey. Fricker et al. (2014) also suggests that the perception of burden was associated with non-salient (less motivated) topics.

For example, participating in a survey on early childhood education that takes a half hour to complete might feel like no burden to a respondent who is concerned about their young children's education, but may seem like a big inconvenience to someone without young children. Therefore, Cannell et al. (1981) suggest that helping a respondent recognize that the intent of the survey is important to them could mitigate survey burden.

For a given survey, the amount of perceived burden that responding to it will induce depends on the current circumstances of the individual, the quantitative measures of burden for the survey, as well as the survey's salience to that individual. This suggests that to estimate the likely burden of a survey, one must account for the characteristics of the individual respondent as well as the objective measures of burden and that this relationship between the factors could be complicated.

Getting an accurate understanding of how perceived burden for a given respondent is affected by changes to the survey design is very important to CE. This could help drive future changes to the survey that reduce burden and slow or stop the decline of response rates. Besides reducing response rates, there is evidence that higher levels of burden are associated with measurement error which also impact data quality (Abayomi et al. 2018; Ashmead et al. 2017). Therefore, understanding and measuring what impacts the burden felt by the respondent is necessary for mitigating the burden put on respondents which is critical for maintaining the quality of the data and the sustainability of the survey.

In order to properly understand the relationship between perceived burden and the objective measures of burden using data collected as part of the survey, one must account for both the differences in respondents as well as the survey's sample design. In this article we demonstrate the use of regression tree models to study relationships between perceived burden and objective burden measures conditioned on characteristics of the survey respondent using CE data. Regression trees are an easily interpreted nonparametric conditional model type that can make it easy to understand interaction effects. In this case the interactions between objective and subjective burden and survey, respondent, and

household characteristics. To obtain these models, we use the $\mathscr{R}$ package **rpms** which builds regression trees that account for the sample design in their estimation (R Core Team 2020; Toth 2020).

When trying to assess the amount of perceived burden a survey puts on a participant, objective measures such as measures of the time and effort required to complete the survey are often used as a proxy (Rolstad et al. 2011). Our analysis will show that the effect of these objective measures of burden on the perceived burden are different for different households in the CE interview survey. Our findings also show the effect of respondent characteristics and the mode of the survey on these relationships.

If the proposed models could "predict" the perceived burden outcome variable accurately conditionally on the household data collected during the first wave of collection, the predicted values could be used to warn of respondents who are likely to experience high levels of burden which could lead to nonresponse or data quality issues in future waves. This could potentially allow survey administrators to intervene before the next data collection or make changes to collection procedures to head off potential issues.

The rest of the article is organized as follows. The CE survey data and how it is collected is described in Section 2. This section includes a description of the variables in the data set used for the analysis, including the created perceived burden measure. Section 3 contains the regression tree model analysis showing how the relationship between subjective burden and the objective burden measures depend on the household and demographic characteristics of the respondent. A discussion of the study results and their potential utility in future survey designs is contained in Section 4.

2. CE Interview Survey Data

The CE data on household spending, demographic and socioeconomic characteristics are collected through two separate household surveys, the CEQ and the Diary Survey. While the Diary Survey collects data on smaller purchases and irregular expenditures, the CEQ is designed to collect data on large and recurring expenditures that consumers can be expected to recall for at least three months, such as rent and utilities. Together, the data from the two surveys cover the complete range of consumer expenditures.

The interview survey is conducted through a structured questionnaire using one of two collection modes: personal visit or telephone. Households are selected to be in the panel using a two-stage cluster sample of addresses where the clusters are geographical regions defined by groups of counties. The 91 clusters are selected using a PPS (probability-proportional-to-size) sample with 23 certainty units and then addresses are randomly selected within each chosen cluster.

Data are collected from the sample of households over four waves. During the first wave, a field representative collects the demographic and social-economic characteristics of the household and the spending during the previous month to use as a baseline. Expenditure data is collected for each household in the sample using a multiple panel questionnaire during the second, third and the final waves. See Yang (2019) and U.S. Bureau of Labor Statistics (2018) for more information on the sample design and data collection procedures.

The response rate is defined as the proportion of eligible sampled housing addresses from which usable interviews were obtained. A sampled housing address is determined to

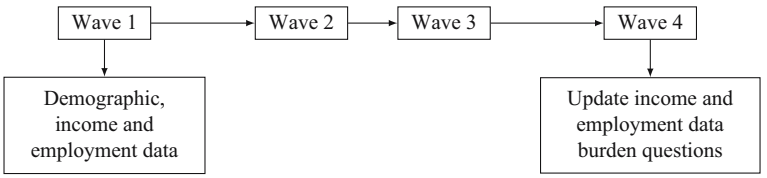


Fig. 2. Illustration of the CE interview survey process.

be ineligible if the house is vacant, under construction, destroyed, abandoned, converted to nonresidential use, or contains temporary residents.

The CE interview survey collects household expenditure data over four waves occurring every three months. In the first wave, demographic questions about the household are asked. Also in the first wave, income and employment information is collected. This information is then updated in the final wave of the survey. After completing the interview in the fourth wave, the respondent was asked four questions aimed at measuring the amount of burden they felt as a result of taking the survey. The burden questions are not part of the usual CE interview survey process and were only asked between the April 2017 and March 2018 study period. In general, the response rate of the interview survey tends to drop from wave 1 to wave 4.

The length and difficulty of a survey is likely to contribute to the attrition of respondents between the first and fourth waves (Kashihara and Ezzati-Rice 2004; Young et al. 2006; Gustavson et al. 2012). For instance, CE lost 22.5% of respondents to the interview survey between wave 1 and wave 4 of the survey in 2017 (Yang 2018). Because the attrition of respondents could impact data quality, it is important to try to reduce attrition by mitigating the burden that the survey imposes on respondents as much as possible through changes to the design and/or collection methods of the survey (Kashihara and Ezzati-Rice 2004; Baird et al. 2008; Cohen et al. 2013).

In order for CE to make meaningful changes to their data collection efforts, it is important to monitor and understand what causes a respondent to feel burdened by the survey and whether or not variables usually thought to be associated with burden, the objective measures of burden, are related to the amount of perceived burden actually felt by respondents. It is important to account for other survey features or respondent characteristics when assessing the relationship between perceived burden and the objective measures of burden because these can affect a respondent's experience of burden. Ignoring these other factors can lead to a misleading interpretation of how certain objective measures of burden are related to perceived burden including making them appear unrelated (Fricker et al. 2014).

We wish to understand how the objective measures of burden, which are obtained as part of the usual CE data collection process, are related to respondents' perceived burden in the CE Interview Survey conditioned on the household characteristics of the respondent. This analysis can help determine which objective burden measures are associated with perceived burden and by how much for different respondents.

One challenge in doing this type of analysis is that in general, surveys do not usually collect measurements of perceived burden (Bradburn 1978). Indeed, in their meta-analysis of studies that examined response rates in relation to a questionnaire's length, Rolstad et al.

(2011) found only three studies that had data that directly asked respondents which questionnaire they preferred and why. Thus, only three of the studies used data that measured respondent burden directly while 25 studies were found that examined the relationship indirectly by means of response rates.

Like most surveys, the CE interview survey does not typically include questions asking respondents directly about the burden they felt filling out the survey. In order to obtain data measuring this more directly, CE engaged in a study with its interview survey between April 2017 and March 2018, in which respondents in the final wave of data collection of the CE interview survey were asked to answer four questions at the end of the questionnaire that were designed to measure the respondent's perception of burden. Respondents were asked to choose answers among "Not at all", "A little", "Somewhat", "Very" or "Extremely" to questions about the amount of burden they felt filling out the survey, how difficult it was to fill out, how sensitive the questions were and if the survey was too long. To get the exact wording of the questions and the possible answers, see U.S. Bureau of Labor Statistics (2017). The relative frequency of responses to the four questions are given in Table 1. For this analysis we use the data from the 6,067 CE interview survey respondents who answered the burden questions at the end of the last wave of their data collection. We can see from these responses that more than half of respondents did not feel the questionnaire was too long or that the questions were very burdensome, difficult or sensitive.

The objective of the analysis demonstrated in this article is to understand the varying effect objective measures of burden have on perceived burden for different household types defined by their demographic and socioeconomic characteristics. To do this analysis, we will need a measure of the perceived burden for each respondent as well as variables that are considered objective measures of burden and variables capturing demographic and socioeconomic information about each respondent and their household.

To quantify the perceived burden felt by a respondent into a single value, we create a composite subjective burden score from the respondents' answers to the four burden questions. First, we assign numeric values to each of the answers of the perceived burden questions, ranging from five for the strongest response, "extremely," to one for the weakest, "not at all," and perform a principal component analysis on these values (Bollen et al. 2001, 2002). However, as Kolenikov and Angeles (2004, 2009) point out, principal component analysis is not well suited for ordinal data where the values are unlikely to

Table 1. The unweighted response rates for each of the response choices to questions on how burdensome, difficult, or sensitive the questions were to answer and how long the survey was in total.

	Relative frequency				
Questions	Not at all	A little	Somewhat	Very	Extremely
Burdensome	34.3%	30.2%	24.1%	7.4%	4.1%
Difficult	44.7%	29.9%	20.3%	3.7%	1.3%
Sensitive	35.3%	26.5%	22.3%	10.4%	5.6%
	Very short	Somewhat short	Neither short nor long	Somewhat long	Very long
Length	4.9%	15.6%	41.8%	28.0%	9.8%

Table 2. Description of explanatory variables. For each numeric variable the mean, median, (standard deviation) and (range) are given. For each categorical variable the percentage of each response is given. All the statistics and percentages given in this table are unweighted.

Variable	Description	Descriptive statistics (unweighted)
NEXP	Number of expenditures reported	31.0, 33.8, (14.2),[0.0, 120.0]
TIME	Interview time (minutes)	63.4, 57.3, (31.7), [6.9, 374.5]
INC	Household income before tax (USD)	62,036.8, 42,000.0, (82,140.8), [-18,572.0, 2865,000.0]
PERL6	Number of people in the household less than 6 years old	0.2, 0.0, (0.5), [0.0, 4.0]
PERL18	Number of people in the household < 18 years old	0.6, 0.0, (1.0), [0.0, 11.0]
PERO64	Number of people in the household ≥ 64 years old	0.4, 0.0, (0.7), [0.0, 3.0]
NCHD	Total number of children	0.7, 0.0, (1.1), [0.0, 11.0]
NUMDK	Number of "Don't Know" responses	0.0, 57.3, (1.6), [0.0, 24.0]
NUMRF	Number of questions not answered	0.1, 0.0, (0.6), [0.0, 16.0]
FAMSIZ	Number of members in the household	2.4, 2.0, (1.5), [1.0, 15.0]
BOOK	If the respondent used the information booklet	Yes 38.0%, No 62.0%
RECS	If the respondent used the financial records	Yes 54.3%, No 45.7%
TENURE	Whether the household owns or rents their home	Homeowner 64.8%, Renter 36.9%
MORT	Whether the household has a mortgage	Yes 36.2%, No 63.8%
MODE	Interview mode	Phone 40.7%, Visit 59.3%
CREF	Respondent initially refused but was persuaded	Yes 13.6%, No 86.4%
DOOR	Door step concerns	No concerns 81.2%, Busy and logistics 8.2%, Privacy and government 7.5%, Other concerns 3.1%
HEDU	Highest education in the household	< high school 6.7%, High school 18.6% Some college 32.3% Bachelors and above 42.4%
FAMT	Family type	Married couple 49.7%, Single father 0.9%, Single mother 4.1%, Other 45.3%
URBAN	Urban or rural area	Rural 18.4%, Urban 81.6% Cannot be Determined 18.4%

follow a normal distribution, leading to estimates that are biased toward zero. For ordinal (or categorical) data where values are assigned in the manner here, Kolenikov and Angeles (2004) found that applying principal component analysis using a polychoric correlation matrix (Pearson and Pearson 1922; Olsson, 1979) rather than the standard correlation matrix corrects for that bias. Using the first component from this principal component analysis, as in Yang (2019), we obtain a single composite measure for each respondent that range in value from 3.4 to 17.0, with a median value of 7.6 and a mean of 8.0.

To help understand what can cause survey participants to feel burden, we model the relationship between this perceived burden measure and several variables collected as part of the survey. Some of these variable are usually associated with burden such as the number of expenditures the respondent had to report (*NEXP*) and interview time in minutes (*TIME*) or *BOOK* and *RECS* which record if the respondent used the information booklet and financial records respectively while answering the questions. These variables can take values of "Yes", or "No". The variable *NEXP* contain values that range from 0 to 120 with a median of 31 and mean of 33.76. The variable *BOOK* had a value of "No" 62% of the time while *RECS* had the value "No" only 46% of the time. These variables that

measure the time and effort required to respond to the survey are often considered objective measures of burden.

Since the effect that objective measures of burden have on the amount of perceived burden can be different for different people in different households, we also include a number of household and demographic variables in the model to help us understand these differences. These variables include household income before tax (*INC*), whether the household owns or rents their home (*TENURE*), whether the household has a mortgage (*MORT*), number of people in the household less than six years old (*PERL6*), less than 18 years old (*PERL18*), and over 64 years old (*PERO64*) and the total number of children (*NCHD*).

Other variables that we include describe different aspects of the data collection process like interview mode (*MODE*), converted refusal indicator (*CREF*), and door step concerns (*DOOR*). The variable *MODE* is the mode of collection used to collect the survey data from the respondent and can be either a personal visit or telephone interview. Since the mode may affect the relationship between the objective measures of burden and perceived burden, we include this variable in the analysis.

The binary variable *CREF* indicates a respondent who initially refused to respond to the survey and has been persuaded by the interviewer. The categorical variable *DOOR* records any concerns that the respondent expressed to the interviewer before taking the survey. The interviewer can code the concerns of the respondent in variable *DOOR* as one of "no concerns", "busy and logistics", "privacy and government concerns", or "other concerns". The category "other concerns" is a catch all for several outcomes from the respondent saying they don't understand the survey to just shutting the door or hanging up the phone on the interviewer. Since the variables *CREF* (McDermott and Tan 2008) and *DOOR* are both associated with a respondents initial attitude toward the survey and thus affect their responses to the burden questions, these variables are also included. The variable *NUMDK* is the number of "Don't Know" responses and *NUMRF* is number of questions to which the respondent refused to answer.

There are several variables about the household collected in the CEQ that could be associated with how busy a respondent is or with their attitudes about government and privacy, so should be accounted for in the analysis. Among these is the variable *HEDU*, which is the highest education level among people in the household. This variable can take one of the values "less than high school", "high school", "some college", or "bachelors and above". The other variables of this type that we include are family type, *FAMT*, which can take one of the values "married couple", "single parent", or "other", *FAMSIZ*, which is the number of members in the household, and the variable *URBAN*, which records if the household is located in an urban or rural area.

3. Regression Tree Model Analysis

In order to understand how different aspects of a survey affect the perceived burden for different groups of respondents, we model the relationship between the perceived burden composite score and various characteristics of the survey, household, and respondent. A model that estimates the value of the perceived burden composite score conditioned on the values to the collected CEQ variables would allow us to see the effect of different measures conditionally on the data collected about the household. In addition, a model that allows us

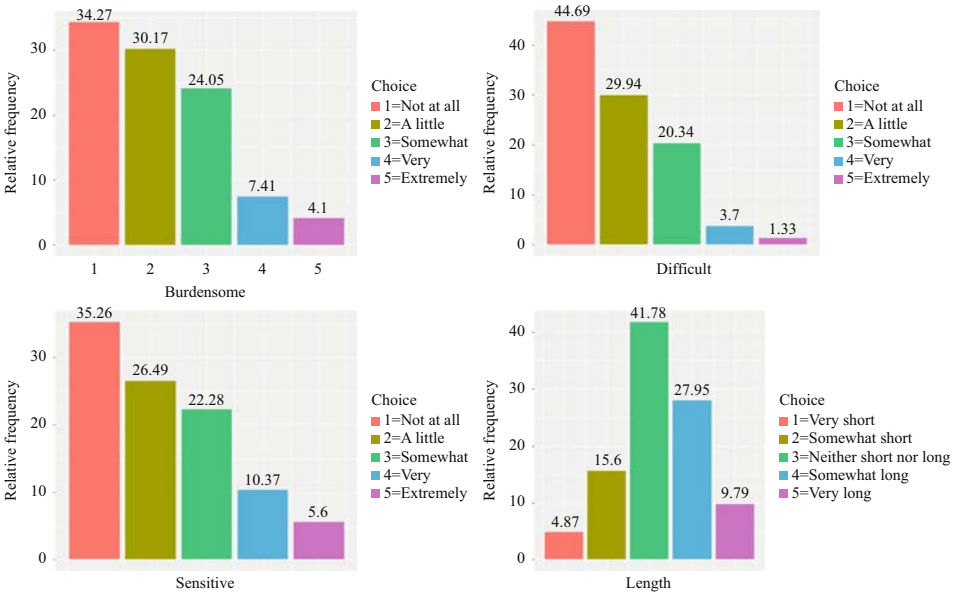


Fig. 3. Bar-charts showing the relative response rates of the choices for each of the burden questions.

to easily understand relationships between the conditional variables could help us better understand what drives the feelings of burden and potentially allow survey administrators to intervene before the next data collection or make changes to collection procedures to head off potential issues. For these reasons we use a recursive partitioning algorithm to create a tree model to do our analysis. These models partition members of the population by splitting them into sub-populations conditionally on their values of the independent variables which can lead to easy interpretation of the model (Toth and Phipps 2014). For examples, Phipps and Toth (2012) and Earp et al. (2018) used these types of models to understand establishment characteristics that affect responses to employment surveys.

The process is termed recursive because each sub-population may in turn be split an indefinite number of times until the splitting process terminates after a particular stopping criterion is reached (Hothorn et al. 2006). One can regard recursive partitioning as producing a model that "predicts" the value of a target variable ("leaf") based on input variables ("branch"). Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the end-nodes of the tree. Typically, within the tree structure, branches represent conjunctions of features that lead to the value of the given end-node.

Since CE data is collected from a sample drawn using a complex design and we would like to generalize our results to the population (Pfeffermann 1996; Pfeffermann and Sverchkov 1999), we use the package **rpms**, *Recursive Partitioning for Modeling Survey Data*, (Toth 2020), in *R* (R Core Team 2020) to estimate the models. This algorithm accounts for the survey design variables and sample weights during the recursive partitioning and parameter estimation to produce a design consistent model. We account for the sample design in our models by including the variables containing the design weights, *FINLWT21*, cluster identifiers, *PSU*, which are the primary sampling units and sample strata, *REGION*,

(Northeast, Midwest, South, West), which are used to stratify the CEQ sample. Since this algorithm uses a design appropriate permutation test to test the statistical significance of each split, it allows us to specify a *p*-value for our analysis. For all models in this article, we specify a *p*-value threshold of 0.05 to test the significance of each split against.

3.1. Conditional Mean Tree Model

To understand the relationship between the value of the perceived burden score (*PB*) and variables that are usually thought to objectively measure burden, we first model *PB* conditionally on the values of those measures. Figure 4 shows the regression tree model of the mean of *PB*, conditioned on several objective burden measures and survey characteristics. The partitioning algorithm selected several different variables for splitting, *TIME*, *NUMEXP*, *CREF*, *BOOK*, *REGS*, *DOOR*, and *MODE*. These variables were identified by the recursive partitioning algorithm to significantly affect the amount of reported burden.

The model identifies whether the respondent expresses concern about their time, government, or privacy as the most influential variable on how much burden a respondent reports feeling. The recursive partitioning algorithm on the variable *DOOR*, where respondents with busy/logistics, privacy/government doorstep concerns reported the highest levels of perceived burden. This indicates that respondents in this group that have these initial opinions (about 15% of the sample), express a higher amount of burden on average than respondents without these concerns. Indeed, they have a *PB* value of 10.09, which is much higher than the 7.6 median and 8.0 mean for respondents overall. The model did not identify any other variable that had a significant effect on the perceived burden reported by this group, which could indicate that there is not much that could be done to change their experience of burden.

The left side of the tree model represents respondents who did not express these concerns. For these respondents, the amount of perceived burden reported was lower overall, but was influenced by the amount of time the survey took to complete; the longer the survey took the higher average reported burden. However, the effect that time has on the amount of perceived burden depends on the mode of the survey.

Respondents that completed the survey through an in-person interview reported a lower overall average amount of burden than respondents that completed the survey over the

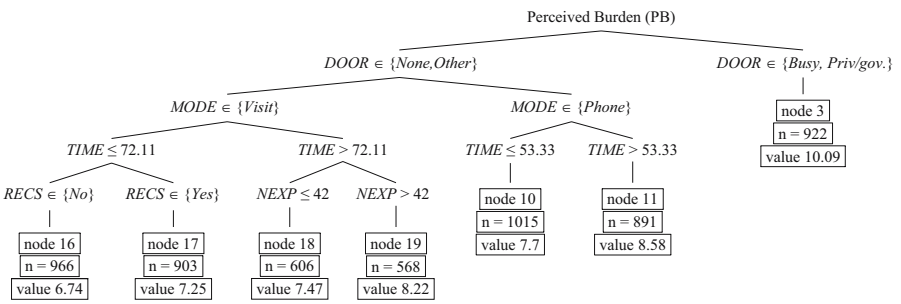


Fig. 4. Regression tree model of the conditional mean of PB: The partitioning algorithm selected the TIME; NUMEXP; CREF; BOOK; RECS; DOOR; and MODE variables for splitting. The mean of PB; conditioned on the splits using these variables, is given in each end node.

phone. For personal interview respondents, the average amount of burden did not significantly increase as long as the time to take the survey was less than 72 minutes, whereas the average reported burden had a significantly higher average for respondents taking the survey by phone after only a little more than 53 minutes. For respondents that had an in-person interview that took longer than 72 minutes, if the number of questions exceeded 42, they reported higher levels of burden than average. While all other respondents that had an in-person interview reported a lower-than-average level of burden, having to refer to records led to a higher reported burden than those who did not. We will see in the regression tree displayed in Figure 7 that the effect of using records on data is more nuanced.

3.2. Conditional Linear Tree Models

Though Figure 4 showed that several objective measures are indeed related to perceived burden, we would like to investigate the effect of each of these objective burden measures for different groups of respondents. Besides modeling the conditional mean of a variable in a tree model, the **rpms** package allows linear models to be fit conditionally on other variables in the tree model (Earp et al. 2018).

We use these models to investigate whether the effects of time, *TIME*, number of expenditures, *NEXP*, and use of records, *RECS*, is different for different groups of respondents. This is done by fitting a linear regression model between *PB* and one of these objective burden measures fit in each end node, while allowing the algorithm to split on the demographic variables when there is an estimated significant difference in the model parameters. By investigating these model parameters conditionally on the demographic information, we hope to understand how the effects of objective measures of burden varies for different types of respondents. Using this type of analysis, we consider how the survey length or the need to consult records or information booklets to answer questions affects perceived burden for different types of respondents.

To analyze the relative effect that time has on the perceived burden we fit the linear model $BURDEN = \beta_1 \times TIME$, while allowing the algorithm to split on any of the variables *MODE*, *INC*, *MORT*, *TENURE*. *PERL6*, *PERL18*, *PERO64*, *NCHD*, *HEDU*, *FAMT*, *FAMSIZ*, or *URBAN* at each step of the algorithm. This model is shown in Figure 5 We fit a linear model with no intercept because we hypothesize that a respondent to a hypothetical interview that took no time ($TIME = 0$ minutes) would report no burden ($PB = 0$). The lowest reported time in the data set is 6.9 minutes, while 75% of all interviews took over 40 minutes to complete.

The resulting regression tree model with this simple linear equation on interview length without intercept confirms what we saw in the previous model, that the effect of time depends on the survey mode (all coefficients are > 0). Though time leads to higher reported level of burden on average, the perceived burden score for a person responding to a personal visit increases by an average of less than one for each 11 minutes of survey length. Meanwhile, the effect of time for respondents answering the survey questions over the phone increases the average perceived burden score at a faster rate, depending on family income. For respondents with a reported family income below USD 25,000, the reported burden score increases by one for every 6.25 minutes they are on the phone, while for respondents with a reported family income greater than USD 25,000, the average

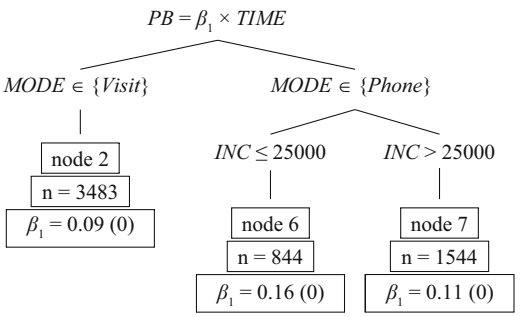


Fig. 5. $PB = \beta_1 \times TIME$: Regression tree model of the conditional relative effect that time has on perceived burden. In each end node the estimated coefficients are given with the (standard error) rounded to three digits.

reported burden score increases by one for each additional nine minutes spent on phone. This difference in rates of reported burden by income is likely due to a difference in the salience of the CE survey for these two groups. Respondents with reported family income below USD 25,000 are unlikely to have as much discretionary income and therefore not many expenditures that they might find interesting to report as compared to families with income greater than USD 25,000.

We also consider this type of tree model analysis for each of the variables *NEXP*, *RECS* and *BOOK* separately. Since each respondent answers demographic questions and questions about the household before the expenditure questions, there is burden associated with responding to the CEQ whether or not the respondent had any expenditures to report, used records, or used the information booklet, so for these variables, we use a linear effect with an intercept term.

Figure 4 shows that, like time, perceived burden increases when more expenditures are reported. Since the number of expenditures and time are correlated, it was natural to see if the effect of *NEXP* is different depending on the mode of survey collection or other demographic variables. To see this, we fit a linear model $PB = \beta_1 \times NEXP + \beta_0$ at each node while allowing the algorithm to split on all the splitting variables used in previous model. Note that, unlike the previous model of time in minutes, we fit a linear model on number of expenditures with a non-zero intercept term, because even if the respondent theoretically reported zero expenditures, they still have to answer questions and so would have burden. The resulting model is shown in Figure 6.

Though the recursive partitioning algorithm found significant differences in average reported perceived burden for different groups of respondents, the effect of *NEXP* was small and about the same for every group. The differences are all in the intercept term. This indicates that the mode of the survey, owning or renting, and the number of children living in the household affects the amount of perceived burden. However, the results show that the perception of burden is not affected by the number of expenditures.

Next, we consider the effect that using records or the information booklet has on perceived burden. Since *RECS* and *BOOK* are indicator variables, fitting linear models $PB = \beta_1 \times RECS + \beta_0$ or $PB = \beta_1 \times BOOK + \beta_0$ at each node while allowing the algorithm to split on any of the variables used in the previous model, leads to an analysis of how the mean-shift effect of these variables changes for different groups of respondents. As in the previous model, Figure 6, we include an intercept term in the linear model which

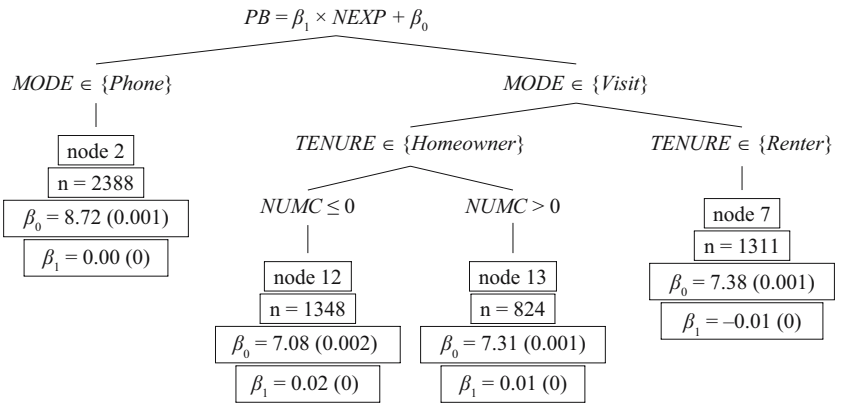


Fig. 6. $PB = \beta_1 \times NEXP + \beta_0$. Regression tree model of the conditional relative effect that the number of reported expenditures has on perceived burden. In each end node the estimated coefficients are given with the (standard error) rounded to three digits.

lets the algorithm differentiate between the effect of the variable being split on and the effect of the variables *RECS* or *BOOK*.

The algorithm did not find any differences in the effect that using the information booklet had on the amount of perceived burden reported. This is not surprising, since the indicator variable *BOOK* does not appear in the tree model relating reported values of perceived burden to objective burden measures (Figure 4). However, the model displayed in Figure 7, analyzing the effect that using records has on perceived burden shows some interesting differences between different groups of respondents.

In the model analyzing the effect of objective burden measures on perceived burden shown in Figure 4, the use of records was associated with a higher reported amount of burden, but only for respondents that responded to the survey through an in-person interview that lasted over an hour (72 minutes). When looking at the difference of this effect by itself among different groups of respondents (Figure 7), we see that though respondents answering the survey questions through a phone interview had higher reported perceived burden, the effect of using records was to lower the amount of perceived burden

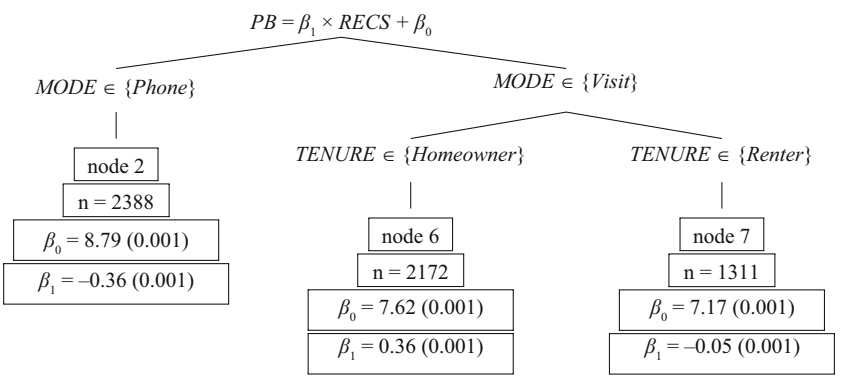


Fig. 7. $PB = \beta_1 \times RECS + \beta_0$. Regression tree model of the conditional relative effect that referring to records to answer survey questions has on the perceived burden. In each end node the estimated coefficients are given with the (standard error) rounded to three digits.

on average. For people responding to the survey through an in-person interview that owned their home, using records raised the reported amount of perceived burden on average. For renters responding to an in-person interview, the effect using records had on the reported perceived burden was negligible. Because consulting records could decrease measurement error in the data, these findings represent trade-offs in some cases between burden and potentially more accurate data. More study on the effect of using records on measurement error in the CEQ should be considered to undergo a complete cost-benefit analysis of asking respondents to consult their records.

The difference in the effect that using records has on the amount of perceived burden between the survey interview modes could potentially be explained by salience. Homeowners often have a larger number of expenditures to report that require records. Locating and using these records to answer questions requires more effort on the part of the respondent, so the positive coefficient is understandable. If the respondent is interviewed in person, it could potentially be more difficult for the respondent to refuse or make an excuse not to get their records. However, respondents interviewed by phone, who usually feel more burdened in general, could more easily say the records are not accessible if they are not interested. Therefore, a respondent's use of records could indicate that they are more interested in answering the survey and so feel less burdened than their counterparts who do not use records in this case.

4. Conclusion

In the above analysis, we demonstrate how regression tree models of the conditional mean can be used to assess the relationship between objective measures and perceived burden. Using respondent's answers to questions directly asking about the burden of CEQ, collected as part of a study by the CE program between April 2017 and March 2018, we were able to model the relationship between the objective measures usually collected as part of the survey and measures of perceived burden. This analysis involves converting the perceived data into a single composite measure obtained from the principal components, which allowed us to use regression tree analysis on the composite measure to see how different survey and respondent characteristics interact with the objective measures of burden to affect perceived burden.

Though there have been mixed findings in the literature on whether these objective measures of burden are related to perceived burden in general, our analysis shows a relationship between perceived burden and some of the objective measures of burden collected in the CEQ survey after conditioning on different demographic and household variables. The tree models also show that the relationship between objective and perceived burden measures are quite affected by the mode of the survey. In general, using a personal interview to collect data seems to ameliorate the effects that most objective measures of burden have on perceived burden.

Using tree models with conditional linear regression estimates at the end-nodes allows us to consider how individual objective measures of burden affect the perception of burden for different groups of respondents, conditioned on the mode of the survey. We modeled three objective measures of burden, interview length (*TIME*), number of expenditures (*NEXP*), and use of records (*RECS*), separately to see the relationship between the values

of these objective measures and the amount of perceived burden. The regression tree model fitting the conditional linear model on *TIME* confirmed a positive association between the number of minutes it takes to complete the survey and the amount of perceived burden a respondent feels, but showed that relative change in perceived burden can be quite different for different groups of respondents. Unfortunately, the finding that door step concerns affect the amount of perceived burden, cautions us that there may be limits to how much changes to a survey or data collection can reduce the burden. This is because door step concerns are indicators of a negative initial attitude that could be very difficult to change.

Meanwhile, the other two tree models using conditional linear predictors, showed that the number of expenditures had almost no effect on perceived burden after conditioning on the survey mode and whether the respondent owned or rented their home, and that record usage can lead to an increase or decrease in the amount of reported burden depending on the mode and whether or not the respondent is a homeowner. The effects of these two variables are likely influenced by whether or not the survey is salient to the respondent. However, the study that collected this data on perceived burden did not directly ask the respondent about the salience of the survey, so we could not test this theory.

Another limitation of this analysis is that the data on burden was collected only from participants who completed the fourth interview. Since burden was not measured for anyone who dropped out before completing the final interview, the findings from this analysis could be misleading if the relationship between objective measures of burden and perceived burden are different for respondents and nonrespondents, even after conditioning on household and survey characteristics.

Despite some unavoidable limitations, the result of the study suggests that interview length, number of expenditures, door step concerns, survey mode, housing tenure and number of children affect perceived burden. Though the relationships between these variables and perceived burden can be complicated, using tree models helped us understand these relationships. By using the package **rpms** which allows us to account for the complex sample design of CEQ, we are able to generalize these results to the full population. This implies that these variables should be included in any model to predict a respondent's perceived burden outcome in future collections and possibly even different surveys.

When constructing a model for prediction, it is not necessary to restrict the model to only the most statistically significant effects as does our tree models nor must we restrict ourselves to models that are easily interpretable. Therefore, in future research, we would like to consider exploring the possibility of using design consistent random-forest models to predict respondent's perceived burden using objective burden measures and characteristics of the survey and respondent.

By constructing a model that can accurately estimate a respondent's anticipated perceived burden, the survey administrator could potentially make changes to the collection mode or survey design early in the data collection process to avoid levels of perceived burden that are likely to lead to nonresponse or possibly be used in an adaptive design. Some of these findings could also be used to guide future changes to the questionnaire or the administration of the survey. Using regression tree models for this analysis is a first step to understanding whether objective measures of burden actually affect perception of burden and to what degree.

5. References

Abayomi, E.J., S. Maliszewski, L. Kreiner, and T. Ballard. 2018. "They spoke, we listened: Reducing respondent burden using previously reported data. In proceedings of the Section on Government Statistics of the 2018 Joint Statistical Meetings, July 28–August 2, Vancouver, British Columbia, Canada. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/conferences/JSM-2018/They_Spoke,_We_Listened-_Reducing_Respondent_Burden_Using_Previously_Reported_Data.pdf. (accessed May 2022).

Ashmead, R., E. Slud, and T. Hughes. 2017. "Adaptive Intervention Methodology for Reduction of Respondent Contact Burden in the American Community Survey." *Journal of Official Statistics* 33(4): 901–919. DOI: https://doi.org/10.1515/jos-2017-0043.

Baird, S., J. Hamory, and E.M. Miguels. 2008. *Tracking, attrition and data quality in the kenyan life panel survey round 1 (klps-1).* Institute of Business and Economic Research Center for International and Development Economics Research University of California, Berkeley Working Paper. Available at: http://emiguel.econ.berkeley.edu/assets/miguel_research/71/attrition_paper_FINAL-CIDER_aug08.pdf (accessed May 2022).

Bogen, K. 1996. "The effect of questionnaire length on response rates: a review of the literature." In Proceedings of the Survey Research Methods Section of the 1996 Joint Statistical Meetings, August 4–8, Chicago, Illinois, USA. U.S. Bureau of the Census. Available at: www.asasrms.org/Proceedings/papers/1996_177.pdf. (accessed May 2022).

Bollen, K.A., J.L. Glanville, and G. Stecklov. 2001. "Socioeconomic status and class in studies of fertility and health in developing countries." *Annual Review of Sociology* 27: 153–185. Available at: http://www.jstor.org/stable/2678618 (accessed May 2022).

Bollen, K.A., J.L. Glanville, and G. Stecklov. 2002. "Economic status proxies in studies of fertility in developing countries: Does the measure matter?" *Population Studies*, 56(1): 81–96. Available at: http://www.jstor.org/stable/3092943 (accessed May 2022).

Bradburn, N.M. 1978. "Respondent burden." In Proceedings of the Survey Research Methods Section of the American Statistical Association: 35: 35–40. American Statistical Association, Alexandria, Virginia, USA. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf. (accessed May 2022).

Cannell, C.F., P.V. Miller, and L. Oksenberg. 1981. "Research on interviewing techniques." *Sociological Methodology* 12: 389–437. DOI: https://doi.org/10.2307/270748.

Cohen, S.B., J.W. Cohen, and K. Davis. 2013 *Longitudinal design options for the medical expenditure panel survey insurance component.* Agency for Healthcare Research and Quality Working Paper. Available at: https://meps.ahrq.gov/data_files/publications/workingpapers/wp_13003.pdf (accessed May 2022).

Connelly, N.A., T.L. Brown, and D.J. Decker. 2003. "Factors affecting response rates to natural resource-focused mail surveys: Empirical evidence of declining rates over time". *Society & Natural Resources* 16(6): 541–549. DOI: https://doi.org/10.1080/08941920309152.

Czajka, J.L., and A. Beyler. 2016. "Background paper declining response rates in federal surveys: Trends and implications." *Mathematica policy research* 1: 1–86. Available at: https://aspe.hhs.gov/sites/default/files/private/pdf/255531/Decliningresponserates.pdf (accessed May 2022).

Earp, M., D. Toth, P. Phipps, and C. Oslund. "Assessing Nonresponse in a Longitudinal Establishment Survey Using Regression Trees." *Journal of Official Statistics* 34(2): 463–481. DOI: https://doi.org/10.2478/jos-2018-0021.

Edgar, J., B. McBride, and A. Safir. 2013a. "Research highlights of the consumer expenditure survey redesign." *Monthly Labor Review* 136:1. Available at: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle = hein.journals/month136&section = 74 (accessed May 2022).

Edgar, J. D.V. Nelson, L. Paszkiewicz, and A. Safir. 2013b. *The gemini project to redesign the consumer expenditure survey: redesign proposal.* CE Gemini Project materials, U.S. Bureau of Labor Statistics. Available at: https://stats.bls.gov/cex/ce_gemini_redesign.pdf (accessed May 2022).

Fekete, C., H. Tough, J. Siegrist, and M.W.G. Brinkhof. 2017. "Health impact of objective burden, subjective burden and positive aspects of caregiving: An observational study among caregivers in switzerland." *BMJ (British Medical Journal)* 7 (9). DOI: http://dx.doi.org/10.1136/bmjopen-2017-017369.

Fricker, S., T. Yan, and S. Tsai. 2014. Response burden: What predicts it and who is burdened out. In Proceedings of the American Association for Public Opinion Research (AAPOR) Annual Conference, May 15–18, 4568–4577, Anaheim, California, USA. Available at: http://www.asasrms.org/Proceedings/y2014/files/400298_500838.pdf (accessed May 2022).

Galesic, M., and M. Bosnjak. 2009. "Effects of questionnaire length on participation and indicators of response quality in a web survey." *Public Opinion Quarterly* 73(2): 349–360. Available at: http://www.jstor.org/stable/25548084 (accessed May 2022).

Groves, R.M. 2006. "Nonresponse rates and nonresponse bias in household surveys." *Public Opinion Quarterly* 70(5): 646–675. Available at: http://www.jstor.org/stable/4124220 (accessed May 2022).

Groves, R.M., D.A. Dillman, J.L. Eltinge and R.J.A. Little. 2001. *Survey Nonresponse*. New York: Wiley Inter-science.

Groves, R.M., E. Singer, and A. Corning. 2000. "Leverage-saliency theory of survey participation: Description and an illustration." *The Public Opinion Quarterly* 64(3): 299–308. Available at: http://www.jstor.org/stable/3078721 (accessed May 2022).

Gustavson, K., T. von Soest, E. Karevold, and E Røysamb. 2012. "Attrition and generalizability in longitudinal studies: Findings from a 15-year population-based study and a monte carlo simulation study." *BMC Public Health* 12(918): 1–11. DOI: https://doi.org/10.1186/1471-2458-12-918.

Hothorn, T., K. Hornik, and A. Zeileis. 2016. "Unbiased recursive partitioning: A conditional inference framework." *Journal of Computational and Graphical Statistics* 15 (3): 651–674. Available at: http://www.jstor.org/stable/27594202 (accessed May 2022).

Kashihara, D., and T.M. Ezzati-Rice. 2004. "Characteristics of survey attrition in the household component of the medical expenditure panel survey (meps)." In Proceedings of the Survey Research Methods Section of the 2004 Joint Statistical Meetings, August 8–12: 3758–3765, Toronto, Ontario, Canada. Available at: http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000706.pdf. (accessed May 2022).

Kolenikov, S., and G. Angeles. 2004. *The use of discrete data in pca: Theory, simulations, and applications to socioeconomic indices*. Chapel Hill: Carolina Population Center,

University of North Carolina: 20: 1–59. Available at: https://www.measureevaluation.org/resources/publications/wp-04-85 (accessed May 2022).

Kolenikov, S., and G. Angeles. 2009. "Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?" *Review of Income and Wealth* 55(1): 128–165. DOI: https://doi.org/10.1111/j.1475-4991.2008.00309.x.

Lynn, P. 2014. "Longer interviews may not affect subsequent survey participation propensity." *Public Opinion Quarterly* 78(2): 500–509. Available at: http://www.jstor.org/stable/24545936 (accessed May 2022).

McDermott, N., and L. Tan. 2008. "The effect of refusal conversion on data quality in the consumer expenditure interview survey." *Consumer Expenditure Survey Anthology*: 23–32. Available at: https://stats.bls.gov/cex/anthology08/csxanth4.pdf (accessed May 2022).

U.S. Bureau of Labor Statistics. 2017. C*onsumer expenditure survey interview questionnaire (ceq).* Technical report, U.S. Bureau of Labor Statistics. Available at: https://www.bls.gov/cex/capi/2 017/2017-CEQ-CAPI-instrument-specifications.pdf (accessed May 2022).

U.S. Bureau of Labor Statistics. 2018. *Handbook of methods.* Technical report, U.S. Bureau of Labor Statistics. Available at: https://www.bls.gov/opub/hom/cex/home.htm (accessed May 2022).

Olsson, U. 1979. "Maximum likelihood estimation of the polychoric correlation coefficient." *Psychometrika* 44(4): 443–460. DOI: http://dx.doi.org/10.1007/BF02296207.

Pearson, K., and E.S. Pearson. 1922. "On polychoric coefficients of correlation." *Biometrika* 14(1/2): 127–156. DOI: https://doi.org/10.2307/2331858.

Pfeffermann, D. 1996. The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research* 5(3): 239–261. DOI: https://doi.org/10.1177/096228029600500303.

Pfeffermann, D., and M. Sverchkov. 1999. "Parametric and semi-parametric estimation of regression models fitted to survey data." *Sankhyā* 61(1): 166–186. Available at: http://www.jstor.org/stable/25053074 (accessed May 2022).

Phipps, P., and D. Toth. 2012. "Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data." *The Annals of Applied Statistics* 6 (2): 772–794. Available at: http://www.jstor.org/stable/41713473 (accessed May 2022).

R Core Team. 2020. *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/. (accessed May 2022).

Rolstad, S., J. Adler, and A Rydén. 2011. "Response burden and questionnaire length: is shorter better? a review and meta-analysis." *Value in Health* 14(8): 1101–1108. DOI: https://doi.org/10.1016Zj.jval.2011.06.003.

Sharp, L.M., and J. Frankel. 1983. "Respondent burden: A test of some common assumptions." *The Public Opinion Quarterly* 47(1): 36–53. Available at: https://www.jstor.org/stable/2748704 (accessed May 2022).

Toth, D., and P. Phipps. 2014. "Regression tree models for analyzing survey response." In Proceedings of the Section on Government Statistics of the 2014 Joint Statistical Meetings: 339–351, Boston, Massachusetts, USA. Citeseer. Available at:

http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.696.8138&rep = rep1&type = pdf (accessed May 2022).

Toth, D. 2020. "*rpms: Recursive Partitioning for Modeling Survey Data*." Available at: https://cran.r-project.org/web/packages/rpms/rpms.pdf. R package version 0.4.0.

Yang, D.K. 2018. "Evaluating perceived burden of household survey respondents." In Presentation of the American Association for Public Opinion Research (AAPOR) Annual Conference, May 16–19, and DC AAPOR and Washington Statistical Society (WSS) Summer Conference July 16, Denver, Colorado, and Washington D.C. USA. Available at: http://app.core-apps.com/aapor_2018/abstract/e28a28d3-f1e7-40e9-b39b-a2cbc76e519, files.dc-aapor.org/slides/summer2018/Yang.pdf. (accessed May 2022).

Yang, D.K. 2019. "Assessing how a household survey is perceived by respondents." In Proceedings of the Section on Government Statistics of the 2019 Joint Statistical Meetings, July 27–August 1. 1–19. Denver, Colorado, USA. Available at: https://www.bls. gov/osmr/research-papers/2019/st190130.htm (accessed May 2022).

Young, A.F., J.R. Powers, and S.L. Bell. 2006. "Attrition in longitudinal studies: Who do you lose." *Australian and New Zealand Journal of Public Health* 30(4): 353–361. DOI: https://doi.org/10.1111/j.1467-842X.2006.tb00849.x.

Received February 2021
Revised November 2021
Accepted May 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1145–1175, http://dx.doi.org/10.2478/JOS-2022-0049

Modeling the Relationship between Proxy Measures of Respondent Burden and Survey Response Rates in a Household Panel Survey

Morgan Earp[1], *Robin Kaplan*[2], *and Daniell Toth*[2]

Respondent burden has important implications for survey outcomes, including response rates and attrition in panel surveys. Despite this, respondent burden remains an understudied topic in the field of survey methodology, with few researchers systematically measuring objective and subjective burden factors in surveys used to produce official statistics. This research was designed to assess the impact of proxy measures of respondent burden, drawing on both objective (survey length and frequency), and subjective (effort, saliency, and sensitivity) burden measures on response rates over time in the Current Population Survey (CPS). Exploratory Factor Analysis confirmed the burden proxy measures were interrelated and formed five distinct factors. Regression tree models further indicated that both objective and subjective proxy burden factors were predictive of future CPS response rates. Additionally, respondent characteristics, including employment and marital status, interacted with these burden factors to further help predict response rates over time. We discuss the implications of these findings, including the importance of measuring both objective and subjective burden factors in production surveys. Our findings support a growing body of research suggesting that subjective burden and individual respondent characteristics should be incorporated into conceptual definitions of respondent burden and have implications for adaptive design.

Key words: Respondent burden measurement; response rates; panel surveys.

1. Introduction

Survey organizations, researchers, and methodologists are concerned with respondent burden, as it can have an impact on response rates to surveys, data quality, attrition in panel surveys, and general survey fatigue (Crawford et al. 2001; Rolstad et al. 2011; Fricker et al. 2014). In an environment where response rates have been on the decline (Brick and Williams 2013), research designed to understand, measure, and help reduce respondent burden is an important, yet understudied topic in the field of survey

[1] U.S. Centers for Disease Control and Prevention, National Center for Health Statistics, Division of Research and Methodology, 3311 Toledo Rd, Hyattsville, MD, 20782, U.S.A. Email: mearp@cdc.gov.
[2] U.S. Bureau of Labor Statistics, Office of Survey Methods Research, Suite 5930. 2 Massachusetts Ave NE, Washington, DC 20212, U.S.A. Emails: Kaplan.Robin@bls.gov and Toth.Daniell@bls.gov
Acknowledgments: We would like to thank our dear friend and colleague Scott Fricker for inspiring, encouraging, and supporting our research as it pertains to respondent burden and survey response in the Current Population Survey. We would also like to thank our colleagues Brandon Kopp, John Dixon, Dorinda Allard, Julie Hatch, and Doug Williams from the U.S. Bureau of Labor Statistics, and Lisa Clement from the U.S. Census Bureau. The work described in this article was performed while Morgan was employed by the U.S. Bureau of Labor Statistics. The findings and conclusions in this article are those of the author and do not necessarily represent the official position of the National Center for Health Statistics, Centers for Disease Control and Prevention.

methodology, as conceptualizations of respondent burden remain underdeveloped and varied.

Bradburn (1978) initially conceptualized respondent burden into four components: interview length, the amount of effort required to answer survey questions, the amount of stress or discomfort experienced during the survey interview, and the frequency of the survey. Despite Bradburn's conceptualization of burden, survey organizations and methodologists still rely on different definitions of burden, namely, whether they emphasize objective features (e.g., survey length and frequency) or incorporate subjective features of burden (e.g., effort and stress). Further, these conceptualizations do not always consider important respondent characteristics, such as motivation and interest (or salience) in the survey topic, which also affect respondent burden and survey response. The sections below summarize these differing conceptualizations of respondent burden, how they have been measured in the literature, and how respondent characteristics moderate the relationship between burden and survey response rates.

1.1. Objective Burden Measures

The U.S. Federal Government and surveys used for official statistics rely primarily on objective measures of respondent burden (also known as "actual burden"). Objective burden is typically defined as an estimate of the time it takes to complete a survey, including survey length, number of questions, and survey frequency. These types of metrics are used to estimate survey burden in the U.S. Federal Government per the Office of Personnel Management (Paperwork Reduction Act Guide 2011).

Objective burden can affect response rates and unit nonresponse, where it is often assumed that longer surveys lead to lower response rates. For example, response rates are higher when respondents anticipate a survey will be relatively short (Crawford et al. 2001; Edwards and Scheetz. 2002; Cunningham et al. 1999). Longer versus shorter survey interviews can lead to higher attrition rates in panel surveys (Galesic 2006), and for respondents to report that they would be less inclined to participate again (Sharp and Frankel 1983). However, the literature on objective burden and response rates has been mixed.

A meta-analysis of studies measuring respondent burden showed only six of 25 studies supported the conclusion that longer surveys lead to lower response rates (Rolstad et al. 2011), and increased objective burden does not necessarily lead to greater survey attrition or poorer data quality (e.g., Sitzia and Wood 1998; Branden et al. 1995). In fact, additional survey waves can actually increase response rates over time, suggesting that objective measures may not always be good indicators of burden (Stanley McCarthy et al. 2006). Objective burden measures are also limited in that they do not consider Bradburn's concepts of effort or stress, nor respondent characteristics that may affect respondents' overall experience of burden (Fricker et al. 2014).

1.2. Subjective Burden Measures

In Bradburn's conceptualization, subjective burden measures (also known as perceived burden) are comprised of effort (e.g., detailed, lengthy, or difficult questions to answer) and stress or sensitivity (e.g., personal questions that cause discomfort). Subjective measures of respondent burden are typically measured via self-reports (e.g., "How

[burdensome/effortful/sensitive] did you find this survey?") (see Fricker et al. 2012, 2014 for examples).

One of the few U.S. government surveys that have asked survey respondents directly about their subjective perceptions of survey burden is the U.S. Consumer Expenditure Survey (U.S. Bureau of Labor Handbook of Methods, Consumer Expenditure and Income 2018; Fricker et al. 2012, 2014). Findings from these studies indicated that objective and subjective burden are not always related, that is, survey length was not associated with how burdensome respondents found the survey (Kaplan and Holzberg 2020). Increased levels of subjective burden have also been linked to survey outcomes such as lower data quality, higher attrition rates in panel surveys, and increased survey fatigue (e.g., Rolstad et al. 2011; Fricker et al. 2014). Thus, subjective burden also has important implications for survey response, where overall respondent burden can be thought of as an interaction between both objective and subjective burden factors (Groves et al. 2000).

More recently, researchers have investigated additional measures that may be related to respondent burden. For example, Yan et al. (2020) developed an innovative model that considers survey features, characteristics of respondents, and subjective measures of burden to determine which contribute to respondent burden and survey response rates. Their model also expands the definition of burden, accounting for subjective burden experienced at three distinct points in time, including burden experienced at the time of the survey request, cumulative burden experienced during the survey interview, and continuous burden experienced at subsequent waves of longitudinal surveys. Importantly, Yan et al. found that respondent characteristics interacted with burden to predict response rates, including respondents' level of motivation, and how difficult, sensitive, or stressful respondents initially found the survey through multiple waves. However, U.S. Federal Government surveys do not routinely collect direct subjective burden measures. When these data are not available, researchers can turn to proxy measures that are available in the existing data (e.g., Kolenikov and Angeles 2009; Montgomery et al. 2000), and known to be related to burden, such as respondent and household-level characteristics.

1.3. Respondent Characteristics Related to Burden, Response, and Data Quality

Respondent characteristics of the individual or household can have important implications for sur-vey outcomes. For example, several studies have measured burden through respondents' interest in the survey topic (e.g., Sharp and Frankel 1983; Galesic 2006). Interest and saliency of the survey topic are positively related to response rates (Groves et al. 2004; Couper 2007). Respondent motivation and level of competence, as well as cognitive ability and higher educational attainment, are associated with greater interest in the survey topic and higher response rates (Haraldsen 2004; Loosveldt et al. 2002). Relatedly, leverage-salience theory suggests that respondents tend to participate in surveys that are personally of interest (or the most salient) to them, and thus the survey topic can be strategically leveraged to increase response by lowering subjective burden (Groves et al. 2000). Thus, individual characteristics of respondents can moderate the relationship between respondent burden and response rates.

Respondent characteristics may also moderate the relationship between respondent burden and data quality. For example, question effort (or difficulty) and stress (or

sensitivity) can increase item nonresponse. Increased question effort can lead to more "don't know" responses, whereas stress can lead to privacy concerns and item refusal (Lee et al. 2004; Bickart et al. 2006; Tourangeau and Yan, 2007; Berinsky 2008). Supporting this, Loosveldt et al. (2002) found that respondents who were not adequately motivated or capable of answering certain questions provided more "don't know" responses and had higher unit nonresponse in future survey waves, whereas those with higher levels of education responded more in future survey waves. Yan and Curtin (2010) also found that once controlling for age, gender, marital status, race, income, and education, respondents who had higher item nonresponse rates had significantly lower re-interview unit response rates. This suggests that not only do respondent characteristics influence subjective burden, but they also predict continuous burden (Yan et al. 2020) and survey response rates over time.

1.4. Overview of Current Article

Taken together, the literature supports the idea of including both objective and subjective features in conceptualizations of respondent burden, as well as taking into account respondent characteristics related to burden and response rate outcomes. Despite this, few measures of respondent burden include all of these dimensions, and survey organizations do not regularly include measures of respondent burden on their surveys.

This article investigates a novel approach to measuring the impact of respondent burden on response rates using proxy measures, when formal measures of burden (e.g., survey length and number of questions administered per respondent) are not made available. Using the Current Population Survey (CPS), we examine the impact of respondent burden, and the moderating role of respondent characteristics, on CPS response rates over time. Specifically, this research investigates multiple factors of Bradburn's model, using proxy factors of both objective burden (e.g., survey frequency and number of months in sample, and number of supplements received) and subjective burden (e.g., effort, stress or sensitivity), along with relevant respondent characteristics at both the respondent and household level measured in the CPS. Because the CPS has eight waves of data collection, it is possible to observe the impact of survey frequency (including supplements) on burden and unit response rates over time.

2. The Current Population Survey (CPS)

The Current Population Survey (CPS; U.S. Bureau of Labor Statistics Handbook of Methods, Current Population Survey 2018) is a monthly longitudinal survey of 60,000 eligible households, jointly sponsored by the U.S. Bureau of Labor Statistics (BLS) and the U.S. Census Bureau, are administered using telephone and in-person interviews. The CPS collects demographic and employment data on all people ages 16 and over residing in the selected household. From the collected data, BLS produces estimates of national unemployment rates, labor force participation rates, full- and part-time employment, and usual weekly earnings. BLS also produces breakdowns by age, sex, race, veteran and disability status, educational attainment, and a variety of other characteristics.

The U.S. Census Bureau collects CPS data in eight waves, where data are collected from a household for four consecutive months, followed by an eight-month break in data

collection. Then data are collected from a household for the same four consecutive months a year later. This allows for a 75% month-to-month overlap and a 50% year-over-year overlap. Census Bureau interviewers ask households to designate one person per household to report about the calendar week (Sunday through Saturday) that typically includes the 12th day of the month. This household respondent reports about all eligible persons in the household and is designated as the current respondent for that month (U.S. Census Bureau 2021). Most months out of the year, the CPS administers additional questions on varying topics – these additional questions are referred to as CPS supplements. Persons within a responding household may qualify to receive a CPS supplement depending on their characteristics. A table of the CPS years, months, and supplement surveys by month in sample discussed in this article is shown in Table 4 in Appendix, Section 6.

2.1. *An Overview of Potential Proxy Indicators of Burden in the CPS*

Few recent studies have directly examined respondent burden in the CPS (e.g., Polivka 1996; Copeland and Rothgeb 1990). DeMaio (1980) found that CPS refusals were highest for collection month one – the introductory month – and month five, when respondents first return after an eight-month break in data collection. Reasons for refusals included invasion of privacy, past experience, and the fact that the CPS is not mandatory.

The CPS public and restricted data do not include objective measures of burden, such as the total length of the survey interview or the number of items each person in the household was eligible to receive. Due to the length and complexity of the CPS it would be nearly impossible to estimate the total number of items each person was eligible to receive, and such estimates would likely be prone to error. While the CPS does not include any direct measures of respondent burden, it does place some unique demands on respondents related to burden that can be used as proxy measures. In terms of objective burden, the CPS is conducted frequently – monthly over eight months. In addition, some subsets of CPS respondents may be more frequently selected for additional supplemental surveys, adding additional length to the survey. Serfling (2004) also found that household size, which can lead to more questions, was a significant predictor of unit nonresponse in future survey waves.

In terms of subjective burden, some respondents may find CPS questions more effortful to answer. The CPS uses proxy response, in which one person responds for all eligible household members. Respondents may have insufficient knowledge to answer for other household members, leading to "don't know" responses (e.g., Cobb 2018). The social distance between household members, or how often they interact, may also cause difficulties in proxy reporting (Bickart et al. 2006). Respondents most familiar with other household members, such as spouses, tend to be better proxies than other relatives or household members (Kojetin and Mullin 1995).

Respondents may also find some CPS questions sensitive to answer, as they might not want to dis-close personal information about themselves or other household members (Sudman et al. 1994). Sensitive questions include those about income, marital status, disability status, education level, or home ownership; depending on these same household characteristics, this can lead to more item-refusals (Fricker 2007; Borlin 2019; Yan and

Tourangeau 2007). Item nonresponse to such sensitive items have also been linked to unit nonresponse in later survey waves in panel surveys (Loosveldt et al. 2002; Kaplan and Holzberg 2020).

On the other hand, some CPS questions or topics may also be more salient to respondents with particular characteristics (Tennant et al. 1991; Mechanic and Newton 1965; Perruccio and Badley 2004). For example, CPS supplements that are personally of interest or most salient may be leveraged to help lower respondent burden (Groves et al. 2000; Zhang et al. 2017). Respondents may experience less burden completing surveys of personal interest to them, regardless of survey length (Marcus et al. 2007; Adua and Sharp 2010). For instance, respondents who own their homes are typically easier to contact, are more likely to be invested in their communities and civic engagement (Groves et al. 1992; Brick and Tourangeau 2017; Abraham et al. 2009; Rohe et al. 2002) and provide more substantive responses (Coder and Scoon-Rogers 1996; Fricker 2007). Because the main topic of the CPS is labor force status – specifically, whether people are employed, unemployed, or not in the labor force – respondents may be differentially motivated to respond to the survey based on their employment status. Employed respondents may be less available but more motivated to respond, but household composition, such as having a spouse present may buffer this because spouses can share the responsibility of responding (Groves 2006). While answering for a spouse may add length, it is often less burdensome than responding for more distant household members (Kojetin and Mullin 1995). Although the literature supports the idea that respondent and household-level characteristics affect survey response, no research has examined the relationship between respondent burden, respondent characteristics, and unit response rates in the CPS.

2.2. The Current Research

The relationship between respondent burden and respondent characteristics remains an important yet understudied topic that could have important implications for survey response and panel attrition over time. Using Bradburn's proxy indicators of subjective burden in the CPS, including effort and stress (or sensitivity) along with objective measures of burden such as survey frequency and length, this research was designed to assess the relationship between respondent burden factors, respondent characteristics, and CPS unit response rates.

3. Identifying Proxy Measures of Burden in the CPS

3.1. Data

For purposes of our research, we used three years of CPS data from 2015 to 2017. Because we were specifically interested in the concept of respondent burden, our analysis began at the time a household first responded to the survey and then followed the same household for the duration of their time in sample.

We restricted our analysis to include all households that were in sample for all eight waves of data collection between 2015 and 2017, starting at the time of their first response to the CPS through wave eight. There were 21 panels of households that were in sample for all eight waves between 2015 and 2017. Of the households that were in sample for all eight

waves between 2015 and 2017, 170,343 re-sponded to the CPS at least once during the eight waves. Only nine of the 170,343 households were flagged as having moved and subsequently replaced by a new household; these households were treated as new households, given that previous respondent characteristics would not apply to replacement households. Of the households was 80% first responded to the CPS during their first month in sample (136,116/170,343). While there is one primary respondent for the household each month, the respondent can change from month to month. For the 170,343 households, there were 245,133 unique household respondents. Our final analysis file includes 170,343 households with 245,133 unique household respondents by month in sample (starting at the first month of household response) totaling 1,115,225 records. The primary respondent may change over the course of eight waves, and while some characteristics (e.g., household characteristics, marital status, and age) will be correlated within a household, others may be unique to each respondent (e.g., gender, education, and employment status).

3.2. Defining Proxy Burden Measures Used

In order to assess the relationship between respondent burden and CPS unit response rates, we defined CPS survey response, identified proxy measures of burden, and identified households and persons sampled or eligible to receive CPS supplements.

3.2.1. Survey Response

Households are in the CPS for eight months total. Each month households agree or refuse to respond to the survey. At the time of the first survey response, each person in the household is designated a person-level identification number ranging from one to sixteen that will carry with them through remaining survey waves, and one of the persons is coded as the household respondent that month (the person designated as the household respondent can vary across waves).

Our study modeled household response using both household and household respondent characteristics. Household response status was determined using the CPS response final outcome code. Full-completes, labor force section completes, and partial completes were coded as respondents. If a respondent completed the labor force questions, but skipped other sections of the CPS, they were coded as a partial complete, and they were treated as a respondent for purposes of this study. We excluded all households that were coded as non-interviews. Non-interviews are households that were temporarily inaccessible due to language barriers or the inability to contact the home, homes that were vacant or were temporarily converted for a non-residential purpose, addresses that no longer existed, were demolished, or were permanently converted for a non-residential purpose. Households that were originally classified as non-interviews, but were later interviewed in subsequent waves, were included at the first month they responded to the CPS.

3.2.2. Factors

The CPS collects data on a variety of household and person-level characteristics that can be used as proxy measures of burden. Using Bradburn's model of respondent burden, we identified several CPS household and person-level items that may correlate with a household respondent's perceived level of effort, stress, as well as survey length and

frequency. Below we describe our hypothesis of how different household- and person-level items are expected to correlate and be used as proxy measures of these burden indicators. The factors described below are not mutually exclusive, nor are they all encompassing. While the authors identify items on the CPS that may be related to each of the following factors of burden, and the authors hypothesize that these proxy indicators will be more highly correlated within the hypothesized factors than across factors, there is no assumption that these items only measure burden or that they measure burden perfectly (for similar approaches, see Fuchs 2005; Van Liere and Dunlap 1980; Below et al. 2012).

3.2.2.1. Effort

As proxy measures of effort, we included household size, number of children, labor force participation, and marital status. Household size is defined as the total number of persons living in the household and can range from one to sixteen. In general, the larger a household is, the more persons the household respondent will be asked to report data for, requiring more effort. Number of children includes the number of children less than 18 years old residing in the household. Depending on the number of children and their ages, the household respondent may be asked to complete additional questions, such as age, race, sex, and ethnicity of children. Number of children may also add to the amount of perceived effort by the household respondent; those with more children may feel like they have less time to complete the CPS. Using the monthly labor force status codes, "employed–at work", "employed–absent", "unemployed–on layoff", "unemployed–looking", "not in the labor force–retired", "not in the labor force–disabled", and "not in the labor force–other", initial regression tree analyses indicated that unit response rates were more similar for those with an monthly labor force status of "employed–at work" compared to other classifications; therefore, we classified respondents as either being employed at work as "At Work" or not "Not at Work" (including all other labor force participation categories). Marital status was self-reported and categorized as married with spouse present, married with spouse absent, widowed, divorced, separated, and never married. Marital status may or may not include cohabitating partners, since the designation is made by the respondent. Marital status with a spouse present was hypothesized to reduce the amount of effort needed to complete the CPS (Groves 2006; Kojetin and Mullin 1995), in comparison to the other marital status levels; therefore, we classified household respondents as either being married with spouse present "Spouse Present" and all other marital status response options as "No Spouse Present". The expectation being that respondents that have a spouse present may be less burdened with household tasks in general, and potentially have more time to complete the survey. Overall, we expected that proxy measures of effort would be correlated, and that increased indicators of effort would be associated with decreased unit response rates over time.

3.2.2.2. Length

The second factor Bradburn describes is length of the survey. While the CPS does not provide a variable measuring the total time spent to complete the survey, there are other aspects of the survey that Bradburn also classifies under the dimension of length: number of months in sample, number of months in data collection, number of waves in data collection, and survey mode. While Bradburn hypothesizes that the perception of survey

length varies by survey mode, we did not include survey mode, since all telephone refusals were reassigned to in-person interviews, and thus all nonrespondents have a mode designation of in-person. Number of months in sample ranged from one to eight, one being the household's first month in the sample, and eight being the household's last month in sample. Not all households responded their first month in sample; therefore, we also measured the total number of months in data collection. Data collection was defined as the total number of months the household had been participating, starting at the household respondent's first month of response. There was an eight-month break in data collection between the first four months in sample and the last four months in sample. We defined all months in data collection prior to the eight-month break as Phase I and all data collection after the eight-month break in data collection as Phase II. If a household did not respond until after the eight-month break in data collection, then they were coded as Phase I, since they did not experience the eight-month break in data collection. We also measured the number of months in data collection within each phase, since month one in sample is thought to have similar characteristics to month five in sample, meaning that both of these initial months tend to have lower unit response rates and ask more key demographic questions. Overall, we expected that proxy measures of survey length would be correlated and that increased survey length would be associated with decreased unit response rates, and that this would be moderated by respondent characteristics. For instance, increased CPS survey salience and availability to respond would be associated with increased unit response rates relative to those with respondent characteristics associated with lower CPS saliency.

3.2.2.3. Stress

The third factor Bradburn (Bradburn 1978) describes is stress and sensitivity, which relate to the re-spondent's perception of how personal they find the survey questions, or the amount of discomfort experienced while responding. As potential proxy measures of stress and sensitivity, we included several item missing/not reported flags. While there are over a hundred item missing flags included in the CPS public file, many items are missing less than one percent of the time. Our model included item missing flags for items thought to be potentially stressful or sensitive that had missing rates of at least one percent. We included item missing flags for education, age, marital status, race, ethnicity, previous or current military service, and disability (a person was considered to have a disability if they reported that they have serious difficulty hearing; sleeping; seeing; concentrating, remembering, or making decisions; walking or climbing stairs; dressing or bathing; or doing errands alone because of a physical, mental, or emotional condition), based on the literature that these items can be sensitive to respondents and are associated with greater item nonresponse (Fricker 2007; Borlin 2019; Kaplan and Holzberg 2020). We hypothesize that item nonresponse across items would be correlated and that increased item nonresponse to sensitive questions would be related to unit nonresponse in later survey waves.

3.2.2.4. Frequency

Lastly, Bradburn (Bradburn 1978) describes the potential moderating effect of survey frequency on survey saliency, hypothesizing that certain subpopulations are sampled/eligible for surveys more frequently than others and that this potentially lessens the positive effects of saliency over time. This is different than being contacted multiple times

during a longitudinal survey, since typically everyone in the sample is surveyed at the same intervals of time. While every household is contacted over all eight waves of data collection in the CPS, only certain households are asked to complete an additional supplement survey each month, and depending on the characteristics of the household, some households may be sampled/eligible for more supplemental surveys than others, which could potentially negate the survey saliency and adversely affect their propensity to respond to future waves of the CPS. For example, in the CPS, a female veteran who is in school would potentially be eligible for at least three supplements (fertility, veterans, and school enrollment), not to mention the supplements that everyone is eligible to receive such as civic engagement and computer and internet usage. While all of these topics may be salient, which could increase their interest and motivation, being sampled for an additional supplemental survey every month could also reduce the impact of saliency on respondent burden and thus reduce survey response rates. In order to assess the potential effect of survey sampling frequency, we examined the effect on CPS unit response rates of the number of additional supplemental surveys a household was previously sampled/eligible for, as well as the number of supplements the household respondent was previously sampled/eligible for. Overall, we expect survey frequency to correlate with proxy measures of survey length and to be associated with decreased response over time, especially for those with respondent characteristics indicative of lower CPS saliency.

In order to estimate frequency, we used supplement sampling flags when available and otherwise we used supplement selection criteria as specified in the corresponding supplement documentation for that given month. We created two supplement frequency sum variables: (1) the total number of supplemental surveys a household was previously sampled/eligible for using the person-level data; and (2) the total number of supplements for which the household respondent was previously sampled/eligible. We used both the household and the household respondent sums as proxy measures of survey frequency. The total number of supplements that a household or household respondent could previously be sampled/eligible for prior to month eight in sample is seven supplements. Due to the fact that a household needs to respond to the CPS in order to be administered a supplement that month, we also controlled for previous month response, by including an indicator of previous month response status, when assessing the effect of previous supplement sampling/eligibility.

We examined the effect of being previously sampled/eligible for the following 13 supplemental surveys administered between 2015 and 2017: Tobacco Use, Unbanked/Underbanked, Computer and Internet Use, Veterans, Volunteer, School Enrollment, Food Security, Displaced Workers/Job Tenure, Child Support, Fertility, Voting, Contingent Workers, and Public Participation in the Arts. Since all households in sample in March are given the Annual Social and Economic Supplement (also known as the March supplement), regardless of their characteristics, we did not examine the impact of having been previously sampled for the Annual social and Economic Supplement in February, March, or April.

3.2.3. Factor Structure

In order to examine relationships among the different burden proxy measures proposed in Subsubsection 2.1.2, we used an exploratory factor analysis (EFA). EFA is used to assess

the relationship among variables and to identify subsets of correlated variables that potentially measure distinct constructs or varying dimensions of constructs referred to as factors (Bates 2017; Brust et al. 2016; Couper et al. 2008). Unlike Principle Components Analysis that is primarily focused on reducing the number of variables, EFA is used to explore the factor structure of an underlying construct. Exploratory factor analysis is typically used to explore the relationship between continuous variables, since it examines the Pearson correlation among each pair of variables. While the Pearson correlation coefficient is typically used to assess the correlation between two continuous variables, it is also equivalent to the Point-Biserial correlation when used to calculate the correlation between a continuous variable and a binary variable, and to the Phi coefficient when used to calculate the correlation between two binary variables, and therefore is used to explore and make sense of the relationships between a mixture of continuous and binary proxy burden indicators described above.

Using exploratory factor analysis, items that are correlated will have higher loadings under the same factor, where items that are not correlated will have higher loadings under other factors. In order to determine factor loadings, items are plotted in a multidimensional space with a separate axis for each factor. The axes of the factors are typically rotated in one of two ways, orthogonal versus oblique, in order to determine where items fall in that multidimensional space. Orthogonal rotation assumes that the factors are not related, whereas oblique rotation assumes that the factors are correlated. For purposes of our research, we use oblique rotation, assuming that the different factors of respondent burden are correlated. The most common orthogonal rotation is varimax rotation. The varimax rotation is shown in Table 1. Promax is the most common oblique rotation and is based on the varimax rotation and is shown in Table A2. Promax raises the varimax factor loadings to the power of two, three or four, which drives down smaller factor loadings without having much of an impact on larger loadings, and can be used to simplify the results, especially when factors are expected to be correlated (Finch 2006). Given that we are interested in modeling respondent burden conditional on being sampled for the CPS, we do not use the design weights in either EFA model.

First, we assessed the appropriateness of the four-factor solution proposed by Bradburn (1978) using Kaiser's (1960) eigenvalue-greater-than-one rule as proposed by Cattell (1966). An eigenvalue represents the amount of variance that a given factor explains in the variables; the lower the eigenvalue, the less variation in the variables the factor explains (Norris and Lecavalier 2010). Using the eigenvalue-greater-than-one rule we settled on a four-factor solution. After the four-factor solution was assessed for appropriateness using the eigenvalues, we examined the factor loadings for each variable. Factor loadings are used to assess how representative a variable is for a given factor (Howard 2016). We retained all items with a factor loading of 0.32 or greater as proposed by Costello and Osborne (2005), Hinkin (1995, 1998), and Tabachnick and Fidell (2001). All items had absolute factor loading values of greater than 0.32, using both the orthogonal and the oblique rotation (see Table 1 for the orthogonal rotation factor loadings and Table 5 in the Appendix, Section 6, for the oblique factor loadings). After identifying the optimal number of factors, we then assessed the item factor loadings. The eigenvalues and factor loadings support the four-factor solution shown in Table 1.

Table 1. Exploratory factor analysis factor loadings of CPS proxy burden measures and factor eigenvalues (λ) using orthogonal varimax rotation.

CPS variable	Effort $\lambda = 1.91$	Length $\lambda = 3.41$	DIS stress $\lambda = 5.65$	DEM stress $\lambda = 1.80$
Household size	0.854	0.003	0.011	-0.036
Number of children	0.732	0.023	0.002	-0.026
Marital status	0.346	0.028	0.016	-0.046
Labor force participation	0.332	-0.010	-0.033	0.007
Disability status	0.330	-0.002	-0.015	0.019
Age	-0.502	0.068	0.029	-0.019
Month in data collection	-0.091	0.972	-0.003	-0.024
Month in sample	-0.015	0.805	0.001	0.016
Data collection phase	-0.049	0.761	0.001	-0.014
Number of previous supplements the household respondent was sam-pled/eligible for	-0.002	0.742	-0.009	-0.030
Number of previous supplements the household was sampled/ eligible for data collection	0.191	0.673	-0.004	-0.037
month w/in phase	-0.049	0.480	-0.004	-0.015
NR – Difficulty dressing or bathing	-0.026	-0.004	0.969	0.190
NR – Difficulty walking or climbing stairs	-0.025	-0.003	0.967	0.189
NR – Difficulty doing errands alone	-0.026	-0.004	0.966	0.190
NR – Blind or difficulty seeing	-0.022	-0.003	0.964	0.191
NR – Difficulty concentrating, remembering, or making decisions	-0.024	-0.004	0.963	0.190
NR – Deaf or difficulty hearing	-0.019	-0.004	0.942	0.190
NR – Ethnicity	0.006	-0.006	0.094	0.642
NR – Race	0.007	-0.005	0.070	0.566
NR – Education Level	-0.020	-0.023	0.227	0.481
NR – Marital Status	-0.051	-0.015	0.096	0.477
NR – Veteran status (ever served in the military)	-0.016	-0.005	0.062	0.381
NR – Age	0.005	-0.024	0.107	0.355
NR – Current military status (currently serving in the military)	0.012	-0.003	0.026	0.339

Overall, the four-factor solution provided evidence for an effort factor, a length factor, and two stress factors. According to the eigenvalues (λ), the stress factor for disability items explained the most variation ($\lambda_{DIS-Stress} = 5.65$) followed by length ($\lambda_{Length} = 3.41$), stress for demographic items ($\lambda_{DEM-Stress} = 1.80$), and effort ($\lambda_{Effort} = 1.91$) (see Table 1). Household size, number of children, being "employed–at work", having a disability, and having a spouse present all had positive correlations with effort, where age had a negative correlation with effort. While we actually expected having a spouse present to decrease the amount of perceived effort, the factor loading is likely positive due to the positive correlation with household size ($\rho = 0.42$, $p < .0001$,

Table 2. Exploratory factor analysis inter-factor correlations (w/ promax rotation).

	Effort	Length	DIS stress	DEM stress
Effort	1.00	-0.02	-0.03	-0.05
Length	-0.02	1.00	-0.01	-0.05
DIS stress	-0.03	-0.01	1.00	0.35
DEM stress	-0.05	-0.05	0.35	1.00

$n = 1{,}115{,}225$). Month in sample, month in data collection (starting at the first month of response), data collection phase, number of previous supplements that the household and the household respondent were sampled/eligible for, and data collection month within phase, were all positively correlated with length. All disability and demographic item missingness flags were positively correlated with the two stress factors.

According to the inter-factor correlations shown in Table 2, both of the stress factors are positively correlated ($\rho = 0.35$), indicating that item missingness for disability questions is related to item missing-ness for demographic questions. The inter-factor correlations provided in the EFA model are not tested for significance and are merely meant to be informative. EFA also does not consider moderating and interaction effects between variables within the same factor and across factors, therefore, the remainder of our article uses regression tree modeling. While we could have used a structural equation model (SEM), regression trees are better suited for handling missing data, moderating and interaction effects, as well as nonlinear effects (Earp et al. 2014, 2018; Morgan and Sonquist 1963; Phipps and Toth 2012; Toth and Phipps 2014). We also could have created factor score variables to be included as inputs in a model of survey nonresponse; however, this would fail to provide us with insight as to how specific household and respondent characteristics interacted and related to survey nonresponse. The EFA model was used to assess construct validity, and ultimately provided evidence that proxy measures of effort are correlated with other proxy measures of effort, proxy measures of length are correlated with other proxy measures of length, and proxy measures of stress are correlated with other proxy measures of stress.

4. Measuring the Impact of Proxy Measures of Burden on CPS Unit Response Rates

Using the burden proxy measures described above and shown in Table 1, we assessed the effect of respondent burden on CPS unit response rates, as well as the linear effect of CPS supplement frequency on CPS unit response rates controlling for other proxy measures of burden and respondent characteristics available in the CPS thought to be indicative of survey saliency. Though we do not have direct measures of saliency, we expected that the saliency of the CPS would be affected by a number of household- and respondent-level characteristics. In addition to the characteristics described above in relation to effort (e.g., marital status, labor force participation, household size, and so on.), additional characteristics thought to impact CPS saliency included household income, home ownership status, race, ethnicity, education level, veteran status, and disability status. If the house was reported as being owned, then home ownership status was set to "own home", but if the house was reported as being rented either for cash or for no payment, then it was classified as "not own home". If a person reported ever having served in the military, they

were classified as being as veteran. Lastly, if a person reported that they had serious difficulty hearing; sleeping; seeing; concentrating, remembering, or making decisions; walking or climbing stairs; dressing or bathing; or doing errands alone because of a physical, mental, or emotional condition disability, then they were coded as having a disability.

Traditionally survey response is modeled using logistic regression models; however, over the last decade there has been an increase in the use of regression tree models to model survey response (Earp et al. 2014, 2018; Toth and Phipps 2014). Regression trees can be advantageous for modeling survey response for a number of reasons: (1) they can automatically identify associated variables as well as their optimal break points or categorical groupings; (2) they can automatically detect statistically significant interaction effects; (3) they can use missing data as a valid input into the model; (4) they are nonparametric and can model nonlinear effects; and (5) the models can be used to easily identify and describe mutually exclusive subgroups with varying levels of unit response rates.

While predictions from a regression tree are not usually as accurate or stable as those from a ran-dom forest or a boosted model, the overall structure of the regression tree and the relative values of the estimate parameters are what is important for inference about the population, and these tend to be stable from sample to sample. Tree models produced using the R Package "Recursive Partitioning for Modeling Survey Data (rpms)" are especially stable since they are based on a permutation test to determine the variables on which to split as well as a when to stop splitting (Toth 2019). In addition, since these other types of models are not easily interpreted, being intended for prediction as opposed to inference, they are not appropriate for this application.

For this article, the regression tree R Package "Recursive Partitioning for Modeling Survey Data (rpms) (Toth 2019) was used to assess the impact of proxy indicators of survey burden on CPS unit response rates. Using the rpms package, we were able to model the impact of proxy burden indicators and CPS supplement frequency on CPS unit response rates. We built two regression tree models using rpms: (1) modeling the effect of proxy burden measures on overall CPS response; and (2) modeling the linear effect of the number of supplements a household respondent was previously sampled/eligible for on CPS response, controlling for the effect of proxy burden measures. The rpms algorithm uses the permutation test that accounts for clustering and unequal probability of selection introduced in Toth (2020), to determine if each split is significant at the level specified by the user. Once the algorithm fails to find a significant split, the recursive splitting is stopped. This stopping rule leads to an unbiased variable selection as opposed to many algorithms which rely on an over-fitting and pruning procedure to build regression trees. The subgroups resulting from each split are referred to as nodes, and the final mutually exclusive groups shown at the bottom a tree a referred to as end nodes. The entire data set before splitting takes place is referred to as the root note, as is labeled node 1 (n). The first left child node of the root node is labeled as node 2 (2n) and the right child is given the node number 3 (2n+1). This leads to every node getting a unique node label.

Two regression tree models are described below. The first regression tree model assesses the effect of proxy measures of burden (including all variables listed in Table 1) on CPS unit response rates, and the second regression tree model assesses the linear effect of frequency of CPS supplements on CPS unit response, controlling for other proxy measures of burden (including all other variables shown in Table 1).

While the rpms package treats missing values for categorical variables as valid values, it does exclude missing values for continuous variables. The results of both regression tree models are based on 976,417 out of 1,115,225 observations that did not have missing values for continuous variables.

4.1. The Effect of Proxy Measures of Burden on CPS Unit Response Rates

See Table 3 for an overview of the components of burden and relevant respondent characteristics, CPS proxy burden measure, its hypothesized impact on unit response rates, and the final results from the regression tree analysis.

Our model of the effect of proxy measures of burden on CPS response rates indicates that proxy measures of subjective burden, including effort, stress/sensitivity, and respondent characteristics, are related to CPS survey response. First, the model indicated that respondents that report owning a home (related to respondent characteristics) (see Figure 1, Nodes 6 and 7) have a higher average response rate ($\bar{z}$) than respondents that do not own their home (rent their home) or do not report whether they own or rent their home (see Figure 1, Nodes 9, 10, 11, 16 and 17). Second, the model indicates that persons that do not report home

Table 3. Overview of respondent burden factors, CPS proxy measures, hypothesized impact on response rates, and final results from regression trees.

Component of respondent burden	Hypothesized CPS related proxy measures	Hypothesized impact on response rates	Regression tree results
Objective burden (Survey frequency and length)	• Survey characteristics known to add time spent on the CPS: • 8+ CPS survey waves • Number of months in sample • Number of months in data collection • Number of waves of data collection (including CPS supplements) • Number of supplements the household was previously sampled/eligible for • Number of supplements the household respondent was previously sampled/eligible for	• Increased survey length and frequency will be associated with decreased unit response rates over time, especially for those who find the survey topic least salient	• Number of additional CPS supplements a household respondent was sampled/eligible for was associated with decreasing response rates for homeowners, employed at work, who refused to report education level.

Table 3. Continued

Component of respondent burden	Hypothesized CPS related proxy measures	Hypothesized impact on response rates	Regression tree results
Subjective burden (Effort)	• Factors known to add effort to the CPS interview due to responding for additional persons in the household or make responding more or less effortful due to availability (i.e., household size, number of children, labor force participation, and marital status)	• Increased indicators of effort will be associated with decreased unit response rates over time. • Those who are currently employed may be busier and less available to respond • Those with a spouse present have more people in the household available to respond	• Labor force participation (being employed at work) and being sampled for additional CPS supplements was associated with decreasing response rates over time • Having a spouse present in the household was associated with higher average response rates over time for those who were currently employed and at work
Subjective burden (Stress/sensitivity)	• Response to questions known to be stressful/sensitive (i.e., income, marital status, education, age, race, ethnicity, military service, disability status)	• Increased item nonresponse to sensitive questions will be associated with unit nonresponse in later survey waves	• Nonresponse to ques-tions on home ownership and education level were associated with lower average response rates
Respondent characteristics (CPS saliency)	• Respondent or household-level characteristics that impact CPS salience, e.g., • Home ownership • Marital status • Labor force participation • Household size • Race • Ethnicity • Education level • Veteran status • Disability status	• Respondent characteristics related to CPS saliency may impact response and moderate the relationship between respondent burden and response rates (e.g., those who own their homes are more interested in the survey topic, and will have increased response rates).	• Home ownership was associated with higher average response rates, except for respondents that refused to report education level; home-owners that refused to report education had lower average response rates over time • Respondents that did not report home ownership had lower average response rates over time.

ownership or education level have lower response rates (related to sensitivity). Overall, we see that not reporting whether a home is owned or rented and not reporting education level are negatively correlated with CPS response rates ($\rho_{NR\text{ - }Home\ Ownership} = -0.07$ and $\rho_{Education} = -0.06,\ p < 0.0001$). Refusing to provide education level is not only correlated with unit nonresponse, but it is also correlated with not reporting other items as indicated by the results of the EFA described above. Third, the model indicates that persons with a spouse who is present have higher response rates than those that do not have a spouse, or their spouse is absent, for those that previously reported being "employed –at work" and did not report whether they owned or rented their home (see Figure 1).

The regression tree also highlights an interesting interaction effect between sensitivity and respondent characteristics. The model indicated a potential interaction effect between marital status and labor force participation within households not reporting whether they own or rent their home. While households where the respondent previously reported being "employed – at work" tended to have lower response rates on average, we found that this effect varied based on marital status. Households where the respondent previously reported being "employed – at work", but had a spouse present (see Figure 1, Node 17, $\bar{r} = 0.896$), had closer response rates to those who previously reported not being "employed – at work" regardless of marital status (see Figure 1, Node 9, $\bar{r} = 0.884$).

4.2. *The Linear Effect of Frequency on CPS Response, Controlling for Other Proxy Measures of Burden*

The second model assesses the effect of objective burden (survey frequency) on response rates, controlling for proxy measures of burden. Here we model the linear effect of the number of supple-ments that a household respondent was previously sampled/eligible for (controlling for all the variables shown in Table 1), in order to examine if being sampled for additional CPS supplements exacerbates or lessens the effect of other proxy burden measures on survey response rates. Where regression trees typically distinguish between groups with different means (e.g., average response rates), the linear regression tree model used here instead distinguishes between groups with different intercepts and slopes. The second model identifies groups that were affected differently by number of supplements. The intercept (β_0), represents the average response rates for household respondents who have never been sampled/eligible for a supplement previously, and the slope (β_1), is the effect on response rates for every additional supplement that the household respondent was sampled/eligible for.

The second model indicates that while average response rates did not vary as a function of labor force participation for households that reported owning their home according to the previous model (see Figure 1), labor force participation is a determinant for home owners that do not report an education level when we consider the number of supplements that the household respondent was previously sampled/eligible for (see Figure 2). The response rates for home owners "employed – at work" decreased by 2.5 percentage points for every additional supplement that the household respondent was previously sampled/eligible for (see Figure 2, Node 10), where the response rate for home owners not "employed – at work" decreased by 1.6 percentage points for every additional supplement that the household respondent was previously sampled/eligible for (see Figure 2, Node 11). Number of supplements did not appear to affect response rates for households that rented their home or

Fig. 1. Regression tree model of the effects of proxy measures of burden on CPS response rates.

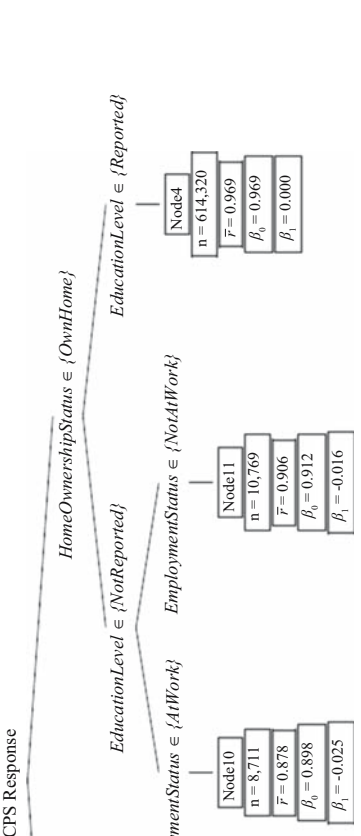


Fig. 2. Linear regression tree model assessing the linear effect of supplemental survey frequency on CPS response, controlling for other proxy measures of burden.

did not report whether they owned or rented their home, as well as for households that reported owning their home, but did report their education level.

Taken together, these findings suggest subjective burden as it relates to effort and stress or sensitivity, and objective burden as it relates to being eligible for additional supplemental surveys, are related to CPS response rates. Further, several interactions between objective and subjective burden factors, including respondent characteristics, sensitivity, and survey frequency provided predictive value in identifying who responded to additional CPS waves and supplements over time.

5. Discussion

Respondent burden has important implications for response rates, panel attrition over time, and data quality, yet remains an understudied topic in the survey methods literature (Crawford et al. 2001; Rolstad et al. 2011). Objective measures of burden, including survey length and frequency have received the most attention and are often used as the primary metric when defining respondent burden in surveys used for official statistics in the U.S. Federal Government (Paperwork Reduction Act Guide 2011). However, relying on objective measures of burden alone does not consider respondents' subjective perceptions of burden and individual respondent characteristics that can have an impact on survey outcomes (Fricker et al. 2014; Yan et al. 2020). Further, studying subjective burden is challenging since U.S. Federal Government surveys do not routinely include measures of subjective burden. Thus, this research used a novel approach to investigate the impact of objective and subjective proxy burden measures based on Bradburn's (1978) framework, and respondent characteristics available in the Current Population Survey, on response rates over time.

Overall, we found that the proxy burden measures, including effort, stress, and respondent charac-teristics correlated as we would have expected given Bradburn's model of respondent burden. House-hold size, marital status, and labor force participation all correlated with items intended to be proxy measures of effort. Item nonresponse and stress or sensitivity tended to be correlated, while household and respondent characteristics such as income, race, and education level intended to measure saliency were often correlated with other proxy measures of saliency.

Interestingly, several proxy measures of burden, as well as interactions between objective and subjective burden factors, were predictive of nonresponse in future survey waves. We found that one of the biggest determinants of future nonresponse was item nonresponse to questions related to household ownership and education level, indicative of survey stress or sensitivity. Respondents that refused to report household ownership or education level also tended to have lower response rates in future waves of data collection, indicating that subjective burden factors such as effort and stress or sensitivity are predictive of CPS response rates. We also saw that respondent characteristics related to effort and respondent characteristics, including labor force participation and marital status, predicted response rates in future waves. This suggests proxy measures of effort, such as being busy at work, reduce the probability of future response, whereas having a spouse present increases the probability of future response.

Lastly, we found that the frequency at which household respondents are sampled or eligible for additional CPS supplemental surveys can also impact their response rates in future waves of

data collection. While owning a home may be correlated with civic engagement and thus increase the overall saliency of the CPS, home owners who are not motivated to report an education level either because the question is considered difficult or sensitive, are also adversely impacted by the frequency of CPS supplements, especially for those who are "employed–at work". Taken together, these findings provide evidence that an interaction between objective burden factors (survey frequency) and subjective burden factors (effort, and sensitivity) related to respondent characteristics impacted propensity to respond to the CPS over time.

5.1. Implications

Our findings support a growing body of literature arguing that respondent burden should be con-ceived of as the interaction between objective and subjective survey factors, and that definitions of respondent burden that only consider objective features are limited. This research suggests that not only do objective features of surveys like frequency contribute to important survey outcomes such as response rates and attrition over time, but subjective factors, including effort, sensitivity, and respondent characteristics are predictive of survey participation over time. We believe these findings are of considerable importance to survey organizations and researchers who want to understand burden when direct measures are not available. Using the approach demonstrated in this article, researchers can use existing data by examining proxy factors and respondent characteristics related to burden to predict response rates.

The findings from this research could be applied in a number of settings. For example, these results could be used in adaptive or responsive survey designs to target those likely to experience greater subjective burden in the CPS via operational interventions (e.g., switching modes or increasing level of effort for those likely to experience the most burden; Brick and Tourangeau 2017) to help mitigate the impact of declining response rates in the future. Researchers could also improve messaging in survey advance letters and within survey instruments to increase saliency.

Because the proxy burden measures used in this study were comprised of general household and respondent level characteristics that are also collected on a variety of other household surveys, these findings likely have implications for other household surveys. Our study provides evidence of interaction effects between proxy measures of burden and respondent characteristics that may be available in existing household survey data where direct measures of respondent burden are not. While the specific results outlined in this research apply to the CPS, other U.S. Federal Government surveys collect similar data and other researchers could apply similar techniques using Exploratory Factor Analysis and regression trees to help predict response rates. This novel approach to measuring and studying respondent burden may be applicable to other surveys to investigate survey response rates in household surveys more generally, but future research is needed to assess whether these results are generalizable to other surveys.

5.2. Limitations and Future Directions

One limitation of this research was that we could not directly measure respondent burden, instead relying on proxy measures of burden. For instance, while we were able to include several proxy measures of length as specified by Bradburn, the CPS does not provide an

actual measure of total survey time, nor does it include any subjective burden measures by asking respondents to report on burden. However, a plus side of that limitation was developing a novel approach, relying on existing measures in the CPS as a proxy for burden to study their impact on response rates.

Although these proxy measures reflected the burden constructs of interest, future research should directly measure respondent burden. This is seldom done, but one example is the Bureau of Labor Statistics' Consumer Expenditure Survey (Fricker et al. 2014), which reports measures of objective burden (length of survey interview in minutes) and includes subjective burden measures of effort, stress, saliency, and length, and could be used to ascertain (a) if the proxy measures proposed in this article are correlated with perceived levels of these dimensions in the Consumer Expenditure Survey; and (b) if the subjective and objective burden measures in the Consumer Expenditure Survey correlate with characteristics thought to be proxy measures of each of the respective respondent dimensions. In the future, these findings could be used to guide adaptive or responsive survey designs in the CPS by identifying subgroups that are more at risk of attrition, as well as those whose response rates are most impacted by receiving additional CPS Supplements.

Future research could also use qualitative and mixed-methods approaches to better understand respondent burden. For example, cognitive interviews could be designed to understand intra-household dynamics that take place when responding to a complex survey like the CPS that involves multiple waves, different household relationships, and proxy responses. The Yan et al. (2020) framework to gain insight on initial, cumulative, and continuous burden could be used to improve messaging in survey advance letters and within survey instruments to increase saliency. In addition, future research could examine paradata looking at where respondents tend to break-off during survey interviews as a proxy measure for questions that might be considered burdensome, and conduct qualitative research on how to reduce item effort and sensitivity for those items.

5.3. Conclusions

This article demonstrates a novel and innovative approach to measuring the impact of respondent burden on response rates when direct measures of respondent burden are not available. We believe our findings support the need to continue research that systematically measures both objective and subjective burden factors to further our understanding of the interplay between survey features and respondent characteristics that affect data quality and response rates. In an ideal world, burden would be directly measured in Federal Government surveys, and we encourage researchers to consider ways to directly measure burden in their production surveys, such as using interviewer observations and post-survey assessments (e.g., Fricker et al. 2012, 2014). When this is not possible, we would encourage researchers to look for ways to measure burden and its impact on response rates using existing data, as well as extend and replicate the findings in this research to other surveys.

As research on respondent burden continues to evolve, we believe these findings support the per-spective that subjective burden factors play an important role in predicting who participates in surveys and thus contributes to official statistics, and research should continue in support of expanding our understanding of how to measure and define respondent burden in the future.

6. Appendix

Table 4. Overview of the CPS schedule by calendar year, month, supplement survey, and month in sample.

Year	Month	Supplement survey	Panel																			
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
2015	January	Tobacco usage	1																				
	February		2	1																			
	March		3	2	1																		
	April		4	3	2	1																	
	May	Tobacco usage		4	3	2	1																
	June	Unbanked/underbanked			4	3	2	1															
	July	Computer and internet usage				4	3	2	1														
	August	Veterans					4	3	2	1													
	September	Volunteering and civic life						4	3	2	1												
	October	School enrollment							4	3	2	1											
	November	Computer and internet usage								4	3	2	1										
	December	Food security									4	3	2	1									
2016	January	Displaced workers/job tenure	5									4	3	2	1								
	February		6	5									4	3	2	1							
	March		7	6	5									4	3	2	1						
	April	Child support	8	7	6	5									4	3	2	1					
	May			8	7	6	5									4	3	2	1				
	June	Fertility			8	7	6	5									4	3	2	1			
	July					8	7	6	5									4	3	2	1		
	August	Veterans					8	7	6	5									4	3	2	1	
	September							8	7	6	5									4	3	2	1
	October	School enrollment							8	7	6	5									4	3	2
	November	Voting and registration								8	7	6	5									4	3
	December	Food security									8	7	6	5									4
2017	January											8	7	6	5								
	February												8	7	6	5							
	March													8	7	6	5						

Table 4. Continued

Year	Month	Supplement survey	Panel 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	April														8	7	6	5					
	May	Contingent workers														8	7	6	5				
	June	Unbanked/underbanked															8	7	6	5			
	July	Participation in the arts																8	7	6	5		
	August	Veterans																	8	7	6	5	
	September	Volunteering and civic life																		8	7	6	5
	October	School enrollment																			8	7	6
	November	Computer and internet usage																				8	7
	December	Food security																					8

[1]While the term panel is not used by the CPS, the authors use it to distinguish between rotating groups of sampled CPS households. The list of supplement surveys is not exhaustive, and only includes the supplement surveys that were controlled for in this article (including the supplements for which data was publicly available, excluding the Annual Social and Economic Supplements administered in March and April for which all households are eligible and thus is considered constant).

Table 5. Exploratory factor analysis factor loadings of CPS proxy burden measures and factor eigenvalues (λ) using oblique promax rotation.

CPS variable	Effort λ = 1.92	Length λ = 3.42	DIS stress λ = 6.04	DEM stress λ = 2.39
Household size	1.000	0.000	0.000	0.000
Number of children	1.000	0.000	0.000	0.000
Disability status	0.995	0.000	0.000	0.000
Labor force participation	0.986	0.000	-0.001	0.000
Marital status	0.965	0.001	0.000	-0.002
Age	-0.969	0.002	0.000	0.000
Month in sample	0.000	1.000	0.000	0.000
Number of previous supplements the household respondent was Sampled/eligible for	0.000	0.998	0.000	0.000
Data collection phase	0.000	0.994	0.000	0.000
Month in data collection	-0.001	0.987	0.000	0.000
Data collection month w/in phase	-0.001	0.984	0.000	0.000
Number of previous supplements the household was sampled/eligible for	0.020	0.888	0.000	0.000
NR – Difficulty walking or climbing stairs	0.000	0.000	1.000	0.007
NR – Difficulty dressing or bathing	0.000	0.000	1.000	0.007
NR – Difficulty doing errands alone	0.000	0.000	0.999	0.007
NR – Difficulty concentrating, remembering, or making decisions	0.000	0.000	0.999	0.007
NR – Blind or difficulty seeing	0.000	0.000	0.999	0.007
NR – Deaf or difficulty hearing	0.000	0.000	0.997	0.008
NR – Current military status (currently serving in the military)	0.000	0.000	0.000	1.000
NR – Race	0.000	0.000	0.002	0.987
NR – Ethnicity	0.000	0.000	0.003	0.979
NR – Veteran status (ever served in the military)	0.000	0.000	0.004	0.969
NR – Marital status	-0.001	0.000	0.008	0.935
NR – Age	0.000	0.000	0.026	0.880
NR – Education level	0.000	0.000	0.082	0.744

7. References

Abraham, K.G., S. Helms, and S. Presser. 2009. "How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States." *American Journal of Sociology* 114 (4): 1129–1165. DOI: https://doi.org/10.1086/595945.

Adua, L., and J.S. Sharp. 2010. "Examining survey participation and response quality: The significance of topic salience and incentives." *Survey methodology* 36 (1): 95–109.

Bates, Nancy. 2017. "The Morris Hansen Lecture. Hard-to-Survey Populations and the US Census: Making Use of Social Marketing Campaigns." *Journal of Official Statistics* 33 (4): 873–885. DOI: https://doi.org/10.1515/jos-2017-0040.

Below, T., B. Khamaldin, D. Mutabazi, D. Kirschke, C. Franke, S. Sieber, R. Siebert, and K. Tscherning. 2012. "Can farmers' adaptation to climate change be explained by socio-economic household-level variables?." *Global environmental change* 22(1): 223–235. DOI: https://doi.org/10.1016/j.gloenvcha.2011.11.012.

Berinsky, A. J. 2008. "Survey non-response." In *The SAGE Handbook of Public Opinion Research*: 309–322.

Bickart, B.A., J.M. Phillips, and J. Blair. 2006. "The Effects of Discussion and Question Wording on Self and Proxy Reports of Behavioral Frequencies." Marketing Letters 17: 167–180. DOI: https://doi.org/10.1007/s11002-006-5232-1.

Borlin, S. 2019. "Data quality of proxy reports: Inconsistent educational information in the German Microcensus Panel." Paper presented at the 8th Conference of the European Survey Research Association, July 16, Zagreb, Croatia. Available at: https://www.europeansurveyresearch.org/conf2019/uploads/654/2802/99/Data_Quality_Proxy_Reports_Boerlin.pdf (accessed November 2022).

Bradburn, N. 1978. "Respondent burden." In Proceedings of the Survey Research Methods Section of the American Statistical Association 35(40). Alexandria, VA, USA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf (accessed November 2022).

Branden, L. 1995. *The Effect of Interview Length on Attrition in the National Longitudinal Survey of Youth*. National Longitudinal Surveys Discussion Paper. Available at: https://files.eric.ed.gov/fulltext/ED406399.pdf (accessed November 2022).

Brick, J.M., and R. Tourangeau. 2017. "Responsive survey designs for reducing nonresponse bias." *Journal of Official Statistics* 33 (3): 735–752. DOI: http://dx.doi.org/10.1515/JOS-2017-0034.

Brick, J.M., and D. Williams. 2013. "Explaining rising nonresponse rates in cross-sectional surveys." *The ANNALS of the American academy of political and social science* 645(1): 36–59. DOI: https://doi.org/10.1177%2F0002716212456834.

Brust, O.A., S. Häder, and M. Häder. 2016. "Is the Short Version of the Big Five Inventory (BFI-S) Applicable for Use in Telephone Surveys?" *Journal of Official Statistics* 32(3) (2016): 601. DOI: https://doi.org/10.1515/jos-2016-0031.

Cattell, R.B. 1966 "The screen test for the number of factors." *Multivariate behavioral research* 1 (2): 245–276. DOI: https://doi.org/10.1207/s15327906mbr0102_10.

Cobb, C. 2018 "Answering for someone else: proxy reports in survey research." In *The Palgrave Handbook of Survey Research*: 87–93. Palgrave Macmillan, Cham. DOI: https://doi.org/10.1007/978-3-319-54395-6_12.

Coder, J., and L. Scoon-Rogers. 1996. *Evaluating the quality of income data collected in the annual supplement to the March Current Population Survey and the Survey of Income and Program Participation*. Working Paper. U.S. Bureau of the Census. Available at: citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.283.6445&rep = rep1&type = pdf (accessed November 2022).

Copeland, K., and J.M. Rothgeb. 1990. "Testing alternative questionnaires for the Current Population Survey."In *Proceedings of the Section on Survey Research Methods 1990*: 63–71. Available at: http://www.asasrms.org/Proceedings/papers/1990_010.pdf (accessed November 2022).

Costello, A.B., and J. Osborne. 2005. "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis." *Practical assessment, research, and evalua-tion* 10(1): 7. DOI: https://doi.org/10.7275/jyj1-4868.

Couper, M., A. Kapteyn, M. Schonlau, and J. Winter. 2007. "Noncoverage and nonresponse in an Internet survey." *Social Science Research* 36 (1): 131–48.

Couper, M.P., E. Singer, F.G. Conrad, and R.M. Groves. 2008. "Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation." *Journal of Official Statistics* 24(2): 255–275. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/risk-of-disclosure-perceptions-of-risk-and-concerns-about-privacy-and-confidentiality-as-factors-in-survey-participation.pdf.

Crawford, S.D., M.P. Couper, and M.J. Lamias. 2001. "Web surveys: Perceptions of burden." *Social science computer review* 19 (2): 146–162. DOI: https://doi.org/10.1177%2F089443930101900202.

Cunningham, J.A., D. Ansara, T.C. Wild, T. Toneatto, and A. Koski-Jännes. 1999. "What is the price of perfection? The hidden costs of using detailed assessment instruments to measure alcohol consumption." *Journal of Studies on Alcohol* 60 (6): 756–758. DOI: https://doi.org/10.15288/jsa.1999.60.756.

DeMaio, T.J. 1980. "Refusals: Who, where and why." *Public Opinion Quarterly* 44 (2): 223–233. DOI: https://doi.org/10.1086/268586.

Earp, M., M. Mitchell, J. McCarthy, and F. Kreuter. 2014. "Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey." *Journal of Official Statistics (JOS)* 30 (4): 701–719. DOI: http://dx.doi.org/10.2478/JOS-2014-0044.

Earp, M., D. Toth, P. Phipps, and C. Oslund. 2018. "Assessing Nonresponse in a Longitudinal Establishment Survey Using Regression Trees" *Journal of Official Statistics* 34 (2) 463–481. DOI: http://dx.doi.org/10.2478/JOS-2018-0021.

Edwards, N.E., and P.S. Scheetz. 2002. "Predictors of burden for caregivers of patients with Parkinson's disease." *Journal of Neuroscience Nursing* 34 (4): 184. Available at: https://www.proquest.com/scholarly-journals/predictors-burden-cregivers-patients-with/docview/219178824/se-2?accountid = 26724 (accessed November 2022).

Finch, H. 2006. "Comparison of the performance of varimax and promax rotations: Factor structure recovery for dichotomous items." *Journal of Educational Measurement* 43(1): 39–52. DOI: https://doi.org/10.1111/j.1745-3984.2006.00003.x.

Fricker, S.S. 2007. *The relationship between response propensity and data quality in the Current Population Survey and the American Time Use Survey*. University of Maryland, College Park. Available at: https://www.proquest.com/dissertations-theses/relationship-between-response-propensity-data/docview/304850876/se-2?accountid = 26724 (accessed November 2022).

Fricker, S., C. Kreisler, and L. Tan. 2012. "An exploration of the application of PLS path modeling approach to creating a summary index of respondent burden." In *JSM Proceedings*: 4141–4155. Available at: http://www.asasrms.org/Proceedings/y2012/Files/304802_73760.pdf (accessed November 2022).

Fricker, S., T. Yan, and S. Tsai. 2014. "Response burden: What predicts it and who is burdened out." In *JSM proceedings* : 4568–4577. Available at: http://www.asasrms.org/Proceedings/y2014/files/400298_500838.pdf (accessed November 2022).

Fuchs, M. 2005. "Children and Adolescents as Respondents. Experiments on Question Order, Response Order, Scale Effects and the Effect of Numeric Values Associated with Response Options." *Journal of Official Statistics* 21 (4): 701–725. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/children-and-adolescents-as-respondents.-experiments-on-question-order-response-order-scale-effects-and-the-effect-of-numeric-values-associated-with-response-options.pdf (accessed November 2022).

Galesic, M. 2006. " Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics* 22 (2): 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf (accessed November 2022).

Groves, R.M. 2006. "Nonresponse rates and nonresponse bias in household surveys." *Public opinion quarterly* 70(5): 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M., R.B. Cialdini, and M.P. Couper. 1992. "Understanding the decision to participate in a survey." *Public opinion quarterly* 56 (4): 475–495. DOI: https://doi.org/10.1086/269338.

Groves, R.M., S. Presser, and S. Dipko. 2004. "The role of topic interest in survey participation decisions." *Public Opinion Quarterly* 68 (1): 2–31. DOI: https://doi.org/10.1093/poq/nfh002.

Groves, R.M., E. Singer, and A. Corning. 2000. "Leverage-saliency theory of survey participation: description and an illustration." *The Public Opinion Quarterly* 64 (3): 299–308. Available at: https://www.jstor.org/stable/3078721 (accessed November 2022).

Haraldsen, G. 2004 "Identifying and Reducing Response Burdens in Internet Business Surveys." *Journal of Official Statistics* 20(2): 393–410. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/identifying-and-reducing-response-burdens-in-internet-business-surveys.pdf (accessed November 2022).

Hinkin, T.R. 1995. "A review of scale development practices in the study of organizations." *Journal of management* 21 (5): 967–988. DOI: https://doi.org/10.1177%2F014920639502100509.

Hinkin, T.R. 1998. "A brief tutorial on the development of measures for use in survey questionnaires." *Organizational research methods* 1 (1): 104–121. DOI: https://doi.org/10.1177%2F109442819800100106.

Howard, M.C. 2016. "A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?." *International Journal of Human-Computer Interaction* 32 (1): 51–62. DOI: https://doi.org/10.1080/10447318.2015.1087664.

Kaiser, H.F. 1960. "The application of electronic computers to factor analysis." *Educational and psychological measurement* 20 (1): 141–151. Available at: https://journals.sagepub.com/doi/pdf/10.1177/001316446002000116 (accessed November 2022).

Kaplan, R., and J. Holzberg. 2020. *Measuring subjective perceptions of burden over time*. Paper presented at the American Association for Public Opinion Research Conference, June 11.

Kojetin, B.A., and P. Mullin. 1995. "The quality of proxy reports on the current population survey (CPS)." In *50th Annual Conference of the American Association for Public Opinion Research.* Available at: http://www.asasrms.org/Proceedings/papers/1995_193.pdf (accessed November 2022).

Kolenikov, S., and G. Angeles. 2009. "Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?" *Review of Income and Wealth* 55 (1): 128–165. DOI: https://doi.org/10.1111/j.1475-4991.2008.00309.x.

Lee, S., N.A. Mathiowetz, and R. Tourangeau. 2004. "Perceptions of Disability: the Effect of Self-and Proxy Response." *Journal of Official Statistics* 20 (4): 671–686. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/perceptions-of-disability-the-effect-of-self–and-proxy-response.pdf (accessed November 2022).

Loosveldt, G., J. Pickery, and J. Billiet. 2002."Item Nonresponse as a Predictor of Unit Nonresponse in a Panel Survey." *Journal of Official Statistics* 18 (4) 4 (2002): 545–557. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/i-tem-nonresponse-as-a-predictor-of-unit-nonresponse-in-a-panel-survey.pdf (accessed November 2022).

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schütz. 2007 "Compensating for low topic interest and long surveys: a field experiment on nonresponse in web surveys." *Social Science Computer Review* 25 (3): 372–383. DOI: https://doi.org/10.1177%2F0894439307297606.

Mechanic, D., and M. Newton. 1965. "Some problems in the analysis of morbidity data." *Journal of Chronic Diseases* 18 (6): 569–580. DOI: https://doi.org/10.1016/0021-9681(65)90078-0.

Montgomery, M.R., M. Gragnolati, K.A. Burke, and E. Paredes. 2000. "Measuring living standards with proxy variables." *Demography* 37 (2): 155–174. DOI: https://doi.org/10.2307/2648118.

Morgan, J.N., and J.A. Sonquist. 1963. "Problems in the analysis of survey data, and a proposal." *Journal of the American statistical association* 58 (302): 415–434. DOI: https://www.tandfonline.com/action/showCitFormats?doi = 10.1080/01621459.1963.10500855.

Norris, M., and L. Lecavalier. 2010. "Evaluating the use of exploratory factor analysis in developmental disability psychological research." *Journal of autism and developmental disorders* 40 (1): 8–20. DOI: https://doi.org/10.1007/s10803-009-0816-2.

Paperwork Reduction Act Guide. 2011. Office of Personnel Management. Available at: https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/-paperwork-reduction-act-guide.pdf (accessed November 2022).

Perruccio, A.V., and E.M. Badley. 2004. "Proxy reporting and the increasing prevalence of arthritis in Canada." *Canadian journal of public health* 95 (3): 168–173. DOI: https://doi.org/10.1007/BF03403641.

Phipps, P., and D. Toth. 2012. "Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data." *The Annals of Applied Statistics*: 772–794. DOI: http://doi.org/10.1214/1 1-AOAS521.

Polivka, A.E. 1996. "Data watch: The redesigned current population survey. *Journal of Economic Perspectives* 10(3): 169–180. Available at: https://www.jstor.org/stable/41713473 (accessed November 2022).

Rohe, W.M., S. van Zandt, and G. McCarthy. 2002. "Home ownership and access to opportunity." *Housing Studies* 17 (1): 51–61. DOI: https://doi.org/10.1080/02673030120105884.

Rolstad, S., J. Adler, and A. Rydén. 2011. "Response burden and questionnaire length: is shorter better? A review and meta-analysis." *Value in Health* 14 (8): 1101–1108. DOI: https://doi.org/10.1016/j.jval.2011.06.003.

Serfling, O. 2004. *The interaction between unit and item nonresponse in view of the reverse cooperation continuum*. No. 2004/02. WWZ Discussion Paper. DOI: https://doi.org/10.5451/unibas-ep61271.

Sharp, L.M., and J. Frankel. 1983. "Respondent burden: A test of some common assumptions." *Public opinion quarterly* 47(1): 36–53. DOI: https://doi.org/10.1086/268765.

Sitzia, J., and N. Wood. 1998. "Response rate in patient satisfaction research: an analysis of 210 published studies." *International Journal for Quality in Health Care* 10 (4): 311–317. Available at: https://academic.oup.com/intqhc/article-pdf/10/4/311/5083940/100311.pdf (accessed November 2022).

Stanley McCarthy, J., D.G. Beckler, and S.M. Qualey. 2006. "An Analysis of the Relationship Between Survey Burden and Nonresponse: If we Bother Them More, Are They Less Cooperative?" *Journal of Official Statistics* 22 (1): 97–112. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-analysis-of-the-relationship-between-survey-burden-and-nonresponse-if-we-bother-them-more-are-they-less-cooperative.pdf (accessed November 2022).

Sudman, S., B. Bickart, J. Blair, and G. Menon. 1994. "The effect of participation level on reports of behavior and attitudes by proxy reporters." In *Autobiographical memory and the validity of retrospective reports*: 251–265. Springer, New York, NY, 1994. DOI: https://doi.org/10.1007/978-1-4612-2624-6_17.

Tabachnick, B.G., and L.S. Fidell. 2001. "Principal components and factor analysis." *Using multivariate statistics* 4 (1): 582–633.

Tennant, A., E.M. Badley, and M. Sullivan. 1991. "Investigating the proxy effect and the saliency principle in household based postal questionnaires." *Journal of Epidemiology & Community Health* 45(4): 312–316. DOI: http://dx.doi.org/10.1136/jech.45.4.312.

Tourangeau, R., and T. Yan. 2007. "Sensitive questions in surveys." *Psychological bulletin* 133 (5): 859–883. DOI: http://dx.doi.org/10.1037/0033-2909.133.5.859.

Toth, D. 2019. *Recursive Partitioning for Modeling Survey Data*. R package version 0.4.0. Available at: https://CRAN.R-project.org/package = rpms (accessed November 2022).

Toth, D. 2020. "A Permutation Test on Complex Sample Data." *Journal of Survey Statistics and Methodology* 8 (4): 772–791. DOI: https://doi.org/10.1093/jssam/smz018.

Toth, D., and P. Phipps. 2014. Regression tree models for analyzing survey response. In *Proceedings of the government statistics section*: 339–351. American Statistical Association.

U.S. Bureau of Labor Statistics Handbook of Methods. 2018. *Consumer Expenditure and income*. Available at: https://www.bls.gov/opub/hom/cex/home.htm (accessed November 2022).

U.S. Bureau of Labor Statistics Handbook of Methods 2018. *Current Population Survey*. Available at: https://www.bls.gov/opub/hom/cps/home.htm (accessed November 2022).

U.S. Census Bureau. 2021. *Current Population Survey Technical Documentation: Methodology*. Available at: https://www.census.gov/programs-surveys/cps/technical-documentation/methodology.html.

Van Liere, K.D., and R.E. Dunlap. 1980. "The social bases of environmental concern: A review of hypotheses, explanations and empirical evidence." *Public opinion quarterly* 44 (2): 181–197. DOI: https://doi.org/10.1086/268583.

Yan, T., and R. Curtin. 2010. "The relation between unit nonresponse and item nonresponse: A response continuum perspective." *International Journal of Public Opinion Research* 22 (4): 535–551. DOI: https://doi.org/10.1093/ijpor/edq037.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response Burden: What Is It and What Predicts It?" In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P. Beatty, D. Collins, L. Kaye, J. Luis Padilla, G. Willis, and A. Wilmot: 193–212. Wiley & Sons. DOI: https://doi.org/10.1002/9781119263685.ch8.

Yan, T., and R. Tourangeau. 2008. "Fast times and easy questions: The effects of age, experience and question complexity on web survey response times." *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 22 (1): 51–68.

Zhang, C., S. Lonn, and S.D. Teasley. 2017. "Understanding the impact of lottery incentives on web survey participation and response quality: A leverage-salience theory perspective." *Field Methods* 29 (1): 42–60. DOI: https://doi.org/10.1177%2F1525822X16647932.

Received February 2021
Revised July 2021
Accepted July 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1177–1203, http://dx.doi.org/10.2478/JOS-2022-0050

Exploring Burden Perceptions of Household Survey Respondents in the American Community Survey

Jessica Holzberg[1] *and Jonathan Katz*[1]

Minimizing respondent survey burden may help decrease nonresponse and increase data quality, but the measurement of burden has varied widely. Recent efforts have paid more attention to respondents' subjective perceptions of burden, measured through the addition of questions to a survey. Despite reliance on these questions as key measures, little qualitative research has been conducted for household surveys. This study used focus groups to examine respondents' reactions to possible sources of burden in the American Community Survey (ACS) such as survey length, sensitivity, and contact strategy; respondents' knowledge, attitudes, and beliefs about burden; and overall perceptions of burden. Feedback was used to guide subsequent selection and cognitive testing of questions on subjective perceptions of burden. Generally, respondents did not find the ACS to be burdensome. When deciding whether it was burdensome, respondents thought about the process of responding to the questionnaire, the value of the data, that response is mandatory, and to a lesser extent, the contacts they received, suggesting these constructs are key components of burden in the ACS. There were some differences by response mode and household characteristics. Findings reinforce the importance of conducting qualitative research to ensure questions capture important respondent burden perceptions for a particular survey.

Key words: Burden measurement; household surveys; federal surveys.

1. Introduction

"Respondent burden" is a multi-faceted concept that includes survey characteristics and how they are perceived by respondents (Bradburn 1978). When respondents find a survey too burdensome, they may engage in undesirable respondent behaviors such as breaking off from a survey early, refusing to respond to a question or survey, or attempting to complete the survey as quickly as possible (i.e., satisficing). Understanding respondent burden and lessening it may decrease these undesirable respondent behaviors, and minimizing burden is also respectful of the respondents upon whom we rely for data.

Survey characteristics impacting burden may include the length of the survey, the effort required to complete it, "stress on the respondent" (sensitivity), and the frequency of survey requests respondents receive in general (i.e., whether they are "over-surveyed"; Bradburn 1978). The effect of these survey characteristics on burden can vary based on

[1] Center for Behavioral Science Methods, U.S. Census Bureau, 4600 Silver Hill Road, Washington D.C., 20233, U.S.A. Emails: Jessica.Holzberg@census.gov and Jonathan.M.Katz@census.gov

Acknowledgments: The authors sincerely thank the U.S. Census Bureau American Community Survey Office (ACSO) for their support of this research. The authors would also like to thank Mary Davis, Elizabeth Poehler, Brian Wilson, Dave Raglin, Joanne Pascale, Paul Beatty, and three anonymous reviewers for their helpful comments on this manuscript. This article is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau. The article has been reviewed for disclosure avoidance and approved under CBDRBFY20-CBSM002-004.

respondent characteristics and their beliefs, knowledge, and attitudes. Respondents may also weigh characteristics differently depending on the nature of a specific survey request. One survey topic may be more sensitive for one respondent than it is for another, and a respondent's belief that a particular survey is important may have more influence on their perceptions of burden than their belief that it is too long (Bradburn 1978). Therefore, it is important to distinguish between measures of objective and subjective burden. Objective burden refers to burden that can be measured by the researcher, such as length or item nonresponse rates, while subjective burden refers to respondents' self-reported perceptions and attitudes (Sharp and Frankel 1983).

In practice, many researchers have focused primarily on objective burden, using breakoffs, nonresponse, or satisficing as outcome measures. While important, this research does not provide insight into respondents' opinions of the survey experience from their own, subjective perspective. There is some evidence that there is no or only a limited relationship between objective and subjective burden (e.g., Yan et al. 2020; Read 2019; Kaplan and Holzberg 2019). As a result, there has been an increase in efforts to measure respondents' subjective perceptions of burden, to be used alone or in combination with objective measures.

Typically, subjective burden is measured through the addition of one or more questions to a survey. However, despite reliance on these questions as key measures, we are unaware of any large-scale qualitative studies on the measurement of respondents' perceptions of burden in household surveys, which are surveys that ask questions about people living in a sampled housing unit. Groves et al. (1999) used vignettes to assess the impacts of incentives, refusal conversion techniques, and survey length on response, but this research was based on response to a hypothetical survey. Focus groups and cognitive interviews have been used to learn about perceptions of burden in establishment surveys, or surveys of organizations and businesses (Hedlin et al. 2005). While there are some parallels between burden in establishment and household surveys, there are also important differences. Since establishment survey respondents respond on behalf of an organization (as opposed to themselves and their household), respondents go through the response process both at an individual level and at an organizational level, which may affect perceptions of burden (Bavdaž 2010). Furthermore, while the impact of some survey and respondent characteristics on perceptions of burden may be fairly consistent across surveys, others may be survey-specific based on the survey's design. There is generally a need for more qualitative research to ensure questions capture the most important respondent perceptions for a particular household survey. A respondent-driven approach to question development helps ensure that respondents are asked about aspects of burden relevant to them, and that questions are well understood and easy to answer.

In this research, we collected initial feedback from respondents about survey burden for a U.S. federal survey that does not currently include subjective burden perception questions, the American Community Survey (ACS). Respondent feedback was collected via focus groups and was used to guide subsequent selection, development, and cognitive testing of questions on subjective perceptions of burden (Holzberg et al. 2021). If implemented on the ACS, subjective burden questions could be used to examine overall perceptions of burden of the ACS and whether perceptions differ over time or by respondent or household characteristics. In this article, we first review the literature to

identify survey and respondent characteristics that may play a role in burden. The goal of this literature review was to select critical topics on which to solicit ACS respondent feedback. We then discuss our methods for data collection and findings from the focus groups. We conclude with a discussion of future areas for research in the measurement of subjective burden.

2. Measuring Burden

In the following section, we focus on burden research conducted on household surveys using various measures of respondent burden, organized by the four components of burden originally outlined by Bradburn (length, effort, sensitivity, and over-surveying). We also discuss other survey features and respondent characteristics that may potentially contribute to burden as well as whether a single overall burden question can be used to capture respondent perceptions. We include examples of survey questions measuring respondent perceptions from the Consumer Expenditure Survey (CE), a survey providing data on consumer spending and, at the time this research was conducted, one of the only federal household surveys in the United States that has collected data on subjective perceptions of burden.

2.1. Survey Characteristics

2.1.1. Length

The Paperwork Reduction Act requires U.S. federal agencies to publicly report the "paperwork burden" a given data collection activity is expected to require of respondents, defined as the estimated number of minutes a typical respondent would need to complete the task (U.S. Office of Personnel Management 2011). Perhaps because of this, survey length has been one of the most researched components of burden, with research showing the relationship between objective length and nonresponse is mixed (Bogen 1996). While objective survey length can be measured using an estimate or available paradata, subjective length can also be examined via questions about respondents' perceptions, as has been asked in the CE (U.S. Bureau of Labor Statistics 2012, 2013):

> *"Do you feel that the length of today's interview was too long, too short, or about right?"*
> *"Would you say that the time and effort you put into answering the survey questions was very well spent, somewhat well spent, a little well spent, or not at all well spent?"*

Studies examining subjective perceptions of length in household surveys have been mixed. Some research has found that respondents who were assigned to a shorter questionnaire reported the survey to be less burdensome, with some improvements in data quality, and that perceptions of length were an important dimension of burden (Fricker et al. 2012; Yu et al. 2015; Fricker et al. 2011; Fricker et al. 2014). On the other hand, there is evidence that length and perceptions of burden were not correlated (e.g., Kaplan and Holzberg 2019). Dahlhamer et al. (2019) found that length was a less important factor in perceptions of burden than difficulty, sensitivity, and other respondent or survey interview characteristics. Sharp and Frankel (1983) also found that while respondents assigned to a

longer 75-minute interviewer-administered survey felt more negatively about the length of the survey than those who took a 25-minute survey, the two groups did not differ in item nonresponse or broad perceptions of survey burden.

2.1.2. Effort

Burden research on respondent effort has generally focused on the difficulty of the task, which can include the amount of thought or consultation with records required to respond, or obstacles to responding presented by a poorly designed survey (Haraldsen 2004; Tourangeau and Plewes 2013). The perceived cognitive burden of the survey topic (e.g., health) may impact respondents' decision to participate in a survey at all (Groves and Couper 1998). Respondents who do participate but find a survey requires more effort than they are willing to give may engage in undesirable behaviors such as speeding (e.g., Malhotra 2008), straightlining (e.g., Kim et al. 2019), and satisficing (Krosnick 1991). The CE (U.S. Bureau of Labor Statistics 2012, 2013): has asked questions about perceived difficulty and effort:

> *"How difficult or easy was it for you to answer the questions in this survey? (Very easy, somewhat easy, somewhat difficult, very difficult)"*
> *"Thinking about the amount of effort that you put into answering today's survey, would you say that you put forth a lot of effort, a moderate amount of effort, or a little effort?"*

Galesic (2006) found that half of the variability in respondents' subjective assessments of burden in a questionnaire was explained by question characteristics that varied in effort required, such as open versus closed-ended format. Burdened respondents were also more likely to drop out of the survey. Researchers have also considered whether asking questions that require respondents to consult records increases burden because it requires more effort (Phipps 2014; Hedlin et al. 2005). Sharp and Frankel (1983) asked some of their respondents to use records when completing a survey. Those asked to consult records did so, but the authors did not find any evidence of this impacting burden perceptions. On the other hand, in a study in which respondents were not told they had to use records, Yang (2015) found a significant association between records use (as observed by the survey interviewer) and burden perceptions. In one recent study, Kunz and Gummer (2019) examined effort and difficulty independently. Respondents tended to rate the survey as more effortful than difficult, suggesting perhaps these are separate concepts. Difficulty seemed to correlate more strongly with other potential measures of perceived burden such as distraction and fatigue. Bradburn (1978) also noted that effort can correlate with length, though some longer surveys do not require much effort and vice versa.

2.1.3. Sensitivity

Bradburn (1978) defined respondent stress as the level of "personal discomfort" or sensitivity a respondent feels when responding to a survey. There is a considerable amount of survey literature on minimizing the impact of sensitive questions due to their potential role in motivated misreporting (e.g., Tourangeau and Yan 2007). The CE (U.S. Bureau of Labor Statistics 2012, 2013) has asked respondents about their perceptions of question sensitivity:

"How sensitive did you feel the questions I asked today were? (Very sensitive, somewhat sensitive, a little sensitive, not at all sensitive)"

However, sensitivity has overall received fairly little research attention as it pertains to subjective burden and has been generally absent from studies of burden in establishment surveys (Hedlin et al. 2005). In one study, Kaplan and Fricker (2017) found that perceptions of question sensitivity predicted overall perceptions of burden. Earp et al. (2019) also found that sensitivity had a consistently negative impact on response rates to Current Population Survey supplements. Bradburn (1978) noted that it is hard to examine how much perceptions of sensitivity might affect nonresponse. Some researchers have discussed that a given survey can be both difficult and sensitive but generally do not examine those characteristics separately (e.g., McCarthy et al. 2006). Haraldsen (2004, 397) also stated that respondent stress seems more like "an aspect of a burdensome feeling" than a cause of burden.

2.1.4. Over-surveying

Researchers have echoed Bradburn's concern that the growing prevalence of survey requests may lead to "survey fatigue" or an "over-surveying effect" (Groves and Couper 1998; Olson 2014; Leeper 2019). This effect may be particularly strong for groups that are frequently oversampled for surveys, such as demographic minorities (Leeper 2019). In the U.S. federal government alone, the number of surveys increased by 50% between 1984 and 2004 (Presser and McCulloch 2011). Researchers have been particularly concerned about the effects of survey fatigue in establishment surveys (Haraldsen 2004; Giesen et al. 2011; Phipps 2014). Many establishment surveys sample some establishments with certainty because of their large influence on estimates; therefore, these establishments may receive many requests. While most evidence of survey fatigue for establishments remains anecdotal, McCarthy and Beckler (2000) generally did not find a relationship between the number of prior survey requests and establishment survey respondent attitudes towards the U.S. National Agricultural Statistics Service (NASS). However, responding establishments who had been contacted more times said they thought it was important to respond more than establishments who had been contacted fewer times did. For household surveys, Sharp and Frankel (1983) found no evidence of a follow-up request impacting attitudes or behavior in a two-interview study and Read (2019) found that the burden of continued participation in a receipt scanning mobile app study was small. While there has also been discussion about the potential for an individual survey's contact strategy to contribute to burden (e.g., Hedlin et al. 2008; Giesen et al. 2011; Hughes et al. 2016; Fricker 2016), there has been little examination of this, other than a question asked on the CE (U.S. Bureau of Labor Statistics 2012, 2013):

"Thinking about the number of calls you received before each interview, would you say it was too many, or did it seem like a reasonable number?"

2.1.5. Other Survey Characteristics

In addition to those discussed by Bradburn (1978), a variety of other survey characteristics may impact both objective and subjective burden. While incentives are often useful for increasing response, it is less clear whether there is a relationship between incentives and

perceived burden (Galesic 2006; Hedlin et al. 2005; Lee and Knappenberger 2019). Other survey characteristics that may contribute to perceived burden include the survey sponsor (Haraldsen 2004; Jones 2012), perceived legitimacy of the request (Dale et al. 2007), mandatory response (Phipps 2014), survey mode (Haraldsen 2004; Yan et al. 2020; Peytchev 2009), how the questionnaire is designed (Peytchev 2009; Crawford et al. 2001), and whether the survey request is sent at a convenient or inconvenient time for respondents (Haraldsen 2004; Hedlin et al. 2005).

2.2. *Respondent Characteristics*

As with survey characteristics, a number of respondent characteristics have been proposed as potential predictors of burden, including demographic and household characteristics (such as age, education, and household size), respondent knowledge and attitudes, and interest in the subject matter of a survey (Hedlin et al. 2005). It is well known that some demographic groups of respondents are more difficult to reach and interview than others, raising concerns for researchers worried about response rates and nonresponse bias. What limited literature on the interaction between respondent demographics and perceptions of burden exists seems to be somewhat mixed. Sharp and Frankel (1983) and Fricker et al. (2011) found little evidence overall of significant differences in perceptions of burden by respondent demographics. On the other hand, Galesic (2006) and Kaplan and Fricker (2017) found that sex predicted perceptions of burden, with male respondents having higher perceptions of burden. Earp et al. (2019) found that education, disability status, and presence of a child in the household consistently affected factor loadings in an exploratory factor analysis on burden. One study using data from the CE found that household size was collinear with other predictors in a burden model and thus excluded it from analysis (Yang 2015).

Respondents' knowledge of and attitudes towards a survey may impact their perceptions of burden. Knowledge and attitudes may be preconceived and "insulate the sample member from processing new information about a specific survey" or they can be "quickly formed based on an impression of a specific [survey] request" (Tourangeau and Plewes 2013, 64). Specific factors identified by researchers as potentially contributing to burden include understanding of the survey's purpose, perceptions of the survey's usefulness, motivation, and general attitudes towards the government or about privacy (Fricker 2016; McCarthy and Beckler 2000; Hedlin et al. 2005; Tortora 2017; Wenemark et al. 2010; Jones 2012; Yan et al. 2020). There have been a few studies examining the role of general attitudes in perceptions of burden. Sharp and Frankel (1983) found that respondents who agreed that surveys are beneficial were willing to be reinterviewed at similar rates regardless of assignment to a less or more burdensome experimental condition. They also found that respondents who held attitudes about surveys being "too personal" had higher burden perceptions. Similarly, in a 1979 study, Jones et al. (1979, 69) found that farmers and ranchers "who are convinced that surveys produce useful and accurate information that serves primarily their own economic interest tend not to feel burdened by even large numbers of surveys. Those who are not so convinced are likely to feel that even one survey request is too many". In a qualitative study, respondents said that the perceived usefulness of survey questions is a component of burden (Yu et al. 2015).

Trust in the survey sponsor is another attitude that may impact perceptions of burden, as it may be associated with respondent's motivation to be cooperative and put forth effort into responding (Fricker et al. 2014). The CE has asked several attitudinal questions of respondents, including (U.S. Bureau of Labor Statistics 2012, 2013):

> *"I see why it was necessary to collect this information. (Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)"*
> *"I trust the U.S. Census Bureau to safeguard the information that I have provided them"*

Respondents' interest may be something that surveys can manipulate; for example, by changing how the survey request is framed to respondents (Haraldsen 2004); interest may also play a role in respondents' willingness to continue with a survey (Galesic 2006). The CE has asked respondents about their interest in the survey (U.S. Bureau of Labor Statistics 2012, 2013):

> *"How interesting was this survey to you? (Very interesting, somewhat interesting, a little interesting, not at all interesting)"*

Some researchers have found evidence that interest significantly predicts overall perceptions of burden (Fricker et al. 2011; Fricker et al. 2012; Kaplan and Fricker 2017). In an exploratory factor analysis, Earp et al. (2019) found that the topic of a Current Population Survey supplement impacted factor loadings.

2.2.1. Measuring Overall Perceived Burden

Some researchers have explored using an overall subjective burden question in addition to or in lieu of more detailed measures assessing survey or respondent characteristics. The CE (U.S. Bureau of Labor Statistics 2012, 2013) has asked:

> *"How burdensome was this survey to you? (Very burdensome, somewhat burdensome, a little burdensome, not at all burdensome)"*

A single measure of overall perceived burden may be sufficient as responses to the question have correlated with other measures of reluctance (Galesic 2006; Yang 2015; Yan et al. 2020). Therefore, while it may be useful to ask respondents a series of questions about their perceptions of survey burden, it may not be necessary. This approach has also been beneficial in that it does not require researchers to tease out individual effects of respondent and survey characteristics and it does not add a lot of time to a survey, therefore avoiding the possibility of increasing the burden of a survey by asking questions about burden (Dale et al. 2007). In a field test comparison of the current National Health Interview Survey (NHIS) questionnaire to a shortened, redesigned questionnaire, responses to an overall perceived burden question differed significantly by condition (Dahlhamer et al. 2019). The percentage of respondents reporting that the survey was "not at all burdensome" was 10% higher in the redesign condition. We note that there has been only minimal pretesting of an overall perceived burden question using the word "burdensome" in household surveys. While questions using "burdensome" have been pretested and used in establishment surveys, this word may not be well-understood by all household survey respondents, some of whom may have lower levels of education.

3. Methods

In spring 2017, we conducted ten focus groups with 46 recent ACS respondents in three cities to learn about their perceptions of the ACS and its burden. Focus groups are a method of qualitative research in which a researcher learns about the opinions, feelings, and beliefs of respondents in a group setting (Morgan 1998). They foster rich, in-depth discussion because respondents can speak to each other as well as to the moderator, and comments made by other respondents may help remind respondents of things they had forgotten. Focus groups are sometimes used as a method of preliminary, exploratory research before new survey questions are written to measure a construct. Usually, multiple focus groups are conducted, and analysis focuses on themes identified across groups, as identified by researchers upon review of notes, recordings, and transcripts. Focus group analyses typically do not rely on reporting exact numbers, as not every respondent in a focus group answers a given question and quiet respondents may or may not agree with other opinions voiced (Ritchie et al. 2003).

Findings from focus groups help researchers learn what is salient to respondents and what words and phrases respondents use to describe concepts of interest, therefore making it easier to develop questions that are relevant and easy for respondents to understand. Sometimes focus groups generate new ideas of which researchers had not previously thought. For this research, taking this initial exploratory step was helpful because subjective burden has been studied less frequently than objective burden. This approach mimicked qualitative research undertaken at national statistical institutes in the United Kingdom, Norway, and Sweden to develop a burden model and perceived burden questions for establishment surveys (Hedlin et al. 2005). We note that this approach did not provide insight into how nonrespondents perceive the burden of the ACS.

As a starting point for the focus group discussion, we used the framework of Bradburn (1978) and findings of the literature review to develop a moderator's guide focused on survey and respondent characteristics that could contribute to burden perceptions in the ACS, including question sensitivity, effort, length, and respondents' knowledge, attitudes, and beliefs. To leverage the exploratory nature of the focus groups, we also included ACS-specific features that we expected could uniquely contribute to burden perceptions in the survey, such as the contact strategy and the fact that response is mandatory. We collected respondent feedback on these topics to aid in development of burden questions that could potentially be added to the ACS. In subsequent research not discussed in this article, we then conducted individual interviews with additional ACS respondents to cognitively test these questions and collect additional feedback on the burden of the ACS (Holzberg et al. 2021).

3.1. The American Community Survey

Formerly administered as the long form in the decennial census, the ACS is conducted by the U.S. Census Bureau. It serves as the premier source for detailed population and housing information about the United States and is used to help allocate USD 675 billion to local communities (Hotchkiss and Phelan 2017). The ACS is particularly unique among U.S. household surveys in both its size and its mandatory response requirement – approximately 3.5 million addresses are sampled over the course of a year and response to the ACS is required by U.S. law. The ACS produces estimates on over 35 demographic,

social, economic, and housing topics such as age, marital status, disability, income, and internet access. These questions vary in their potential perceived sensitivity and effort/difficulty. Except for questions on housing, most ACS questions are asked at a person-level about most or all members of the household; therefore, the amount of time spent answering the survey can vary depending on the size of the household. On average, the U.S. Census Bureau estimates it takes 40 minutes to complete the ACS.

At the time the present research began, all mailable sampled addresses received two pieces of mail that ask a recipient to complete the survey online on behalf of the household. If no internet self-response is received, sampled addresses are provided a paper questionnaire in a third mailing and are also sent a fourth mailing reminder. If a response was not received by internet or mail before the fifth mailing, the address was either sent a fifth mailing, or the U.S. Census Bureau attempted to complete the ACS by computer-assisted telephone interview (CATI) using telephone numbers obtained by external sources. A subsample of addresses for which an internet, paper, or CATI response was not received were then selected for computer-assisted personal interview (CAPI) follow-up. Respondents could still self-respond online or by paper during these latter phases of data collection. The CATI operation was eliminated in late 2017 (U.S. Census Bureau 2017). More detailed information about the ACS methodology is available online (U.S. Census Bureau 2014).

3.2. Sample and Recruitment

Focus group respondents responded to the ACS in the one to two months immediately preceding the focus groups. Nonrespondents were not in scope for this research. Since nonrespondents had not responded to multiple ACS mailings, and in some cases, phone calls or personal visits from interviewers, the expected effort associated with recruiting nonrespondents was higher than our resources allowed for this research. Groups ranged in size from three to eight respondents, and each focus group respondent received USD 75. The groups were conducted by one of two experienced moderators who were trained on the moderator's guide before the first focus group and were also audio and video recorded. The first two focus groups were also attended by several observers.

Because of the sequential data collection mode strategy, many ACS respondents who self-respond do so earlier in the data collection period than others and therefore may be more eager to respond to the ACS. ACS respondents who complete the survey with an interviewer respond later (i.e., after multiple contacts) and therefore may generally be more reluctant to respond. This reluctance to participate may be related to respondents' perceptions of burden of the ACS, so to ensure both perspectives were captured we separated the focus groups by response mode. Five of the ten groups were with those who responded in a self-administered mode (on the internet or paper), and five were with those who responded in an interviewer-administered mode (by CATI or CAPI; see Table 1).

We recruited prior ACS respondents for the focus groups by telephone, using a list of all addresses that had responded to the ACS within the prior two months in three different geographic areas. Recruitment primarily focused on mode of response to get sufficient participation from each group. The demographic profile of respondents is shown in Table 2.

Table 1. Number of focus group respondents by mode.

Mode	Number of Respondents
Self-response	
Paper	9
Internet	17
Interviewer-administered response	
CATI	7
CAPI	13
Total	**46**

Table 2. Respondent demogra phic characteristics by mode

Respondent Characteristic	Self-response (n = 26)	Interviewer-administered response (n = 20)	Total (n = 46)
Household Size			
1	6	11	17
2	9	3	12
3	7	2	9
4+	4	4	8
Sex			
Male	16	9	25
Female	10	11	21
Age*			
18–29	5	1	6
30–44	4	7	11
45–60	8	8	16
Over 60	9	3	12
Race*			
White	16	4	20
Black or African American	8	14	22
Other	2	1	3
Education			
High school degree or less	3	6	9
Some college, no degree	3	4	7
Associate's degree (AA/AS)	3	2	5
Bachelor's degree (BA/BS)	9	5	14
Post-bachelor' s degree	8	3	11

*There is one response missing from each of these characteristics for the interviewer -administered groups.

Across all focus groups, respondents were fairly balanced on characteristics such as sex, race (white versus nonwhite), and education (less than Bachelor's degree versus Bachelor's or higher). Most of our respondents lived in one to two person households. The demographics of respondents also differed between the self-response (SR) and interviewer-administered (IA) groups. Compared to IA respondents, SR respondents were older and had higher levels of education. More SR respondents were white, male, and lived in three to four person households. Response mode differences for race, education, and age align with those typically observed in the ACS (Joshipura 2008; Nichols et al. 2015).

3.3. Protocol

Upon arrival to the focus groups, respondents completed a short handout with demographic questions and a consent form. Respondents were told that information they provided would be confidential, that the group would be video- and audio-recorded, and that observers were present (when applicable). After respondents signed the consent forms, focus group moderators introduced the topic, set ground rules for the discussion, and facilitated respondent introductions. Respondents were told that these focus groups were being conducted to learn more about people's survey experiences.

After a discussion of their general survey experiences to get a sense of whether respondents felt over-surveyed, respondents were asked what they recalled about their experience with the ACS. We started with open-ended questions so that we could learn what features of the ACS were most salient to respondents and therefore might contribute most to their perceptions of burden. After respondents were given enough time to spontaneously recall their experience, we asked probes from the moderator's guide as needed about what they remembered about the ACS. Probes were focused on the survey and respondent characteristics identified by Bradburn (1978) and the broader literature as well as ACS-specific features identified by researchers. Moderators used more probes in focus groups where spontaneous recall was low. Probes within each section of the moderator's guide started general and then grew increasingly specific.

After probing respondents on their ACS experience, we asked them to respond to a written survey question about the burden of the ACS that was adapted from the CE: "How burdensome was the ACS to you?" (U.S. Bureau of Labor Statistics 2017). We asked respondents to discuss how they came up with their answers and what "burdensome" means to them. Respondents also answered and discussed 12 additional questions, most of which were adapted from Sharp and Frankel (1983) and the 2012/2013 Consumer Expenditure Survey (Yang 2015). Questions were on topics such as the length of the ACS, question difficulty and sensitivity, and broader knowledge and attitudinal topics such as the importance of the ACS, see Figure 1 in the Appendix (Section 6). The questions were used to encourage further discussion on earlier topics; the focus groups were not used to cognitively test these questions. To conclude the groups, we asked respondents if there were any changes that could make the ACS less burdensome.

3.4. Analysis

Following data collection, focus group recordings were transcribed except for the introduction, off-topic comments, and moderator probes, which were generally summarized rather than transcribed. Personally identifiable information such as names were not included in the transcriptions; respondents were referred to using numbers instead. Transcripts were reviewed individually by the two focus group moderators as well as three additional researchers for recurring themes and patterns across focus groups and by focus group type (SR versus IA response; Willis 2015). We used the topics of the moderator's guide as a preliminary list of themes and added themes as necessary. Comments that pertained to more than one theme were included in both themes. We also analyzed the transcripts for evidence of respondent or household-level characteristics that may have influenced responses. We then created a summary document with consensus

findings organized by sections of the focus group moderator's guide that compiled these results.

4. Results

In this section, we discuss common themes that emerged across the focus groups. We also describe notable differences between focus groups and by response mode and respondent characteristics. Overall, focus group respondents mostly commented on aspects of the ACS questionnaire, including the response mode, question topics, the length of the survey, difficulty, and sensitivity. Respondents also made some comments about the contact attempts they received, though this received less attention than the questionnaire did. While respondents reported that they did not remember the messaging of these contacts, some respondents still recalled that response was mandatory and some details about how ACS data would be used, which may have been gleaned from the mailings, or from the interviewer for those who responded by CATI or CAPI. Some respondents also commented that ACS data is important and discussed their knowledge of and attitudes and beliefs towards the ACS and the U.S. Census Bureau. Overall, focus group respondents did not find the ACS to be very burdensome. When asked how they decided whether the ACS was burdensome, respondents tended to think about the process of responding to the survey itself (e.g., time, effort, question topics), the value of the data, that response is mandatory, and to a lesser extent, the number of contacts they received.

In the following sections, we discuss the survey characteristics that were most salient to respondents as well as respondents' knowledge, attitudes, and beliefs. We then discuss respondents' reactions to the overall perceived burden of the ACS and responses to probes about which of these features contributed to their perceptions of burden. For simplicity, results are grouped by topic because most of the comments made across sections of the moderator's guide were similar. Findings distinguished by section of the moderator's guide can be found in Holzberg et al. (2021). Throughout these sections, we also embed findings related to respondents' demographic and household characteristics and response mode where applicable.

4.1. Questionnaire

4.1.1. Response Mode and Question Topics

Focus group respondents were able to correctly recall their mode of response to the ACS unprompted. A few respondents said they appreciated being given the option to respond by mail; one of these respondents did not have internet access. While some respondents responded early in the field period, a few respondents talked about starting the ACS online or on paper and coming back to it later, often after receiving a reminder:

> *"I remember getting it in the mail and setting it aside and then filling it out when I got nudged."*

Most of the respondents in the IA response focus groups indicated that they liked responding with an interviewer, in some cases finding it preferable to SR because of the personal interaction and because they thought it was easier. A few of these respondents

discussed starting the ACS online or on paper and then neglecting to finish, with one respondent saying she thought if she ignored the mail she received, the ACS would "go away." A few IA focus group respondents indicated their response was reluctant:

> *"I got caught off guard because I didn't know what I was about to get myself into. Had I known, I would not have taken that phone call. But once I was there, I didn't want to be rude and say, 'hey, I have to go."*

The topic of the questions on the ACS came up in about half of the focus groups unprompted. In a few groups, respondents described the ACS as asking basic questions "about the household" or "just basic facts that they need to collect." One respondent remembered some questions included a reference period (e.g., "in the past 12 months") and said the questions were specific. Some of the questions recalled by respondents included name, age, race, sex, property value, health insurance, education, and income. A few respondents incorrectly recalled being asked about topics such as square footage. Respondents were not always sure why certain questions were being asked; for example, one respondent thought the utilities questions were particularly strange. A respondent in an IA response group said that not all the questions applied to her but "they had to ask (them) anyway."

4.1.2. Length

At least a few respondents in most of the groups said unprompted that they thought the ACS was too long and/or too repetitive. Some respondents were unable to provide examples of how it was repetitive, while others commented on answering questions about each household member. On the other hand, a few respondents, mostly in IA response focus groups, said the experience was quick. These respondents said that their response seemed shorter than self-responding on the paper form would have been:

> *"It's quicker on the phone... if you take the form, it's going to take longer because of the reading and writing."*

When we asked respondents whether the length of the ACS was too long, too short, or just right, about half of the respondents in the SR focus groups said it was too long and about half said it was about right. Most IA response focus group respondents said the length was about right. This may be because IA focus group respondents tended to think the ACS took less time than those in the SR groups and/or because more SR focus group respondents lived in larger households. Respondents who thought the length was about right said it seemed appropriate given the importance of the ACS and that this is a survey for the U.S. Census Bureau. Other than mode, perceptions of length seemed to be influenced by having a large household, having trouble answering the questions, being unsure about the purpose of some of the questions, and needing to look up information or get help from someone else:

> *"I think a lot depended [on] if you had three [people] in your household versus seven in your household."*

Only a few respondents mentioned being told in advance how long it would take. When we asked probing questions about the length of the ACS, respondents indicated they were

not sure how long it took and were guessing. Responses varied widely from ten to 90 minutes, though most said it took between 20 and 45 minutes.

4.1.3. Difficulty and Effort

Most respondents in both the SR and IA response groups spontaneously mentioned that they thought the ACS was "easy," "straightforward," "basic," and "simple" as a whole, though some found individual questions to be confusing. In one IA response group, respondents said the interviewers were able to explain any questions that confused them. A few people said that while the questions were not difficult, they sometimes made them think about why they were being asked. A few people also talked about trying not to estimate. In a few groups, respondents mentioned needing to look up information or consulting with a household member to respond; one respondent looked up her mortgage and another said she looked up her income. A few respondents mentioned that responding on the behalf of other household members required them to speak to other household members to get their information:

> *"The questions were fairly easy; they weren't difficult at all. But:.. had to keep, you know, asking....yelling back in the room, 'hey, where do you work? How many hours do you work? What do you do?"*

Most of the respondents making these comments lived in more complex households, including arrangements such as having stepchildren, adult children who sometimes live with the respondent, and roommates. One respondent referred to this experience as "frustrating but also interesting." Some other respondents speculated that their experience may not have been as easy or pleasant if they lived in larger households:

> *"I could see [how] if you're [not in] a nuclear family... that it might be difficult."*

We also asked respondents how much effort they put into responding to the ACS. Most respondents across groups said they put in a moderate amount of effort into responding to the ACS, as opposed to a lot of effort, a little effort, or no effort. Overwhelmingly, respondents said that answering the ACS required at least a little effort because they tried to be precise in their responses. A few respondents said that responding took a little bit of time and thus effort; others said that it took some effort for them to get started but that once they did, the ACS was not that bad. Respondents also answered a question about whether the time and effort responding to the ACS was well-spent. There were not any respondents who said their time and effort on the ACS was not at all well-spent. However, responses to this question were split across the other categories of this question (a little well spent, somewhat well spent, and very well spent). Respondents who said a little or somewhat well spent tended to do so because they were unsure of the ultimate data use or purpose of the survey. Others were more optimistic about the uses of the data and therefore thought their time and effort was well spent.

4.1.4. Sensitivity

Respondents in about half of the groups felt that at least some of the questions on the ACS were sensitive, most frequently describing it as "personal." One respondent joked, "You might as well have a chip in me when this is over." Another respondent said he only

answered the questions because it was for the U.S. Census Bureau; he would not answer them for anyone else. Income was mentioned a few times as a sensitive question; however, most respondents who found the ACS sensitive did not cite any specific questions as particularly sensitive. Respondents who found the ACS sensitive cited lack of knowledge about data use, concerns with certain questions such as income, and proxy response on behalf of household members as making response more sensitive. Most respondents who found the ACS sensitive were the same ones who said they found response difficult because of their complex households:

"I felt kind of awkward answering questions about them because... I didn't know how it was going to be used. I just felt like it wasn't my information to give, I guess."

One of the respondents with roommates said it was sensitive because her roommates were not U.S. citizens. Another respondent said that she and her roommates each individually logged into the survey online to answer about themselves:

"At first I started filling it out and then when it started asking these details about them, I was like, 'I don't know.' I basically had to then coordinate and be like, 'here's the PIN guys...' We had to do these separate logins... we know each other well enough, but I was kind of thinking... we could have scrolled through and seen [how much money] we made..."

People who did not find the ACS sensitive tended to either say the questions were not personal in any context, or that questions they would usually consider to be too personal were okay on a U.S. Census Bureau survey. A few respondents mentioned that they knew the data would be aggregated which made it less sensitive.

4.2. Contact Strategy

4.2.1. Number and Mode of Contact Attempts

Most respondents said they felt like the number of contacts they received was reasonable, commenting they did not feel "flooded" or "bombarded." Only a few respondents said that the number of times they were contacted was a lot or felt excessive. Most respondents remembered receiving contacts by mail and a few remembered receiving phone calls; however, this varied by focus group type with more SR respondents recalling mail contacts. Some IA response respondents remembered only one piece of mail or did not remember receiving any mail. This was somewhat surprising, since the U.S. Census Bureau sends addresses five mailings before an interviewer attempts to contact someone living at that address. It is possible that these respondents were not paying attention to their mail or that someone else in their household saw the mail instead.

Some respondents across groups were able to provide a general description of the types of mail they received when we asked. Among those who recalled receiving mail across the groups, some respondents commented that they received reminder mailings because they had not yet responded, and some said reminders were helpful. One respondent said the timing of a reminder surprised them.

"It was just the right amount for me because I kept forgetting."
"I got letters saying to do it online, and then three days later I got a letter that said, 'You

haven't filled it out yet,' and I was like 'Holy cow! It's only been three days! It's not even the weekend yet!' (laughter)"

Other than those who responded by mail, most respondents did not recall receiving a paper questionnaire, including IA respondents who should have received a paper questionnaire before being contacted by an interviewer. Respondents in IA response groups mentioned having interactions with interviewers by phone, in person, or both. Some respondents recalled receiving a notice of visit from an interviewer or a missed phone call, while others only recalled a single contact attempt in which they also responded to the ACS. A few respondents said they were surprised to be contacted on the phone or in person, with one respondent who did not recall receiving mail even suggesting that they be warned or notified somehow ahead of time.

4.2.2. Messaging of Contact Attempts

Only a few respondents in either the SR or IA response groups made explicit comments about the messaging of the contact attempts they received. In about half of the groups, respondents made comments about response to the ACS being mandatory. When we specifically asked respondents about this, most respondents recalled that it was mandatory though some respondents, mostly in IA response focus groups, thought response was voluntary. Interviewers are instructed to gain cooperation with respondents by using methods other than pushing that response is mandatory, so respondents may have been unaware of it. The reaction to the ACS being mandatory was mixed, with some respondents disliking it and others saying it communicated to them that the survey must be important. A few people said they would have still done it if it was not mandatory.

Only a few respondents specifically mentioned other types of messaging in the contact attempts. One respondent admitted he did not read most of the material. No one said they could recall messaging in contact materials about how the data would be used or why the ACS is conducted, though comments made by some respondents on these topics during other parts of the focus groups indicated that perhaps some had retained this messaging and simply forgot where they had learned this information (see Subsection 4.3). A few respondents recalled messaging about how they would continue to be contacted until they participated, with one respondent in a SR group saying they were motivated to complete the ACS so an interviewer would not contact them:

"(It said) you better finish your survey in like five days or we're going to come visit ya."

4.3. Knowledge, Attitudes, and Beliefs

Overall, respondents in nearly all groups spontaneously articulated a vague sense that the ACS was useful and valuable, and a few respondents were knowledgeable about the survey. At the end of the groups, we also asked respondents additional questions about their knowledge, beliefs, and attitudes towards the ACS. When asked about their overall impressions of the survey, respondents across groups said that the ACS is used to plan for the future, designate funds for community resources like schools and hospitals, create population statistics and "reports" or graphs, and monitor changes in the population over

time. One respondent referred to the ACS as "statistics jury duty," and others similarly said responding was a civic duty:

"It will help the city and the town, and I think that's very important."
"It made me feel good that I was able to contribute to something like that."

Other respondents stated generic uses such as that the government would "review it." However, respondents in about half of the groups said that they did not exactly know how the data would be used or expressed some sort of skepticism that it would be used for its intended purpose. These respondents who were unsure about how the data are used had trouble deciding how important they thought the survey was.

"I know it's for data, statistics but... I'm more concerned about the outcome of the demographic research. I feel like the questions they were asking, there was legitimacy, but... it would be nice to know... well, in answering this, these things will be affected."
"I hope they use this for good analysis (and don't) sell it off for marketing out to the public."
"As far as what effect this has on us, I have no idea."
"I don't think much will be done in the community."

At least one respondent in nearly all the focus groups conflated the ACS with the U.S. decennial census, such as by commenting that the ACS is administered every ten years or that ACS data are used for congressional redistricting. There was no clear evidence of conflation with the census differing between the SR and IA response groups. Respondents also asked questions such as how ACS data are kept secure, why ACS information could not be collected from other records, and how ACS respondents are selected.

Most respondents said that they trust the U.S. Census Bureau to safeguard their information, though there were more respondents in the SR focus groups than in the IA response focus groups who said they neither trusted nor distrusted the U.S. Census Bureau. Across groups, most respondents who were not very trusting said it was because everything today is hackable, including U.S. Census Bureau data.

"I don't know where it could be floating around. Look what happens."

Other respondents said they thought the information might otherwise be leaked or misused. Some respondents were considering their level of trust in the government more broadly. In a few groups, respondents discussed the potential for data sharing between government agencies. In one of these groups, respondents assumed their data was shared but did not mind because they had "nothing to hide"; in another group, respondents strongly trusted the U.S. Census Bureau because they said they knew their individual information could not be shared.

Respondents' answers to a question about how interesting they found the ACS to be varied, but most said the ACS was a little or somewhat interesting. People who thought the ACS was interesting often said it was because the ACS is important, and the data are used widely. On the other hand, some respondents said the ACS was interesting because they did *not* understand how the data are used; therefore, they were curious about the purpose of the survey. Similarly, a few respondents made comments about feeling curious about the rationale behind specific questions. Respondents who did not find the ACS interesting said

this was because it was something they just had to do. One respondent said he did not see how the ACS would be interesting to anyone.

We also asked respondents about their attitudes and experiences with surveys other than the ACS to get a sense of whether they felt over-surveyed. Most focus group respondents had taken surveys other than the ACS and most said they felt like they are asked to participate in surveys "constantly," particularly customer satisfaction surveys and political polls. When asked how they decide whether to respond to these survey requests, respondents mentioned considering what good would come out of their participation, the topic, sponsor, and length. Throughout the focus groups, respondents clearly thought of the ACS as very different from other surveys they are often asked to do. The mandatory response requirement, topic, and sponsor were all distinguishing factors.

4.4. Overall Perceived Burden

After hearing about respondents' experiences with the ACS and learning about their knowledge, attitudes, and beliefs, we administered a survey question assessing overall perceptions of burden. We asked respondents to discuss how they came up with their answers and what the word "burdensome" means to them. None of our focus group respondents said the ACS was very burdensome, and only seven said it was "somewhat burdensome" (see Table 3). More IA response focus group respondents evaluated the ACS as "not at all burdensome."

When we asked respondents how they came up with their answer to this question, most focus group respondents commented that the ACS was "not a big deal" and "not a huge imposition" on their lives. One respondent asked if we were conducting this research because something had "gone wrong" with the ACS.

"Why would they even ask this question? It's not a burden."

Respondents considered numerous factors when deciding how to answer. However, the most common responses were related to how respondents felt about responding to the survey itself, especially the time needed to complete it and complications associated with proxy response. Other factors mentioned related to the questionnaire, including the level of effort required to complete it, difficulty and sensitivity of the questions, and needing to check records. A few respondents said the fact that the ACS is not something they are asked to do frequently made it less burdensome. Some respondents commented on weighing the fact that response to the ACS is mandatory against the fact that the data are valuable. Interestingly, some respondents who thought that the survey was too long or required a moderate amount of effort said that the ACS was not at all or only a little

Table 3. Respondents' perceptions of ACS burden: "Hoe burdensome was the ACS to you?"

Response option	Self-response (n = 26)	Interviewer-administered response (n = 20)
Very burdensome	0	0
Somewhat burdensome	5	2
A little burdensome	12	5
Not at all burdensome	9	13

burdensome. It appeared that perceptions about the importance of the survey and a focus on the survey being a one-time request outweighed negative perceptions of length and effort when making this determination. Slightly less often, respondents commented that they considered how they felt about the contacts they received such as mail, and in the case of the IA focus groups, interactions with interviewers. These comments were consistent across groups, except for proxy response and response being mandatory, which came up more frequently in SR focus groups. Since these survey features were also discussed extensively earlier in the focus groups, this suggested that these might be the most appropriate topics on which to focus burden questions for the ACS.

When asked what being "burdensome" or a "burden" means to them, focus group respondents most frequently thought of it as a nagging responsibility or chore that they need to complete:

"A thorn in your side."
"It's another thing on my plate that I have to do."

Most respondents also said that a burden is something that requires a lot of time and/or effort, with some respondents saying a task becomes especially burdensome if it interrupts your usual routine. Other common synonyms were annoying, a hassle, inconvenient, and bothersome. A few respondents who were not native English speakers were not sure what "burdensome" meant. A few respondents also said that "burdensome" did not feel like the right word to use to describe the survey:

"Took a little time for me to do it, but I wouldn't have exactly called it burdensome. That seems like a strong word to use. Inconvenient? Sure. But burdensome almost seems like an extreme word to use for filling out a survey."
"It wasn't a burden to me... it's not something you have to do every week, every month. It's something that needs to be done so it wasn't a burden to me at all."
"It didn't hurt me or cause me any conflict."

When we asked respondents how the ACS could be less burdensome, they did not have much to offer in terms of suggestions. A few respondents across groups suggested that the ACS should better communicate how the data are used. Other less frequent suggestions included making questions easier and less sensitive, asking fewer questions, giving a deadline for responding, and explaining the penalty for nonresponse. A few respondents in IA response focus groups said they thought ACS respondents should be able to choose their response mode. Some of these comments illustrated respondents' lack of memory for the mailing materials they should have received.

5. Discussion

The findings of our focus groups support the idea that burden is a complex, multidimensional concept likely to be affected by both survey and respondent characteristics (Bradburn 1978). Respondents were able to give an assessment of their overall perceived burden of the survey, but they tended to think of different aspects of the ACS when assessing the overall burden. Most thought about the process of responding to the survey itself (e.g., length, effort, sensitivity), the value of the data, that response is

mandatory, and to a lesser extent, the contact attempts they received. This suggested that these are the most important aspects of burden on which to ask questions in the ACS, though the importance of each aspect varied by the individual respondent. Some perceptions varied by response mode or respondent characteristics such as household size and composition.

Some of the themes that emerged during these focus groups aligned with those suggested by Bradburn (1978) and found in qualitative burden research with establishment survey respondents, such as the process of responding to the survey itself and the value of the data. However, new themes also emerged, and this research was a good opportunity to delve into unique features of the ACS and understand respondents' attitudes. For example, it was useful to learn that respondents did not think the ACS was like other surveys and therefore effects of over-surveying did not seem as relevant as they might be for other surveys. We also were able to confirm that the mandatory nature of this household survey was salient to respondents and learned that proxy response may be an important understudied component of effort and burden.

Our findings reinforce the importance of capturing respondents' perceptions, as perceived burden did not always align with objective burden. Using only objective measures of burden (e.g., number of contacts) may lead researchers to focus their efforts on changing or improving survey characteristics that may not actually lessen burden from the respondents' perspective. In our research, some focus group respondents who received a small number of contacts from the U.S. Census Bureau said they received too many contacts while others who received more did not say they received too many. We recommend a qualitative research approach be taken before adding subjective burden questions to help identify burden themes relevant to an individual survey.

5.1. Limitations

We note that prior respondents who were amenable to participating in a focus group may differ from those who did not want to participate in their experiences with the ACS or in their knowledge, attitudes, and interest about the ACS and the U.S. Census Bureau. We were also unable to learn about perceptions of burden from people who chose not to respond to the ACS. While we tried to prioritize respondents who had completed the ACS more recently during recruitment, some respondents still had difficulty recalling their experience with the ACS because the time between survey completion and research participation was at least one month. However, there was no clear evidence of recall differing between the SR and IA response groups. There was also a gap between when respondents completed the questionnaire and when they answered the burden questions. If burden questions were to be administered on the ACS, they would be asked immediately at the end of the questionnaire. It is possible that respondents' perceptions may differ when they are given an opportunity to provide feedback with their experience fresh in their minds. Finally, the impact of some survey and respondent characteristics on perceptions of burden may be specific to the ACS. We expect our findings would be most applicable to other household surveys sponsored by the U.S. government of similar length, effort, and sensitivity, but we emphasize that it is important to conduct qualitative research before adding subjective burden questions.

5.2. Future Research

While this research was a useful first step to learn about the most salient aspects of burden for a specific survey, identifying a short series of questions to capture the most important perceptions of burden is difficult because burden is multidimensional. Further qualitative testing and field testing can help narrow down options identified through initial qualitative research. In addition to examining the response distributions to the burden questions themselves, looking at the correlations between burden items, using them as predictors in models, and examining whether responses to burden questions differ by respondent characteristics would be useful. This type of work is being conducted on surveys such as the CE.

It is important to note that qualitative and quantitative research on subjective burden, including the present study, is typically conducted with people who responded to a survey request. Therefore, we do not know much about where the perceptions of burden may be highest—those who never respond at all (Hedlin et al. 2005). In 2018, U.S. Census Bureau staff conducted an analysis of complaints received about the ACS (Fox et al. 2018). While this analysis serves as a useful starting point and similar analysis can be conducted with other surveys, these complaints still only reflect the perceptions of people who proactively reached out to provide their feedback. Future qualitative research on burden should prioritize nonrespondents, perhaps by leveraging input from interviewers for CATI and CAPI modes.

Reactions to the term "burdensome" were not universally positive in this study, raising some concerns about the appropriateness of this widely used burden measure that are worth further examining. Though respondents were still able to understand and answer the question using this word and thought about different aspects of the survey while doing so, some respondents thought it was an odd or severe term to describe a survey experience. A clearly superior alternate term for a question about overall perceived burden was not identified in these focus groups or subsequent cognitive interviews (see Holzberg et al. 2021). Other household surveys using an overall perceived burden question with the word "burdensome" have still found meaningful differentiation in response options (e.g., Dahlhamer et al. 2019; Yan et al. 2020), and a more severe word may be more useful for identifying respondents whose burden perceptions are particularly strong. However, future research should continue to explore alternate terms that may be more appropriate, especially as the word "burdensome" was also not well understood by a few respondents who were not native English speakers. To our knowledge, no U.S. household survey has conducted qualitative research about subjective burden questions in languages other than English. Cognitive testing in languages such as Spanish would provide better insight into potential issues in other languages. While a benefit of the overall perceived burden question is a reduction in the number of subjective burden questions asked, an alternative solution to a question with the word "burdensome" may be to ask a series of questions teasing out individual effects of respondent and survey characteristics. We recommend researchers consider a multi-question approach in future studies of burden.

Overall, research into subjective perceptions of burden on household surveys is still in its relative infancy, with few surveys in the United States regularly asking questions on this topic. There is generally a need for more qualitative research to ensure questions capture important respondent perceptions for a particular survey and are understood as intended.

6. Appendix

Set 1

How burdensome was the ACS to you?
- o Very burdensome
- o Somewhat burdensome
- o A little burdensome
- o Not at all burdensome

Set 2

1. Do you feel that the length of the American Community Survey (ACS) was too long, about right, or too short?
   - o Too long
   - o About right
   - o Too short

2. How interesting was the ACS to you?
   - o Very interesting
   - o Somewhat interesting
   - o A little interesting
   - o Not at all interesting

3. How difficult or easy was it for you to answer the questions in the ACS?
   - o Very easy
   - o Somewhat easy
   - o Somewhat difficult
   - o Very difficult

4. Thinking about the contacts you received for the ACS, would you say it was too many, a reasonable number, or not enough?
   - o Too many
   - o A reasonable number
   - o Not enough

5. How sensitive did you feel the questions on the ACS were?
   - o Very sensitive
   - o Somewhat sensitive
   - o A little sensitive
   - o Not at all sensitive

6. How agreeable would you be to take another survey like the ACS in the future?
   - o Very agreeable
   - o Somewhat agreeable
   - o Somewhat disagreeable
   - o Not at all agreeable

7. If we had to extend the ACS for another 15 minutes, how willing would you have been to continue?
   - o Very willing
   - o Somewhat willing
   - o Somewhat unwilling
   - o Very unwilling

8. Thinking about the amount of effort that you put forth into answering the ACS, would you say that you put forth:
   - o A lot of effort
   - o A moderate amount of effort
   - o A little effort
   - o No effort

9. Would you say that the time and effort you put into answering the ACS survey questions was:
   - o Very well spent
   - o Somewhat well spent
   - o A little well spent
   - o Not at all well spent

10. How motivated were you to complete the ACS as quickly as possible? Would you say that you were:
    - o Very motivated
    - o Somewhat motivated
    - o Somewhat unmotivated
    - o Very unmotivated

11. How much do you trust the U.S. Census Bureau to safeguard the information you have provided them on the ACS?
    - o Strongly trust
    - o Somewhat trust
    - o Neither trust nor distrust
    - o Somewhat distrust
    - o Strongly distrust

12. How important do you think it is for the U.S. Census Bureau to collect the information on the ACS?
    - o Very important
    - o Somewhat important
    - o Neither important nor unimportant
    - o Somewhat unimportant
    - o Very unimportant

Fig. 1. Respondent burden survery questions.

7. References

Bavdaž, M. 2010. "The multidimensional integral business survey response model." *Survey Methodology* 36(1): 81–93.

Bogen, K. 1996. "The Effect of Questionnaire Length on Response Rates—A Review of the Literature." In *Proceedings of the Survey Research Methods Section*, May 16–19,

American Statistical Association, 1020-5. Alexandria, VA, USA. Available at: http://www.asasrms.org/Proceedings/papers/1996_177.pdf (accessed November 2021).

Bradburn, N. 1978. "Respondent burden." In *Proceedings of the Survey Research Methods Section:* American Statistical Association: 35–40. Alexandria, VA, USA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf (accessed November 2021).

Crawford, S.D., M. Couper, and M.J. Lamias. 2001. "Web Surveys: Perceptions of Burden." *Social Science Computer Review* 19(2): 146–162. DOI: https://doi.org/10.1177/089443930101900202.

Dahlhamer, J., A. Maitland, B. Zablotsky, and A. Galinsky. 2019. "Does Reducing the Length Matter? An Assessment of Respondent Burden with the Redesigned U.S. National Health Interview Survey." *Presentation at the DC-AAPOR Respondent Burden Workshop*, October 21, Washington D.C., USA. Available at: http://files.dc-aapor.org/slides/burden2019/Galinsky.pdf (accessed November 2021).

Dale, T., J. Erikson, J. Fosen, G. Haraldsen, J. Jones, and O. Kleven. 2007. *Handbook for Monitoring and Evaluating Business Survey Response Burdens.* Technical Report, European Commission, Eurostat. Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/12-HANDBOOK-FOR-MONITORING-AND-EVALUATING-BUSINESS-SURVEY-RESONSE-BURDEN.pdf/600e3c6d-8e8d-44f7-a8f5-0931c71d9920 (accessed November 2021).

Earp, M.S., B. Kopp, and J. Dixon. 2019. "Exploring the Relationship between Burden Factors and Survey Response." *Presentation at the DC-AAPOR Respondent Burden Workshop*, October 21, Washington, DC, USA. Available at: http://files.dc-aapor.org/slides/burden2019/Earp.pdf (accessed November 2021).

Fox, M., K. Johns, A. Kline, D. Raglin, and R.C. Sawyer. 2018. *Analysis of Complaints Received.* Technical Report, U.S. Census Bureau 2018 American Community Survey Research and Evaluation Report Memorandum Series #ACS18-RER-08. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2018/acs/2018_Fox_01.pdf (accessed November 2021).

Fricker, S. 2016. "Defining, Measuring, and Mitigating Respondent Burden." *Presentation at the Workshop on Respondent Burden in the American Community Survey*, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, March 8, Washington, DC, USA. Available at: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_173169.pdf (accessed November 2021).

Fricker, S., J. Gonzalez, and L. Tan. 2011. "Are you burdened? Let's find out." *Presentation at the Annual Conference of the American Association for Public Opinion Research*, May 14, Phoenix, AZ, USA. Available at: https://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2011/05-14-11_3B_Fricker.pdf (accessed November 2021).

Fricker, S., C. Kreisler, and L. Tan. 2012. "An exploration of the application of PLS path modeling approach to creating a summary index of respondent burden." In *Proceedings of the Survey Research Methods Section of the American Statistical Association:* 4141–4155, July 30, Alexandria, VA, USA. Available at: http://www.asasrms.org/Proceedings/y2012f.html (accessed November 2021).

Fricker, S., T. Yan, and S. Tsai. 2014. "Response Burden: What Predicts It and Who is Burdened Out?" In *Proceedings of the Survey Research Methods Section of the American Statistical Association:* 4568–4577, May 16, Alexandria, VA, USA. Available at: http://www.asasrms.org/Proceedings/y2014f.html (accessed November 2021).

Galesic, M. 2006. "Dropouts on the Web: Effects of Interest and Burden Experiences During an Online Survey." *Journal of Official Statistics* 22 (2): 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf.

Giesen, D., M. Bavdaž, and G. Haraldsen. 2011. "Response Burden Measurement: Current Diversity and Proposal for Moving toward Standardization." In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, March 22-23, Heerlen, Netherlands: 125–134. Available at: https://www.cbs.nl/-/media/imported/documents/2011/25/2011-06-20-giesen-bavdaz-haraldsen-2011-response-burden-measurement.pdf (accessed November 2021).

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Surveys.* New York: Wiley and Sons.

Groves, R.M., E. Singer, and A. Bowers. 1999. "A Laboratory Approach to Measuring the Effects on Survey Participation of Interview Length, Incentives, Differential Incentives, and Refusal Conversion." *Journal of Official Statistics* 15(2): 251–268. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-laboratory-approach-to-measuring-the-effects-on-survey-participation-of-interview-length-incentives-differential-incentives-and-refusal-conversion.pdf.

Haraldsen, G. 2004. "Identifying and Reducing Response Burdens in Internet Business Surveys." *Journal of Official Statistics* 20 (2): 393–410. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/identifying-and-reducing-response-burdens-in-internet-business-surveys.pdf.

Hedlin, D., T. Dale, G. Haraldsen, and J. Jones. 2005. *Developing Methods for Assessing Perceived Response Burden*. Technical Report, European Commission, Eurostat. Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/10-DEVELOPING-METHODS-FOR-ASSESSING-PERCEIVED-RESPONSE-BURDEN.pdf/1900efc8-1a07-4482-b3c9-be88ee71df3b (accessed November 2021).

Hedlin, D., H. Lindkvist, H. Backstrom, and J. Erikson. 2008. "An Experiment on Perceived Survey Response Burden among Businesses." *Journal of Official Statistics* 24 (2): 301–318. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-experiment-on-perceived-survey-response-burden-among-businesses.pdf.

Holzberg, J., J. Katz, and M. Davis. 2021. *Measuring Respondents' Perceptions of Burden in the American Community Survey (ACS)*. Technical Report, U.S. Census Bureau Research Report Series Survey Methodology: 2021(4). Available at: https://www2.census.gov/adrm/CBSM/rsm2021-04.pdf (accessed November 2021).

Hotchkiss, M. and J. Phelan. 2017. *Uses of Census Bureau Data in Federal Funds Distribution: A New Design for the 21St* Century. Technical Report, U.S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/working-papers/Uses-of-Census-Bureau-Data-in-Federal-Funds-Distribution.pdf (accessed November 2021).

Hughes, T., E. Slud, R. Ashmead, and R. Walsh. 2016. *Results of a Field Pilot to Reduce Respondent Contact Burden in the American Community Survey's Computer Assisted Personal Interviewing Operation.* Technical Report, U.S. Census Bureau 2016 American Community Survey Research and Evaluation Report Memorandum Series #ACS16-RER-07. Available at: https://www.census.gov/content/dam/Census/library/-working-papers/2016/acs/2016_Hughes_01.pdf (accessed November 2021).

Jones, C., P. Sheatsley, and A. Stinchcombe. 1979. *Dakota Farmers and Ranchers Evaluate Crop and Livestock Surveys.* Technical Report, National Opinion Research Center Report No. 128. Chicago. Available at: https://www.norc.org/PDFs/publications/NORCRpt_128.pdf (accessed November 2021).

Jones, J. 2012. "Response Burden: Introductory Overview Lecture." *Presentation at the Fourth Annual International Conference on Establishment Surveys*, June 12, Montreal, Quebec, Canada. Available at: https://ww2.amstat.org/meetings/ices/2012/papers/302289.pdf (accessed November 2021).

Joshipura, M. 2008. *2005 American Community Survey Respondent Characteristics Evaluation*. Technical Report, U.S. Census Bureau DSSD American Community Survey Research and Evaluation Memorandum Series, Chapter #ACS-RE-2. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2008/acs/2008_Joshipura_01.pdf (accessed November 2021).

Kaplan, R. and J. Holzberg. 2019. "What Does it Mean to be Burdened? Exploring Subjective Perceptions of Burden." *Presentation at the DC-AAPOR Respondent Burden Workshop*, October 21, Washington D.C., USA. Available at: http://files.dc-aapor.org/slides/burden2019/Kaplan.pdf (accessed November 2021).

Kaplan, R. and S. Fricker. 2017. *Objective and Subjective Burden Measures: Which Survey Features and Respondent Characteristics Contribute to Both?* Paper presented at the Annual Conference of the American Association for Public Opinion Research, May 20, New Orleans, LA, USA.

Kim, Y., J. Dykema, J. Stevenson, P. Black, and D.P. Moberg. 2019. "Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail-Web Mixed-Mode Surveys." *Social Science Computer Review* 37: 1–20. DOI: https://doi.org/10.1177/0894439317752406.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. DOI: https://doi.org/10.1002/acp. 2350050305.

Kunz, T. and T. Gummer. 2019. "Perceived and Actual Respondent Burden and the Effects on Data Quality in Web Surveys." *Presentation at the European Survey Research Conference*, July 17, Zagreb, Croatia. Available at: https://www.europeansurveyresearch.org/conferences/programme?sess=203#573.

Lee, Y.A. and C. Knappenberger. 2019. "Exploring monetary incentives and respondent burden in the Consumer Expenditure Interview Survey." *Presentation at the DC-AAPOR Respondent Burden Workshop*, October 21, Washington D.C., USA. Available online at: http://files.dc-aapor.org/slides/burden2019/Lee.pdf (accessed November 2021).

Leeper, T.J. 2019. "Where Have the Respondents Gone? Perhaps We Ate Them All." *Public Opinion Quarterly* 83: 280–288. DOI: https://doi.org/10.1093/poq/nfz010.

Malhotra, N. 2008. "Completion Time and Response Order Effects in Web Surveys." *Public Opinion Quarterly* 72: 914–934. DOI: https://doi.org/10.1093/poq/nfn050.

McCarthy, J.S., D.G. Beckler, and S.M. Qualey. 2006. "An Analysis of the Relationship Between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative?" *Journal of Official Statistics* 22 (1): 97–112. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-analysis-of-the-relationship-between-survey-burden-and-nonresponse-if-we-bother-them-more-are-they-less-cooperative.pdf.

McCarthy, J.S. and D.G. Beckler. 2000. "Survey Burden and its Impact on Attitudes toward the Survey Sponsor." *Paper presented at the Second Annual International Conference on Establishment Surveys*, June 17-21, Buffalo, NY, USA. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/allreports/Survey_Burden_and_Its_Impact.pdf.

Morgan, D.L. 1998. *The Focus Group Guidebook.* Thousand Oaks, California: SAGE Publications.

Nichols, E., R. Horwitz, and J.G. Tancreto. 2015. *An Examination of Self-Response for Hard-to-Interview Groups When Offered an Internet Reporting Option for the American Community Survey.* Technical Report, U.S. Census Bureau 2015 American Community Survey Research and Evaluation Report Memorandum Series Chapter #ACS15-RER-10. Available at: https://www.census.gov/content/dam/Census/library/-working-papers/2015/acs/2015_Nichols_01.pdf (accessed November 2021).

Olson, C.A. 2014. "Survey Burden, Response Rates, and the Tragedy of the Commons." *Journal of Continuing Education in the Health Professions* 34: 93–95. DOI: https://doi.org/10.1002/chp. 21238.

Peytchev, A. 2009. "Survey Breakoff." *Public Opinion Quarterly* 73(1): 74–97. DOI: https://doi.org/10.1093/poq/nfp014.

Phipps, P. 2014. "Defining and Measuring Respondent Burden in Establishment Surveys." *Presentation at the Federal Committee on Statistical Methodology (FCSM) Statistical Policy Seminar*, December 16, Washington D.C., USA.

Presser, S. and S. McCulloch. 2011. "The growth of survey research in the United States: Government-sponsored surveys, 1984-2004." *Social Science Research* 40: 1019–1024. DOI: https://doi.org/10.1016/j.ssresearch.2011.04.004.

Read, B. 2019. "Respondent burden in a mobile app: Evidence from a shopping receipt scanning study." *Survey Research Methods* 13: 45–71. DOI: https://doi.org/10.18148/srm/2019.v1i1.7379.

Ritchie, J., L. Spencer, and W. O'Connor. 2003. "Carrying out Qualitative Analysis." In *Qualitative Research Practice: A Guide for Social Science Students and Researchers*, edited by J. Ritchie and J. Lewis: 219–262. London: Sage Publications.

Sharp, L.M. and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 47: 36–53. DOI: https://doi.org/10.1086/ 268765.

Tortora, R.D. 2017. "Respondent Burden, Reduction Of." In *Encyclopedia of Statistical Sciences*, edited by S. Kotz, C.B. Read, N. Balakrishnan, B. Vidakovic, and N.L. Johnson. New York: John Wiley and Sons.

Tourangeau, R. and T.J. Plewes. 2013. *Nonresponse in Social Science Surveys: A Research Agenda.* Panel on a Research Agenda for the Future of Social Science Data

Collection, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. DOI: https://doi.org/10.17226/18293.

Tourangeau, R. and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133: 859–883. DOI: https://doi.org/10.1037/0033-2909.133.5.859.

U.S. Bureau of Labor Statistics. 2012. *Consumer Expenditure Surveys Quarterly Interview CAPI Survey 2011–2012.* Available at: https://www.bls.gov/cex/capi/2011/cecapihome.htm

U.S. Bureau of Labor Statistics. 2013. *Consumer Expenditure Surveys Quarterly Interview CAPI Survey 2013.* Available at: https://www.bls.gov/cex/capi/2013/cecapihome.htm

U.S. Bureau of Labor Statistics. 2017. *Consumer Expenditure Survey Interview Questionnaire (CEQ) – 2017.* Available at: https://www.bls.gov/cex/capi/2017/2017-CEQ-CAPI-instrument-specifications.pdf (accessed November 2021).

U.S. Census Bureau. 2014. *American Community Survey Design and Methodology (January 2014).* Technical Report. Available at: https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_report_2014.pdf (accessed November 2021).

U.S. Census Bureau. 2017. *ACS Notification: Change in Methodology for Nonresponse Follow Up Data Collection.* Available at: https://www.census.gov/programs-surveys/acs/news/updates.2017.html (accessed November 2021).

U.S. Office of Personnel Management (OPM). 2011. *Paperwork Reduction Act (PRA) Guide Version 2.0.* Available at: https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/paperwork-reduction-act-guide.pdf (accessed November 2021).

Wenemark, M., G.H. Frisman, T. Svensson, and M. Kristenson. 2010. "Respondent Satisfaction and Burden among Differently Motivated Participants in a Health-Related Survey." *Field Methods* 22: 378–390. DOI: https://doi.org/10.1177/1525822X103 76704.

Willis, G.B. 2015. *Analysis of the Cognitive Interview in Questionnaire Design.* New York: Oxford University Press.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response Burden: What Is It and What Predicts It?" In *Advances in Questionnaire Design, Development, Evaluation, and Testing* edited by P.C. Beatty, D. Collins, L. Kaye, J. Padilla, G.B. Willis, and A. Wilmot: 193–212. Hoboken, NJ: John Wiley and Sons.

Yang, D.K. 2015. "Compiling Respondent Burden Items: A Composite Index Approach." *Presentation at the Consumer Expenditure Survey Methods Symposium*, July 14, Washington D.C., USA. Available at: https://www.bls.gov/cex/respondent-burden-index.pdf (accessed November 2021).

Yu, E.C., S. Fricker, and B. Kopp. 2015. "Can Survey Instructions Relieve Respondent Burden?" *Presentation at the Annual Conference of the American Association for Public Opinion Research*, May 15: Hollywood, FL, USA. Available at: https://www.bls.gov/osmr/research-papers/2015/pdf/st150260.pdf (accessed November 2021).

Received March 2021
Revised June 2021
Accepted December 2021

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1205–1234, http://dx.doi.org/10.2478/JOS-2022-0051

Determination of the Threshold in Cutoff Sampling Using Response Burden with an Application to Intrastat

Sašo Polanec[1], *Paul A. Smith*[2], *and Mojca Bavdaž*[1]

Statistical offices frequently use cutoff sampling to determine which businesses in a population should be surveyed. Examples include business surveys about international trade, production, innovation, ICT usage and so on. Cutoff thresholds are typically set in terms of key variables of interest and aim to satisfy a minimum coverage ratio–the share of aggregate values of reporting units. In this article we propose a simple cost-benefit approach to determination of the sampling cutoff by taking into account the response burden. In line with existing practice, we use the coverage ratio as our measure of accuracy and provide either analytical or numerical solutions to cutoff determination. Using a business survey on response burden of reporting trade flows within the EU (Intrastat), we present an application that illustrates our approach to cutoff determination. An important practical implication is the possibility to set industry-contingent cutoffs.

Key words: Accuracy; cost-benefit analysis; reporting costs; social benefits; social welfare function.

1. Introduction

As long ago as 1978, Bradburn (Bradburn 1978) raised the issue of "respondent burden" because of an increasing 'number and complexity of sample surveys'. Since then, burden has grown into an even more urgent matter, especially in business surveys, as with global competition, pressures on achieving higher productivity and the spread of lean approaches to eliminate any 'waste' activity, working time has become more precious and the opportunity cost of survey response higher. Businesses have been loudly questioning the need to provide data to the government, which has contributed to programs of administrative burden reduction and increased attention to burden in mandatory business surveys. National statistical institutes (NSIs) have taken different measures to reduce response burden; some of them try to help businesses to carry the burden by providing guidance and improving processes and others try to reduce the burden by reducing sample

[1] School of Economics and Business, University of Ljubljana, Ljubljana, Slovenia. Emails: saso.polanec@ef.uni-lj.si and mojca.bavdaz@ef.uni-lj.si
[2] Department of Social Statistics and Demography, University of Southampton, Southampton, UK. Email: P.A.Smith@soton.ac.uk

Acknowledgments: We are grateful to the Associate Editor and three anonymous Referees for their comments. We would also like to thank the Statistical Office of the Republic of Slovenia (SURS) and Simon Perše for access to data, and for the past research which inspired this analysis. Mojca Bavdaž was partly funded by the Slovenian Research Agency under the research programme Digitalization as Driving Force for Sustainability of Individuals, Organizations, and Society (no. P5-0410). The 'Survey on current response burden in Intrastat' (SURS 2014) received funding from Eurostat under the project Modernising Intrastat carried out by SURS (January 2014–March 2015) under grant agreement no. 06142.2013.001-2013.521

sizes, numbers of questions, and so on (Bavdaž et al. 2015, Giesen et al. 2018). This article mainly focuses on this second approach, controlling burden through sample sizes.

Economic activity is typically concentrated, with a large proportion of activity in a few, large businesses, and a relatively small proportion amongst many small businesses. Moreover, the activity of small businesses is often strongly correlated with that of surveyed large businesses through vertical supply linkages. This makes it possible to completely disregard the business units with the smallest values (a take-none stratum, hereinafter *E*, for "excluded") at the cost of a small bias. This sampling procedure is known as cutoff sampling. The part of the population which is included may be completely enumerated (take-all, *C*, for "completely enumerated"), sampled (take-some, *S*, for "sampled"), or divided into different strata in which one of the preceding options is used. The reason for choosing cutoff sampling is mainly pragmatic, to substantially reduce the cost of the survey while only modestly decreasing accuracy (Knaub 2007). Research on cutoff sampling has been relatively scarce. Some studies have focused on decreasing the potential bias in estimation because of excluded units by using auxiliary information (e.g., Haziza et al. 2010; Guadarrama et al. 2020). Other studies have focused on establishing optimal thresholds between take-none, take-some and take-all strata (e.g., Benedetti et al. 2010). We found only one study, Lee and Shin (2016), that explicitly considers the cost of data collection, and that has a quite different context. It doesn't consider the burden imposed on businesses but rather the cost to the statistical agency; it optimises the trade-off in sample size between an expensive take-all stratum and a less expensive take-some stratum; and it does not include a take-none stratum. Indeed, Lee and Shin (2016)'s optimisation uses variance estimates which do not exist in a take-none stratum.

Besides reducing the cost on the NSI side, another advantage of cutoff sampling is the reduction of response burden, typically among the smallest units that are most sensitive to it. To give some examples where cutoff sampling is used in the European Statistical System: the Prodcom Regulation states that NSIs are not required to survey enterprises with fewer than 20 employees (Eurostat 2020d); the Labour Cost Survey typically includes units with at least ten employees (Eurostat 2020b); the population for the Community Innovation Survey consists of all enterprises with ten or more employees (Eurostat 2020c); the population for the survey on ICT usage in enterprises consists of enterprises with ten or more persons employed (Eurostat 2020a), and so on. These examples demonstrate two subtly different strategies–the first two conceptually cover the whole population of businesses but use a cutoff design, whereas the second two redefine the population of interest to exclude businesses below the cutoff (Dalén 2005).

In Intrastat, the system for collecting information and producing statistics on the trade in goods between Member States of the European Union, coverage of at least 93% of the total value of intra-EU imports, and 97% of the total value of intra-EU exports is prescribed by Regulations (EC) No 638/2004 and No 222/2009. EU Member States have to set the exemption threshold for Intrastat so that it satisfies the prescribed coverage. Thresholds are therefore country specific and may change annually (thresholds for 2017–2020 are given in the Annex of Eurostat 2021b). The required minimum coverage of intra-EU imports was reduced three times in the past because of the high response burden in what is the largest business survey in the EU. With the implementation of the so-called FRIBS (Framework Regulation Integrating Business Statistics) regulation (Commission Implementing Regulation (EU) 2020/1197, 2020), the production of import statistics will not be

regulated at the EU level any longer, while more data will have to be collected on intra-EU exports from units above the exemption threshold and exchanged confidentially with relevant NSIs of other EU Member States; that is, intra-EU trade flows may only be measured in one direction.

The cutoff value is set according to a detailed technical procedure that also takes into account nonresponse, fluctuations in economic activity and response burden (Eurostat 2021a, Sec. 5). Scenarios with different coverage rates (i.e., different exemption thresholds) for 2005 showed that the impact on response burden (measured by the number of providers of statistical information) varied considerably across the Member States (Eurostat 2007a). Compared to a census, the impact would be larger when reducing the coverage of exports than imports for most Member States, with some notable exceptions (e.g., Germany, Italy, Austria), while compared to the actual coverage rate, reducing the coverage of exports would be preferred to imports in even more scenarios. Reducing the coverage rates would primarily affect businesses that report trade flows for just a few products, as their number and percentage among reporting units would considerably decrease, and it would also lower the quality of data at disaggregated (product, country) levels (Eurostat 2007b).

Despite the detailed presentation of scenarios, the selection of a specific coverage rate (and consequently, the exemption threshold) appears to be experiential rather than theoretically founded. Our article aims to address this gap by offering a simple theory of cutoff determination built around two key concepts from official statistics, namely accuracy and response burden. The research exposes trade-offs that are implicit in the cutoff determination, thus contributing to an increased transparency that is also advocated by the European Statistics Code of Practice (Eurostat 2018). Unlike Lee and Shin (2016), our approach focuses on costs incurred by businesses and uses cost functions that allow both fixed and variable costs. Heterogeneity of costs across units is common in business surveys and it is also clearly present in the Intrastat data used for illustration.

Our research was inspired by Abowd and Schmutte (2019) who analyse the trade-off between two competing mandates of NSIs – ensuring sufficient accuracy and privacy protection of responding units – as a problem of resource allocation; a decrease in accuracy is required to increase privacy protection. They solve an optimization problem where the objective is a social welfare function that depends on both accuracy and privacy, while taking into account a transformation curve as a constraint. Cutoff determination addresses a similar trade-off: a decrease in accuracy is required to free businesses of response burden; to increase accuracy, businesses incur higher opportunity costs as productive time is lost on reporting. However, for a social welfare function, we would have to model social preferences defined over accuracy of official statistics as the public good, and over consumption (arising from output generated from resources not used for statistical reporting) as the private good. Even though the use of social welfare functions is recommended for an analysis of costs and benefits (Adler and Posner 1999; Adler 2019), this approach goes beyond the scope of this article. We therefore narrow down the problem to the comparison of benefits from accuracy and costs incurred by businesses, and treat them in the framework of a cost-benefit analysis. A cost-benefit analysis is a technique for making decisions based on enumeration and evaluation of all relevant (direct and indirect) costs and benefits (Prest and Turvey 1965; Layard and Gleister 1994). We will conduct a simple cost-benefit analysis given our focus just on the main (but most sensitive)

component of costs, and therefore disregard cost components such as privacy costs, or costs to NSIs of running the survey. Optimal levels of response burden and accuracy will be determined when the marginal cost of the additional response burden equals the marginal benefit of the additional accuracy.

The article continues in Section 2 with a discussion of related literature, and Section 3 proposes a theory of cutoff determination, which also includes the operationalisation of relevant concepts. Section 4 presents an empirical application using intra-EU trade data and Section 5 concludes with a discussion of the findings and some proposals for future research.

2. Related Literature

In this section we briefly discuss the literature that provides the two main building blocks for our proposed theory of cutoff determination. The first block is the theory of cutoff sampling and second is the cost-benefit analysis. We discuss these in turn.

2.1. Cutoff Design

The literature on designing cutoff samples is rather sparse even though they are widely used. It distinguishes between two cases – one where there are two strata, a take-all stratum (C) and a take-none stratum (E); and a second where there is additionally a take-some stratum (S). These two cases will be labelled EC and ESC respectively.

In EC designs, there is strictly no (design-based) sampling variance, whereas in ESC a sampling variance arises in the S stratum. E can be regarded as containing negligible activity, so that it may be ignored, or, more usually, a model based on some auxiliary information is used as the basis for estimation in E. Knaub (2007) suggests the use of the variance from the model (a model-based accuracy measure) as the variance component, and this can be applied in both EC and ESC designs as the basis of a variance component of (model-based) mean squared error (henceforth MSE).

The approaches to defining the cutoff values (the boundary(ies) of the strata) are then:

1. The cutoff is set according to a minimum coverage criterion (Yorgason et al. 2011), so that at least $x\%$ of the population is covered by responses (for Intrastat this percentage is prescribed by Regulations (EC) No 638/2004 and No 222/2009). The idea is to make the bias (from either neglecting E or using a modelled estimate for E) sufficiently small. But this is not solvable through a minimisation, since the bias is minimised when C covers the whole population. So cutoffs set in this way rely on judgement. For further approaches it is necessary to generate an expression for the MSE (either the design-based MSE in ESC or the model-based MSE in either EC or ESC), which is followed by:
2. Minimising the MSE analytically with respect to the stratum boundaries. This is the basis of the Lavallée and Hidiroglou (1988) approach, which aims to optimise the placement of the boundary between C and S (without considering E).
3. Minimising the MSE numerically with respect to the stratum boundaries. Benedetti et al. (2010) use simulated annealing for this purpose to deal with ESC designs with multiple auxiliary variables, and Bee et al. (2011) use simulated annealing for this purpose to deal with ESC designs with multiple auxiliary variables, and Bee et al.

(2011) use the same approach to the univariate case in longitudinal surveys, focusing on determining the sample size required to meet a fixed MSE for estimating a ratio, and for estimating a total in the presence of nonresponse. This kind of approach has been used more widely in stratified designs (see Smith and Yung (2023) for an overview), but needs to be adapted to cutoff sampling to include the bias from neglecting or modelling E in the MSE.

4. Chaudron and Carlier (2015) do a grid search over the stratum boundaries in an ESC design and then choose the smallest MSE, although their description is not comprehensive.

There are therefore rather few approaches to objective optimisation of the cutoff with a univariate predictor. In this article we consider an EC design, similar to Yorgason et al. (2011). However, instead of relying on judgement in determining the cutoff, we develop an objective approach based on the trade-off between the costs and benefits of response in the survey.

2.2. *Cost-Benefit Analysis*

The second building block of our approach is cost-benefit analysis (Prest and Turvey 1965; Layard and Gleister 1994), which allows cutoff determination without reliance on arbitrary judgment. The cost-benefit approach is in our view more general as it does not only focus on minimization of MSE (subject to some constraints), but instead on choosing accuracy that yields the highest net benefits. In order to be able to calculate such net benefits, Layard and Gleister (1994, 2) note: "we should be willing to assign numerical values to costs and benefits, and arrive at decisions by adding them up and accepting those projects whose benefits exceed their costs."

The overall government goal is to maximize the social welfare subject to constraints given by preferences, technology and resources. In line with this aim, the government should use the cost-benefit approach to decisions by measuring the costs and benefits of individual agents, and then aggregating them. In the context of surveying, the costs are born by: (1) taxpayers, who pay for surveys conducted by national statistical institutes and (2) responding firms, whereas the benefits from surveying are accrued by residents, either directly or indirectly through ownership of firms. The residents enjoy the benefits of better governance (better economic policies) that stem from research activities and improvements in general knowledge, such as understanding economic relationships (a form of public good) that may lead to higher economic growth, but also macroeconomic and microeconomic policies designed to alleviate the implications of various transitory economic shocks. According to the cost-benefit analysis principle (Layard and Gleister 1994) we should appropriately determine the relative values of costs and benefits, taking into account the time, risk and individual valuations of incomes (due to income differences marginal utilities or valuations vary). While the most convenient way of expressing these costs and benefits is in monetary terms, this is not necessary (Layard and Gleister 1994).

In our framework both costs and benefits of reporting increase with accuracy – defined as the coverage ratio – which in turn depends on the cutoff that determines units included (and excluded) from the survey (The coverage ratio reflects the bias component of accuracy from sampling, since there is no sampling variance in a cutoff design). The main

trade-off is between social benefits of accuracy and reporting costs of survey respondents. While in practice the costs and benefits may accrue in different time periods and may be uncertain, we assume these complications away, but allow for differences in income inequality that may generate decreasing marginal benefits of accuracy. The proposed approach is different from the existing statistical literature, which has usually considered the social cost of inaccuracy alone (Spencer 1985; Spencer and Moses 1990; Trottini and Fienberg 2002; Spencer and Seeskin 2015).

Our approach is also quite different from the analysis recently considered by Abowd and Schmutte (2019). Unlike our problem of maximization of net social benefits, they frame the sample determination as a social welfare maximization problem with a trade off between the accuracy of statistical outputs and the privacy of survey respondents. While we could formulate our problem as a social welfare optimization problem, with accuracy as the public good and consumption as the private good, the limited information on the parameters of social welfare functions makes this route impractical, so we do not pursue it here. The second important difference from Abowd and Schmutte (2019) is in the arguments of the social welfare function. While we consider the costs of reporting, they focus on the cost of privacy loss. Lastly, their measure of accuracy is the expected squared error loss, which allows the error between the published estimate and the true population value to be stochastic due to randomness in the data publication mechanism, whereas our measure is a simpler coverage ratio that is not random.

3. A Theory of Reporting Cutoff Determination

3.1. Theoretical Framework

For cutoff determination within the cost-benefit framework, we use the net social benefits function as the objective function. It depends on two elements: (1) the social benefits ($B(A)$) and (2) reporting costs of survey respondents ($C(A)$), both of which depend on accuracy (A). As these two components do not have the same measurement units, we use weights in the social benefits function that allow comparison between costs and benefits. Specifically, we assume the following functional form for net social benefits:

$$NSB(A) = B(A) - C(A) = \lambda_0 A^{\lambda_1} - C(A), \tag{1}$$

where $\lambda_0 > 0$ converts the measure of accuracy into monetary units like costs; its positive value implies that the marginal benefit of accuracy is positive. λ_1 determines the shape of the marginal benefits function. If $\lambda_1 = 1$, the marginal benefit of accuracy is constant, while if $\lambda_1 < 1$, the marginal benefit is decreasing with accuracy.

In this article, the measure of accuracy is defined in terms of the reporting variable (V) and depends (i.a.) on a reporting or exemption cutoff $\underline{V}$. In empirical illustration of our approach to cutoff determination, the reporting variables are firm-level values of Intra-EU exports and imports, and cutoffs are determined in terms of these variables. Values of exports and imports in our empirical example come from administrative sources, but from a preceding period; we assume a strong relationship to the reporting variable, but leave the challenge of explicitly accounting for this relationship for further research. We do not consider defining a cutoff on variables that are not on the frame.

We now assume that the reporting variable (V) is a random variable with a continuous distribution. Then aggregate accuracy is defined as the proportional coverage of the population quantity:

$$A(\underline{V}) = \frac{\int_{\underline{V}}^{V_{max}} V f(V) \mathrm{d}V}{\int_{V_{min}}^{V_{max}} V f(V) \mathrm{d}V} \tag{2}$$

where f denotes the probability density function of V, V_{min} and V_{max} denote the lower and upper bounds of the distribution, and $\underline{V}$ denotes the reporting cutoff (sometimes also known as the exemption cutoff). Note that V_{min} is non-negative for most reporting variables (e.g., employment, turnover, investment etc. all have a minimum value of zero).

What is the interpretation of our measure of accuracy and how is this measure related to the standard measure of mean squared error? To provide the answers to these questions, note that the numerator is a partial expectation, which can be expressed as a product of the conditional mean for reporting firms, $E(V|V \geq \underline{V})$, and the share of reporting firms, $F(V \geq \underline{V})$, whereas the denominator is the expected value of the reporting variable or unconditional mean, $E(V)$. (Note that the measure of accuracy pertains to the total quantity or aggregate, so both the numerator and denominator should be multiplied by the same mass of firms (denoted N), which cancels and is thus not shown.) Hence our measure of accuracy is a product of two components (1) a ratio between the conditional mean and the unconditional mean, which reflects the bias of the conditional expectation, and (2) the share of reporting firms. As the reporting cutoff typically exceeds the lowest values, the conditional mean exceeds the unconditional mean (the ratio is above 1), and the share of reporting firms is below 1, with the product assuming values below 1.

Our exposition is for the estimation of a total, but the accuracy measure (2) is also related to the quality of other statistics such as means or ratios, though in these cases the reduction in accuracy from increasing the threshold would often be smaller since these statistics implicitly compensate for the cutoff part of the population. Inasmuch as Equation (2) is a good reflection of the effect of changing coverage on the quality of the estimate, the same arguments can be used for other outputs.

As EC cutoff sampling does not feature any design-based variance for the estimates, the design-based root mean squared error (RMSE) for the aggregate value of V is just equal to the bias of the aggregate value and can be directly expressed as a function of accuracy:

$$RMSE = N(E(V|V \geq \underline{V})F(V \geq \underline{V})) = N\,E(\mathrm{V})(1-A(\underline{V})), \tag{3}$$

Moreover, the RMSE expressed relative to the aggregate value of V is just $1-A(\underline{V})$.

Next we turn to aggregate costs of reporting; these are defined as the following partial expectation of cost function:

$$C(\underline{V}) = N \int_{\underline{V}}^{V_{max}} c(V) f(V) dV \tag{4}$$

where unit-level reporting costs c are allowed to vary with the reported value. In line with discussion below regarding the normalization of the net social benefits function, the mass of firms, N will be normalized to 1.

The cutoff is then determined using optimization methods by maximizing the net social benefits function. The solution to this problem will generally depend on the specific assumptions regarding the continuity and functional form of the probability density function of the variable of interest $f(V)$, the functional form for the cost function $c(V)$, and the preference parameters. In this article we assume that all functions are twice continuously differentiable with respect to $\underline{V}$.

The solution to the optimisation of the net social benefits function depends on the following derivative

$$NSB'(\underline{V}) = B'(A(\underline{V}))A'(\underline{V}) - C'(\underline{V}),$$

which has both corner and interior solutions. Two corner solutions refer to situations where no units are in the sample (no survey) or all units are in the sample (a census). The first case arises when the marginal social benefit of reporting is negative even at the upper bound ($\lim_{\underline{V}\to V_{max}} NSB'(\underline{V}) < 0$), and a census is a solution when at the lower bound the marginal net social benefit is still positive ($\lim_{\underline{V}\to V_{min}} NSB'(\underline{V}) > 0$). While these two cases are clearly relevant, we focus on determining interior solutions for which $NSB'(\underline{V})$ is equal to zero.

3.2. *Cutoff Determination for Reporting Variables with Log-Normal Pdfs*

The optimal reporting cutoff from the first-order condition $NSB'(\underline{V}) = 0$ depends on (1) the functional forms for the pdf $f(V)$, the function that maps accuracy to social benefits $B(A)$ and the function for the costs of reporting $c(V)$, and (2) the specific values of the parameters for these functions. The purpose of this article is not to provide a wide range of theoretical results for different functional forms, but instead to derive results for functional forms which are commonly used in survey and economic analysis, and to apply these to the Intrastat reporting of Slovenian business units (see Section 4).

As our main theoretical case we assume a log-normally distributed random variable V (i.e., ln V is normally distributed). We motivate this assumption with empirical evidence on business revenues whose logarithms are often bell-shaped and may be reasonably approximated using this distribution (e.g., Cabral and Mata 2003; Head et al. 2014; Bee and Schiavo 2018; Fernandes et al. 2018). We also consider an alternative assumption of the Pareto distribution (in Appendix, Subsection 6.1) that is used in theoretical models of international trade with heterogeneous businesses (e.g., Melitz 2003) and also documented in empirical analysis of business size distributions for variables such as revenues and employment (e.g., Axtell 2001; Gabaix 2009). For the cost function we assume a loglinear functional form in this section and Appendix, Subsection 6.1. (e.g., Nerlove 1963; Christensen and Green 1976). Finally, for the function that links benefits to accuracy we consider a power function with either constant or decreasing marginal social benefits of accuracy. The constant marginal social benefits require that agents are both risk neutral and society features no inequity aversion, while either risk aversion or inequity aversion would imply declining marginal social benefits (Stern 1976). An alternative motivation for declining marginal benefit of accuracy may be declining value of information for policy makers (e.g., Bikhchandani and Mamer 2013).

Based on the assumption of normality for ln V with mean μ and variance σ^2, we can express accuracy given in Equation (2) as a function of the reporting cutoff:

$$A(\underline{V}) = \frac{\frac{1}{\sqrt{2\sigma}} \int_{\underline{V}}^{V_{max}} e^{\ln V} e^{-\frac{1}{2}(h(\underline{V}))^2} \mathrm{d}V}{\frac{1}{\sqrt{2\sigma}} \int_{V_{min}}^{V_{max}} e^{\ln V} e^{-\frac{1}{2}(h(\underline{V}))^2} \mathrm{d}V} = \Phi\left(h\left(\underline{V}\right)\right) \tag{6}$$

Here Φ denotes the cumulative distribution function of the standard normal distribution and $h(\underline{V}) = (\mu + \sigma^2 - \ln \underline{V})/\sigma$. As the CDF of the standard normal distribution is strictly increasing and $h(\underline{V})$ is a decreasing function of $\underline{V}$, raising (lowering) the reporting cutoff reduces (improves) the accuracy.

Next, we determine the benefits of accuracy. These are harder to model as empirical evidence is rather scant. Social benefits pertaining to individual agents may feature either linear functions (Abowd and Schmutte 2019) or concave functions (Samuelson 1954). A sum of individual utilities even in the linear case can give a concave social benefits function (see e.g., Stern (1976) for an explanation of two sources of concavity of social welfare functions). Hence, as in Equation (1), we assume a simple increasing and concave power function with two parameters that is commonly used in economics:

$$B(A(\underline{V})) = \lambda_0 A(\underline{V})^{\lambda_1}, \tag{7}$$

where $\lambda_0 > 0$ and $0 < \lambda_1 \leq 1$. Note that a higher value of the scalar λ_0 implies higher social benefits for given accuracy and vice versa. $\lambda_1 < 1$ implies that the marginal social benefits of accuracy decline with increasing $A(\underline{V})$, whereas $\lambda_1 = 1$ implies that marginal social benefits are constant.

Using expression (6) in the social benefits function allows us to determine the marginal social benefit (the derivative of the social benefits function) of changing the reporting cutoff:

$$\frac{\mathrm{d}B}{\mathrm{d}A}\frac{\mathrm{d}A}{\mathrm{d}\underline{V}} = -\lambda_0\lambda_1\Phi(h(\underline{V}))^{\lambda_1 - 1}\phi(h(\underline{V}))(\underline{V}\sigma)^{-1} \tag{8}$$

where $\phi(.)$ denotes the density of the standard normal distribution. The marginal social benefit of increasing (decreasing) the reporting cutoff is negative (positive) for all values, as such changes reduce (increase) the accuracy. Further note that the marginal social benefits vary with cutoff non-monotonically as the second derivative $\mathrm{d}B^2/\mathrm{d}\underline{V}^2$ is either positive or negative. Namely, the second derivative of the benefits function with respect to the cutoff is $\mathrm{d}^2B/\mathrm{d}\underline{V}^2 = \lambda_0\lambda_1\Phi^{\lambda_1 - 1}\left((\lambda_1 - 1)\Phi^{-1}\phi^2(h')^2 + \phi'(h')^2 + \phi h''\right)$, where ϕ and Φ are shorthand notations for the functions that depend on function $h(\underline{V})$, which is also denoted shortly as h. The first term in the parentheses is always negative for assumed $\lambda_1 < 1$ (or 0 for $\lambda_1 = 1$), the second term is either negative or positive, whereas the third term is always positive. However, in our empirical examples, the second derivative tends to be positive, which implies that the marginal social benefit is higher at higher values of the reporting cutoff.

In order to calculate the aggregate or social cost of reporting, $C(\underline{V})$, for the loglinear cost function $c(V) = e^{\ln c_0 + c_1 \ln V}$, we calculate the following partial expectation (normalizing N to 1):

$$C(\underline{V}) = \int_{\underline{V}}^{V_{max}} c(V) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\ln V - \mu}{\sigma}\right)^2} \mathrm{d}V = e^{\ln c_0 + c_1\mu + \frac{c_1^2\sigma^2}{2}} \Phi\big(g(\underline{V})\big)$$

where $g(\underline{V}) = (\mu + c_1\sigma^2 - \ln \underline{V})/\sigma$. The derivative of the social cost of reporting with respect to the reporting cutoff, which can be interpreted as the marginal social cost of reporting, is also negative:

$$\frac{\mathrm{d}C(\underline{V})}{\mathrm{d}\underline{V}} = -e^{\ln c_0 + c_1\mu + \frac{c_1^2\sigma^2}{2}} \phi\big(g(\underline{V})\big)\big(\underline{V}\sigma\big)^{-1} \tag{9}$$

This expression implies that the social costs of reporting decrease (increase) when the reporting cutoff increases (decreases), although at a different rate than the marginal benefits of accuracy. Equating the two derivatives with respect to the cutoff given in Equations (8) and (9) yields:

$$\lambda_0\lambda_1\Phi(h(\underline{V}))^{\lambda_1 - 1}\phi(h(\underline{V})) = e^{\ln c_0 + c_1\mu + \frac{c_1^2\sigma^2}{2}} \phi\big(g(\underline{V})\big) \Rightarrow$$

$$\lambda_0\lambda_1\Phi(h(\underline{V}))^{\lambda_1 - 1} = e^{\ln c_0 + \mu + \frac{1}{2}\sigma^2 - (1 - c_1)\ln \underline{V}} \tag{10}$$

The optimal reporting cutoff is determined by this implicit function. Before turning to discussion of the effects of changing key parameters, note that the solutions to this equation are invariant to simultaneous re-scaling of both cost and preference parameters λ_0 and c_0 (which is also the reason we could normalize N to 1). The solution only depends on the ratio between the two, which gives some flexibility to use various measures of costs either in terms of monetary units or units of time. As long as the costs of reporting per unit of time are homogeneous across businesses (i.e., an hour of reporting bears a similar monetary cost regardless of unit characteristics), the optimal reporting cutoff will be the same.

Turning to analysis of the determinants of the optimal reporting cutoff in Equation (10), note that analytical solutions are attainable under the assumption of constant marginal social benefits of accuracy, that is, $\lambda_1 = 1$, which we also use below in our empirical application. Focusing on unique interior solutions, which exist for $c_1 < 1$ (costs of reporting are subject to economies of scale) and finite values of λ_0, the optimal reporting cutoff is:

$$\underline{V} = \exp\left(\frac{\ln c_0 + \mu + \frac{1}{2}\sigma^2 - \ln \lambda_0}{1 - c_1}\right)$$

Evidently the optimal reporting cutoff increases in both cost parameters (c_0 and c_1) and parameters that determine the mean of the lognormal distribution ($\mu + \frac{1}{2}\sigma^2$), and decreases with the parameter that reflects the weight given to accuracy in the social benefits function (λ_0).

For decreasing marginal benefits of accuracy ($\lambda_1 < 1$), we rely on numerical solutions and illustrate the determination of the optimal reporting cutoff graphically. Figure 1 shows the curves for the marginal social benefits and marginal social costs for a set of specific values from our application below in Section 4. It is evident that both the marginal social

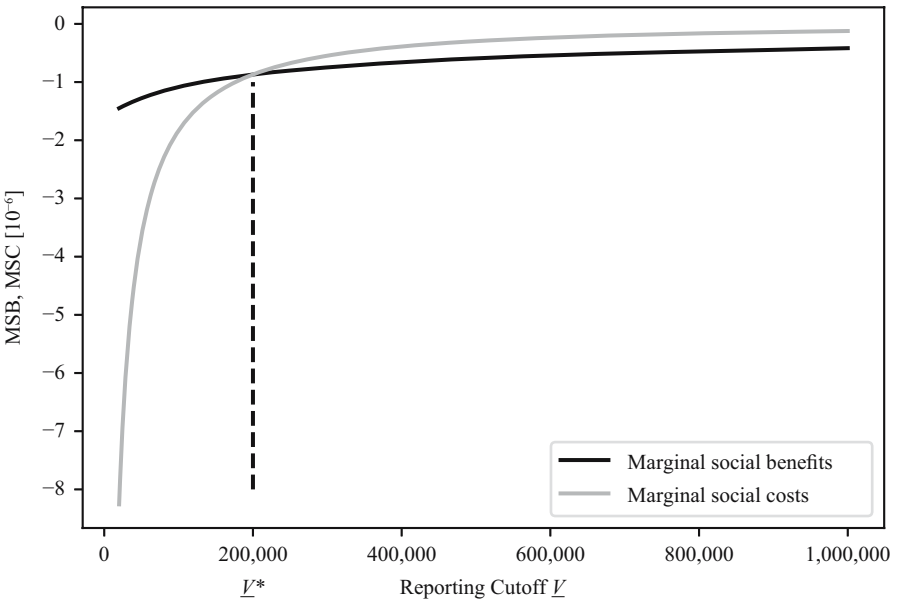


Fig. 1. Marginal social benefits and marginal social costs of reporting.
Notes: For this figure it is assumed that ln V is distributed normally with parameters $\mu = 9.233$ and $\sigma^2 = 8.111$. The cost function is loglinear with parameters $c_0 = 0.105$ and $c_1 = 0.240$. The parameters for converting accuracy into social benefits are $\lambda_0 = 7.841$ and $\lambda_1 = 0.800$. The optimal reporting cutoff $\underline{V}^*$ is at 200,000 and the corresponding accuracy is 0.9644.

benefits and costs are negative and increase with the reporting cutoff. As the marginal social costs are lower than the marginal social benefits at low cutoff values and the former also increase with cutoff at higher rates, the two curves intersect (at point $\underline{V}^*$) and determine the socially optimal cutoff. Alternatively, starting at higher values of the reporting cutoff where the marginal benefits exceed the marginal costs in absolute terms, lowering the reporting cutoff increases net social benefits (but at a decreasing rate) until the optimal cutoff is reached.

Although Figure 1 is used for illustrative purposes only, note that four (out of six) parameters are obtained from the Intrastat application in Section 4. In particular, we use the fitted parameter values for the business unit-level distribution of log (annual) values of total exports of goods from Slovenia to other EU countries in 2013 ($\mu = 9.233$ and $\sigma^2 = 8.111$). The parameters of the loglinear cost function are estimated using a survey (the 'Survey on current response burden in Intrastat' (SURS 2014); see also Perše (2016) for details) measuring the costs of reporting, completed by a set of Slovenian business units (mostly firms and sole proprietors) that were engaged in intra-EU trade in goods: $c_0 = 0.105$ and $c_1 = 0.247$. The parameters for converting accuracy into social benefits are $\lambda_0 = 7.841$ and $\lambda_1 = 0.800$, which are selected to yield an optimal reporting cutoff equal to EUR 200,000, which corresponds to the actual reporting cutoff set in Slovenia for intra-EU export flows in 2013. The corresponding accuracy is 0.9644, which is slightly below 0.97–the official lower bound for accuracy (see Regulations (EC) No 638/2004 and No 222/2009). It is important to note again that these preference parameters are not unique and we could find an alternative set of parameters that would yield the same cutoff.

For example, if we restricted λ_1 to 1, then $\lambda_0 = 6.319$ would also replicate the actual cutoff and therefore produce the same accuracy.

Next we discuss the role of changes in key parameters on the marginal social benefits and marginal social costs, and the optimal reporting cutoff values. These variations offer further insights regarding how the cutoff determination changes over time (e.g., due to shifts in all the parameters) and across different groups of respondents (e.g., businesses in different industries).

Figure 2 depicts the effects of an increase in μ, while keeping all the other parameters at the benchmark values used in Figure 1. The motivation for this increase may be found in the dynamics of values like business revenues over time, but also in differences across industries. An increase in μ shifts both marginal social benefits and marginal social costs. The higher mean implies that the distribution of V shifts to the right, which in turn lowers the marginal social benefits as less accuracy is lost at given values of the reporting cutoff. Hence the marginal social benefits curve shifts downwards. Marginal social costs are also affected by this increase as costs of reporting increase at all levels. The combined effect of these shifts leads to an increase in the reporting cutoff from 200,000 to 388,000 with a reduction in optimal accuracy by 0.0047 (from 0.9644 to 0.9597). While analysis of a shift in σ^2 is not presented in detail, an increase in variance σ^2 decreases marginal social benefits and increases marginal social costs for all cutoff values, leading to an increase in

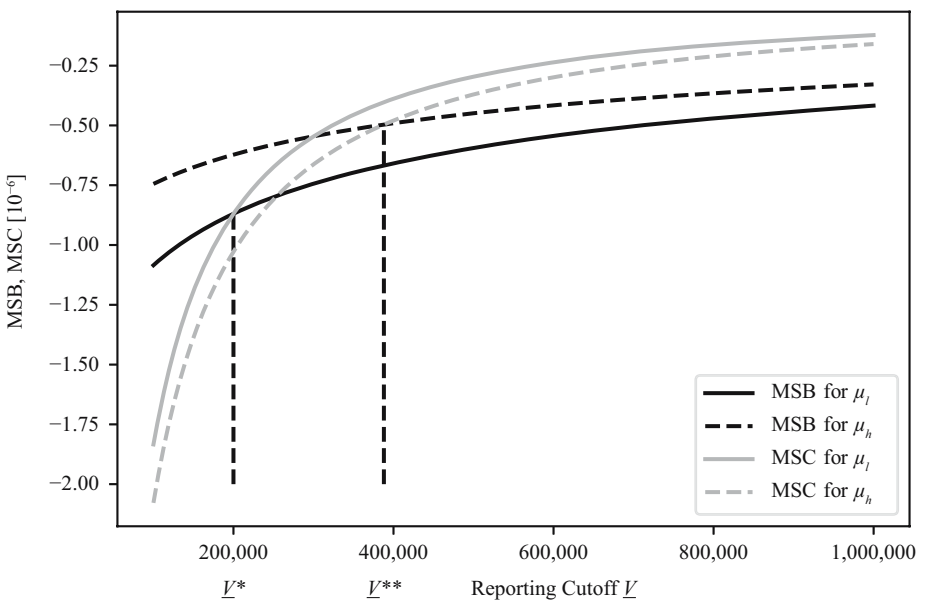


Fig. 2. The effects of an increase in μ from μ_l to μ_h on marginal social costs and benefits, and the optimal reporting cutoff.

Notes: This figure shows the impact of an increase in the mean of the lognormal distribution from $\mu_l = 9.233$ (the benchmark case) to $\mu_h = 9.733$, while keeping the variance of the distributions the same as in the benchmark case, that is, $\sigma^2 = 8.111$. The cost function is loglinear with parameters $c_0 = 0.105$ and $c_1 = 0.247$. The parameters converting accuracy into social benefits are $\lambda_0 = 7.841$ and $\lambda_1 = 0.800$. The optimal reporting cutoff $\underline{V}^*$ is 200,000 and the corresponding accuracy is 0.9644, whereas reporting cutoff $\underline{V}^{**}$ for $\mu_h = 9.733$ is significantly higher at 388,000 and the corresponding accuracy is lower at 0.9597.

the optimal reporting cutoff as well. This illustration suggests that upward (downward) shifts in means over time should be met with increases (decreases) in reporting cutoffs. Moreover, it suggests that reporting thresholds across industries should differ–industries with higher means should set higher reporting thresholds. In our empirical application we revisit these practical implications for cutoff setting.

Variation in the cost parameters is considered next. Such changes are interesting as they are likely to take place over time, but also because different groups of businesses may exhibit heterogeneous cost structures. Changes in cost parameters over time may be motivated by general technological progress, changes in the content and data collection procedures of the surveys, and changes in costs of inputs such as staff time or computer costs, whereas disparities across industries may be due to differences in economic structure that may affect reporting costs. Figure 3 shows the impact of an increase in c_0, which increases the (geometric) mean of the cost function (a shift in elasticity has a qualitatively similar effect). The marginal social cost schedule exhibits a downward shift that increases the optimal reporting cutoff from 200,000 to 282,500, and thereby leads to a reduction of accuracy by 0.0106 from 0.9644 to 0.9538. From this example we can infer that our model suggests that the reporting cutoff should be lower when technological improvements reduce the costs of reporting, higher when surveys become more detailed (increasing costs), and increase when input costs such as wages increase. Moreover, the model suggests that if cost parameters differ across industries (assuming similar

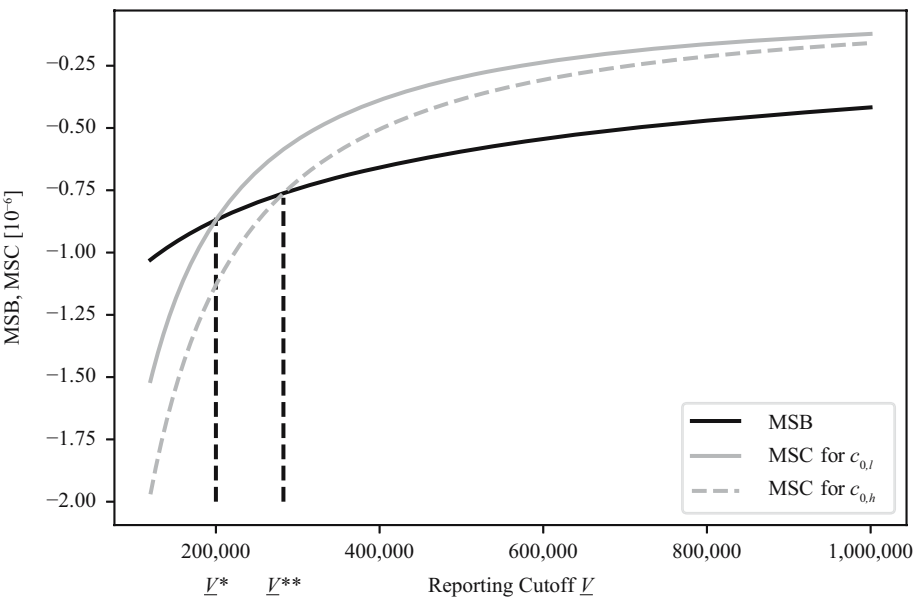


Fig. 3. The effects of an increase in c_0 from $c_{0,l}$ to $c_{0,h}$ on marginal social costs and the optimal reporting cutoff. Notes: This figure shows the impact of an increase in the cost parameter c_0 from the benchmark case $c_{0,l} = 0.105$ to $c_{0,h} = 0.1365$. All other parameters assume their benchmark values. Elasticity of costs with respect to value is $c_1 = 0.247$, ln V is distributed normally with parameters $\mu = 9.233$ and $\sigma^2 = 8.111$, and the social benefits function parameters are $\lambda_0 = 7.841$ and $\lambda_1 = 0.800$. The optimal reporting cutoff $\underline{V}^*$ is 200,000 and accuracy is 0.9644 for $c_{0,l}$, whereas reporting cutoff $\underline{V}^{**}$ for $c_{0,h}$ is 282,500 and accuracy is lower at 0.9538.

distributions $f(V)$), the reporting cutoffs should also differ. Finally, if businesses with more experience in reporting exhibit lower costs, the reporting cutoff for such businesses should be lower as well.

Our final illustration shows the impact of shifts in the parameters of the social benefits function. Figure 4 depicts an increase in λ_0; qualitatively the effects of an increase in λ_1 are similar and therefore not shown. From this figure it is evident that such a change shifts the marginal social benefits function downwards and thus lowers the reporting cutoff. While these changes are harder to motivate with practical examples, the mere fact that various surveys (e.g community innovation survey, industrial production survey) achieve different accuracies may be partly attributed to differences in the social valuations of different data sets.

4. Empirical Application: Intrastat Reporting

This section provides an application of the method of optimal cutoff determination developed in Section 2. In our illustration we discuss cutoff determination for intra-EU trade flows using data for Slovenian businesses, for which cutoff sampling is already used. As our method requires information on several parameters, an important advantage of this application is the availability of data on trade flows that allows estimation of the parameters of the distributions $f(V)$ and the cost of reporting. The key limitation of our empirical application is an inability to estimate the parameters of the function that links

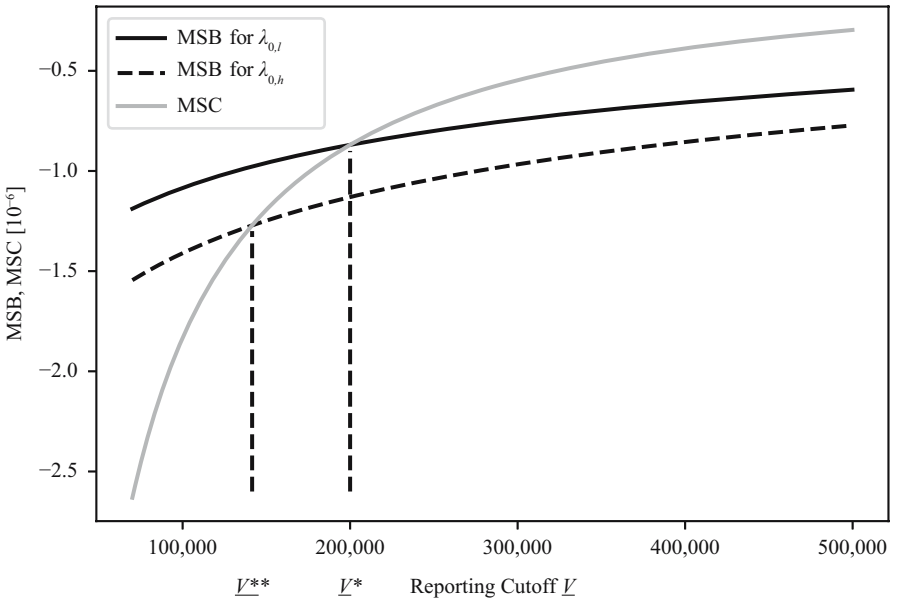


Fig. 4. The effects of an increase in λ_0 from $\lambda_{0,l}$ to $\lambda_{0,h}$ on marginal social benefits and the optimal reporting cutoff.

Notes: This figure shows the impact of an increase in the social preference parameter λ_0 from the benchmark case $\lambda_{0,l} = 7.841$ to $\lambda_{0,h} = 10.193$, while keeping $\lambda_1 = 0.800$. All other parameters assume their benchmark values. The distribution of ln V is normal with parameters $\mu = 9.233$ and $\sigma^2 = 8.111$. The cost function is loglinear with parameters $c_0 = 0.105$ and $c_1 = 0.247$. The optimal reporting cutoff $\underline{V}^*$ is 200,000 and the accuracy is 0.9644 for $\lambda_{0,l}$, whereas reporting cutoff $\underline{V}^{**}$ for $\lambda_{0,h}$ is 141,500 and accuracy is higher at 0.9729.

benefits to accuracy, which we further discuss below in Subsection 4.2. However, even without this information, using calibrated values raises important questions regarding the optimality of cutoffs used in practice.

4.1. Institutional Features of Intrastat Reporting in Slovenia

Our application uses Slovenian business unit-level data on intra-EU trade flows and a survey of a subset of these business units on response burden that was implemented for the year 2013. Therefore, we begin by setting out background information on the institutional setting for Intrastat in Slovenia. To avoid confusion, we will refer to exports and imports of goods throughout, although official terminology refers to dispatches for intra-EU exports and arrivals for intra-EU imports.

The reporting cutoffs for exports and imports in goods are set by the Statistical Office of the Republic of Slovenia (SURS). As Slovenia is a member state of the EU, SURS is obliged to comply with the aforementioned Regulations No 638/2004 and No 222/2009, which determine the lower bounds on accuracy for intra-EU reporting set by Eurostat. In 2013 the lower bounds on accuracies for intra-EU exports and imports were set to 97% of total exports and 95% of all imports of the taxable persons covered. Based on these EU-wide constraints on accuracy, SURS determined by October 31, 2012 the set of business units that were obliged to report export and import flows of goods. A business was required to report exports (imports) if it (1) was obliged to fill in and submit Value Added Tax returns and (2) had total intra-EU exports (imports) exceeding EUR 200,000 (120,000) in 2012. A business could also be added to a group of businesses reporting intra-EU trade flows within any month of the year if their 12-month cumulative trade flows (based on information on the Intrastat trade flows from the value added tax forms) exceeded the reporting cutoffs. Finally, additional reporting cutoffs were set for the collection of additional variables like type of international commercial terms, transport means and statistical value of trade, which is the net domestic value (fair market value) of a product when it crosses a border. These were set to EUR 9,000,000 for exports and EUR 4,000,000 for imports.

4.2. Data Sets

We estimate the parameters needed in determination of the optimal reporting cutoff using three main data sets obtained from three distinct sources.

A comprehensive study of response burden was conducted in May 2014 among all Slovenian businesses included in Intrastat (excluding only businesses with their seat abroad or which became part of Intrastat in 2014) (SURS 2014). For this analysis, we used data from businesses that reported data to Intrastat themselves (from $n = 3{,}579$ units, 2,141 responded, yielding a response rate of 59.9%; see Perše (2016) for more details). These data were used to estimate the parameters of the cost functions. Survey respondents were also asked to state their tax office registration number, which allowed SURS to match the data from the response burden survey with other key business-unit-level information. For our illustration we use information on the value of exports and imports in 2013, broad sectors of economic activity (manufacturing, trade and other industries), the numbers of records or distinct reporting items (i.e., distinct records are due to differences in country of

consignment/destination, delivery terms, nature of transaction, mode of transport, and country of origin), trading partners, countries and products (the latter based on the Standard International Trade Classification, SITC); and past obligation to report intra-EU trade flows up to five years back.

Businesses that report through a declarant ($n = 2{,}675$) and declarants themeselves ($n = 927$) were surveyed separately and are not included. Compared to the Intrastat population, our sample has relatively fewer micro units and relatively more larger units with more trade (5% of responding units were large compared to 4.1% among units that report themselves and 1.9% among units that report through a declarant). Additionally, we have to take into account that units below the officially determined cutoffs for reporting intra-EU trade were not surveyed. The moments of such a truncated distribution are severely biased as estimators of the full distribution, which is the motive for the use of an alternative dataset on Intrastat trade flows that is collected jointly by SURS and the Slovenian Financial Authority (SURS and FURS 2014). While the same cutoffs should apply to Intrastat trade, many business units nevertheless report trade volumes below the cutoff, partly because obligation to report is determined on the basis of past trade flows. From this source we use total values of exports and imports. Using business-level identifiers we also add broad sectors of economic activity from annual business accounts (Agency of the Republic of Slovenia for Public Legal Records and Related Services, henceforth AJPES).

Finally, as the estimated moments of the distributions from Intrastat are nevertheless likely to be biased, we also consider a data set on intra-EU trade from the Annual Business Accounts of businesses (AJPES 2014). From this source we obtain total values of exports within the EU. Although this data set does not suffer from truncation, it excludes information on exports of sole proprietors and other business units, and contains information on sales of both goods and services (whereas Intrastat covers only goods).

4.3. Descriptive Statistics

Table 1 shows descriptive statistics for the responding sample of business units with information on response burden, separately for export and import flows, that is used in estimation of the cost functions. Of 2,141 business units, 872 reported total time needed to fill the Intrastat forms on export flows, and 1,855 reported on import flows (see Appendix 3 in Perše (2016) for more information on the representativeness of the sample).

Table 1 contains descriptive statistics for original variables and log-transformed variables for the reference year 2013. From the top panel of this table it is evident that the average surveyed exporting (importing) business needed 6.6 (9.3) hours per month or roughly 80 (111) hours per year to complete the Intrastat questionnaires. The average value of intra-EU exports (imports) within this truncated sample was EUR 7.9 (4.1) million. Note that these values are significantly higher than the corresponding averages reported in Table 3 for the sample of all 2013 Intrastat reporters–these were EUR 0.87 (0.403) million. Due to this discrepancy, we expect that the estimated cost function parameters are not consistent and possibly suffer from selection bias. Table 1 also shows the descriptive statistics for the numbers of records, products, partners and countries of reported flows. The average surveyed exporting business unit within this sample sells 21

Table 1. Summary statistics for key variables for 2013 from response burden sample

	Exports			Imports		
	Observations	Mean	S.d.	Observations	Mean	S.d.
		Levels				
Reporting time (hrs)	872	6.6	10.6	1,855	9.3	15.0
Value (m EUR)	911	7.9	36.4	1,925	4.1	23.4
Products	911	20.9	38.9	1,936	55.9	93.5
Partners	930	36.6	90.6	1,925	33.2	55.8
Countries	911	6.9	6.1	1,919	5.8	4.1
Records	911	342.6	1,191.7	1,925	590.7	2,272.8
		Logs				
Reporting time	872	1.17	1.18	1,855	1.49	1.20
Value	906	13.94	1.85	1,919	13.29	1.70
Products	906	2.13	1.36	1,919	3.18	1.41
Partners	930	2.71	1.30	1,936	2.87	1.11
Countries	906	1.53	0.96	1,919	1.49	0.79
Records	911	4.43	1.59	1,925	4.93	1.72

Source: Survey on current response burden in Intrastat, May 2014 (SURS 2014).

products to 37 partners in seven countries and has 342 corresponding records. In contrast, its importing peer buys on average 56 products from 33 partners in six countries and has 590 records of these activities. We can see that import flows are significantly more dispersed than export flows, which likely contributes to the higher reporting time–even for lower transaction values.

The bottom panel of Table 1 also shows the descriptive statistics for log-transformed variables, which are used in cost function estimations. In order to give some indication of the direction and strength of correlations between these variables, Table 8 in Appendix, Subsection 6.2, also contains correlation coefficients, separately for exports and imports. All variables are positively correlated; most importantly, reporting time is particularly strongly correlated with numbers of products and records, which suggests that businesses more engaged in reporting also have more disaggregated trade flows.

As already mentioned above, Perše (2016) merged the data from the SURS survey (SURS 2014) on response burden with administrative data providing the broad sector of the respondent's main economic activity (from AJPES), distinguishing between manufacturing, trade, and other industries. For these sectors Table 2 presents further summary statistics for the two main variables of interest: value of trade flows and reporting time. From these statistics it is evident that business units in manufacturing businesses have the highest mean values for both values of trade flows and reporting times. However, differences in mean trade values do not seem to translate into differences in reporting times, suggesting important differences in the structure of trade flows. For example, the mean number of records (not shown in the tables) is the highest for 'manufacturing' among exporters, but not among importers.

The next set of descriptive statistics features estimates of the moments of the distribution of the value of transactions from the full sample of Intrastat reporters in 2013. As discussed above, we do not use the estimates of the mean and variance from Table 1 for the value of trade due to truncation of the distributions below the cutoffs set at EUR 200,000 for exports

Table 2. Trade flows and reporting times (in logs) for 2013 by sector.

	Exports				Imports			
	Value		Reporting time		Value		Reporting time	
Sector	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
All	13.945	1.847	1.173	1.177	13.296	1.699	1.486	1.203
Manufacturing	14.405	1.819	1.306	1.202	13.667	1.803	1.572	1.208
Trade	13.256	1.645	1.077	1.151	13.262	1.585	1.523	1.191
Other	13.527	1.827	0.838	1.040	12.698	1.640	1.217	1.195

Source: Survey on current response burden in Intrastat, May 2014 (SURS 2014).
Note: All variables are log transformed.

and EUR 120,000 for imports. The full sample of Intrastat reporters also suffers from truncation, as many businesses below the cutoffs do not report trade flows. However, the samples of exporting and importing businesses above the cutoffs are 2,538 and 5,181 business units, whereas the full samples used for the estimations of the moments given in Table 3 are 16,575 and 33,978 business units, respectively. The average value of exports (imports) for these larger samples are EUR 0.87 (0.40) million, which is significantly lower than the averages reported in Table 1 for the sample of businesses filling the response burden survey. Nevertheless, given that reporting under the cutoffs was still incomplete, the parameters for the lognormal distribution are likely still biased. The estimates of means (variances) are likely upward (downward) biased. In cutoff determination below we use the means and squared standard deviations for the log-transformed variables. In our basic cutoff determination exercise we use the estimates of means and variances for log transformed values given in Table 3 (these values were also used in the illustrations of the method in Subsection 3.2). The distributions of log transformed values of exports and imports for all observations are depicted in Figure 5. Evidently, these are fairly closely approximated with normal distributions. The three economic sectors have similar shapes of distributions, but different values of mean and dispersion.

The final set of statistics refers to the sample of businesses for which we have information on all intra-EU exports (including exports of services) that is collected in annual business accounts. This sample contains a complete set of businesses, but excludes

Table 3. Summary statistics for the values of imports and exports (in logs) for the full sample of Intrastat reporters, 2013.

	Exports			Imports		
	Observations	Mean	S.d.	Observations	Mean	S.d.
All	16,575	9.233	2.848	33,978	8.913	2.701
Manufacturing	4,268	10.530	3.110	6,178	9.721	2.856
Trade	5,213	9.498	2.680	10,181	10.049	2.598
Other	7,094	8.258	2.417	17,619	7.979	2.356

Source: SURS and FURS (2014) and AJPES (2014).
Note: All variables are log transformed.

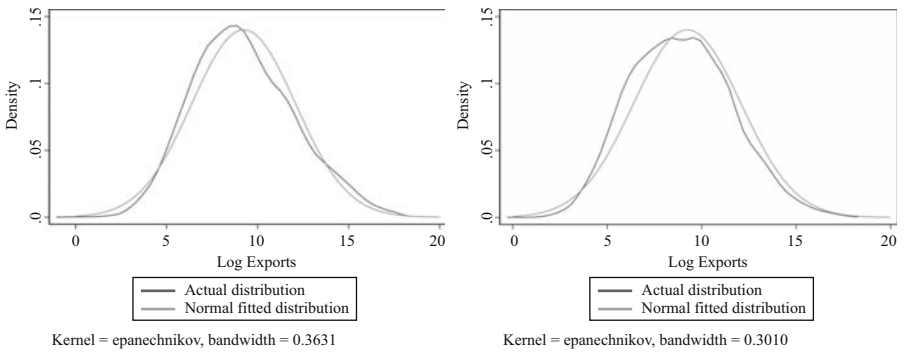


Fig. 5. Actual and fitted distributions of the values of exports and imports (in logs) for the full sample of Intrastat reporters, 2013.
Source: SURS.
Notes: All variables are log transformed. The actual distributions are smoothed kernel density plots.

Table 4. Summary statistics for the values of exports of goods and services (in logs) for the full AJPES sample, 2013.

	Observations	Mean	S.d.	Accuracy
All	12,797	10.487	2.743	0.981
Manufacturing	2,912	11.597	2.995	0.993
Trade	4,113	10.183	2.562	0.959
Other	5,752	10.143	2.584	0.964

Source: AJPES (2014).
Note: All variables are log transformed.

sole proprietors. In the construction of the sample used for calculation of the moments of the distribution, we follow the EU regulations and exclude businesses that are not obliged to pay value added tax. This reduces the final sample of businesses from 15,382 to 12,797 with total value of exports EUR 18.667 billion. For these businesses we again calculate means and standard deviations for all businesses and three economic sectors (Table 4). Comparison of the moments reveals that the exclusion of sole proprietors and addition of service flows yields higher means and mostly lower standard deviations. In addition to these statistics we are able to calculate the measure of accuracy using the full samples and samples restricted by the EUR 200,000 reporting cutoff. Note that we apply both VAT status and inclusion in reporting based on 2013 data, which makes the accuracy calculated here comparable to that reported in the determination of the optimal cutoff below. From these measures of accuracy we can observe that actual accuracy varies across sectors between 95.9% in trade and 99.3% in manufacturing, while overall accuracy is 98.1%. Accuracy in trade and other sectors is thus below the required 97%. Not surprisingly, application of the same cutoff at industry level reveals even greater variation in accuracy, some industries achieving accuracy close to 80%.

4.4. Cost Function Estimation and Cutoff Determination

In this section we estimate the cost functions using the sample of observations from the SURS survey (SURS 2014). Based on these estimates, the estimated moments of the

distributions for the values of trade flows and the assumed benefits functions, we determine the optimal reporting cutoffs.

The parameters of the cost function (c_0 and c_1) are estimated from the following loglinear equation:

$$\ln C_i = \ln c_0 + c_1 \ln V_i + \varepsilon_i \tag{11}$$

which can be done using the ordinary least squares method. We estimate this equation separately for different types of trade flows (exports and imports) and different sectors of economic activity (all sectors, manufacturing and trade) in order to allow for heterogeneity of parameters.

Table 5 shows the estimated coefficients based on constant marginal benefits of accuracy, $\lambda_1 = 1$, in Equation (7). The elasticity of reporting time (parameter c_1) is between 0 and 1 for both types of flows and all economic sectors, while the estimated constant is always negative and the corresponding parameter c_0 is small, but positive. The most interesting observation is that the parameters are significantly different between the sectors (and the same direction of flow), with manufacturing having a lower constant than trade (the difference for exports is $-3.266-(-0.765) = -2.501$ with $p<0.001$, whereas for imports it is $-3.044-(-1.716) = -1.328$ with $p = 0.004$) and higher elasticities (the difference for exports is $0.320-0.139 = 0.181$ with $p<0.001$ and for imports it is $0.339-0.245 = 0.094$ with $p = 0.006$). The estimates for the two trade flows within a sector are similar, and significantly different only in elasticity for the trade sector (the difference is $0.139-0.245 = -0.106$ with $p = 0.003$).

Table 6 shows optimal reporting cutoffs calculated with the estimated values of parameters for the cost functions and the moments of the distributions. Given our ignorance regarding the values of parameters of social preference for accuracy, we first calibrated parameter λ_0, while assuming λ_1 to be equal to 1. Calibration was done for estimated paramaters of costs and trade value distributions for all sectors jointly to replicate the official cutoffs, that is, EUR 200,000 for exports and EUR 120,000 for imports. Calibrated λ_0's are equal to 6.319 (6.849) for exports (imports) with corresponding accuracies at 0.9644 (0.9526).

Table 5. Estimates of cost function parameters.

	Exports			Imports		
	All sectors	Manufacturing	Trade	All sectors	Manufacturing	Trade
c_1	0.247**	0.320**	0.139**	0.268**	0.339**	0.245**
	(0.021)	(0.027)	(0.042)	(0.015)	(0.024)	(0.023)
log c_0	-2.251**	-3.266**	-0.765	-2.073**	-3.044**	-1.716**
	(0.288)	(0.393)	(0.562)	(0.206)	(0.335)	(0.310)
Observations	855	480	265	1,836	599	923
Adj. R^2	0.144	0.222	0.036	0.142	0.243	0.107

Source: Own calculations based on Survey on current response burden in Intrastat, May 2014 (SURS 2014).
Notes: This table shows the estimates of parameters from a regression of log reporting time on log of values of exports (imports), using Equation (11).
***,**,* indicate statistical significance level at 0.1,1 and 5%.

Table 6. Optimal reporting cutoffs by sectors and direction of trade flows.

	Exports			Imports		
Parameters	All sectors	Manufacturing	Trade	All sectors	Manufacturing	Trade
λ_0	6.319	6.319	6.319	6.849	6.849	6.849
λ_1	1.000	1.000	1.000	1.000	1.000	1.000
μ	9.233	10.530	9.498	8.913	9.721	10.049
σ^2	8.111	9.673	7.182	7.295	8.157	6.750
c_1	0.247	0.320	0.139	0.268	0.339	0.245
c_0	0.105	0.038	0.465	0.125	0.047	0.179
V^*	200,000	3,533,930	193,170	120,000	630,650	423,200
$\overline{A}(V^*)$	0.9644	0.9503	0.9538	0.9526	0.9434	0.9305
Relative RMSE	0.0356	0.0497	0.0462	0.0474	0.0566	0.0695

Source: Own calculations based on AJPES (2014), SURS (2014) and SURS and FURS (2014).

The optimal reporting cutoffs for the manufacturing and trade sectors calculated with the calibrated preference parameters are quite different from the official cutoffs, and consequently also produce different accuracies. For the manufacturing sector in particular we observe significantly higher cutoffs, which we can attribute primarily to the larger mean and variance of the distribution. For example, the optimal cutoff for reporting intra-EU export flows in the manufacturing sector is around EUR 3.5 million, which is primarily due to the larger mean and variance of the distribution of export values (for the cost parameters obtained for all businesses, the optimal cutoff would be EUR 3.15 million). Despite the higher cutoffs, we have a rather modest decline in accuracy by roughly one percentage point. More generally, we can observe that optimal cutoffs may increase or decrease. For exporting businesses in the trade sector these should increase, whereas for importing businesses in both the manufacturing and trade sectors they should decrease.

As a robustness test we also consider cutoff determination based on total intra-EU exports of goods and services (see Table 7). We follow the same approach as above and initially calibrate the parameters of social benefits of accuracy to data for all sectors using the actual cutoff (EUR 200,000). Based on this cutoff we can see that for the lognormal distribution accuracy is very close to the actual value (0.981 versus 0.983), which suggests that use of the lognormal

Table 7. Optimal reporting cutoffs by sectors using intra-EU exports of goods and services.

Parameters	All sectors	Manufacturing	Trade
λ_0	16.510	16.510	16.510
λ_1	1.000	1.000	1.000
μ	10.487	11.597	10.183
σ^2	7.524	8.970	6.564
c_1	0.247	0.320	0.139
c_0	0.105	0.038	0.465
V^*	200,000	2,464,000	243,560
$\overline{A}(V^*)$	0.9828	0.9746	0.9800
Relative RMSE	0.0172	0.0254	0.0200

Source: Own calculations based on AJPES (2014), SURS (2014) and SURS and FURS (2014).

distribution as an approximation is a reasonable choice. Based on the calibrated value of social preference for accuracy, which is now somewhat higher (16.51 versus 6.319), we again determine the cutoffs for manufacturing and trade using the same estimated parameters for cost functions and distributional moments based on the full sample of exporting businesses. We can see that the optimal reporting cutoff for both manufacturing and trade is higher than the official cutoff, although for manufacturing the difference is again extremely large. Our approach to cutoff determination suggests that taking into account the costs and benefits, reporting in manufacturing in particular should have a higher reporting cutoff and lower accuracy.

5. Discussion

5.1. Implications and Operational Considerations

In this article we have proposed a simple theory for determining the threshold in cutoff sampling that uses the trade-off of the costs and benefits of accuracy (defined as the proportional coverage of the total value of the variable of interest). The cutoff is determined by an optimisation that maximizes the net social benefits. For specific distributional assumptions on the variable of interest, the respondents' cost function and the social benefits function, we are able to derive analytical expressions for the cutoffs or (if that is not possible) provide numerical solutions.

We show that the optimal cutoffs increase with the mean and variance of the distributions and the preference for accuracy, and decrease with the reporting costs of survey respondents. This research exposes the trade-offs that are implicit in the cutoff determination in surveys of this type, thus contributing to an increased transparency that is also advocated by the European Statistics Code of Practice (Eurostat 2018).

We show how our approach could be used in practice by utilising a survey on Intrastat respondents' costs in Slovenia that allows us to estimate the cost function parameters. By combining the estimated cost parameters with estimates of the mean and variance of the distribution of the variable of interest and the calibrated social benefits function parameters, we are able to calculate the optimal cutoffs. By calibrating the social benefits function parameter to match the actual cutoffs used in Intrastat in Slovenia in 2013, we find that the optimal cutoffs in specific industrial sectors differ significantly. In the manufacturing sector the costs of reporting are higher and the mean and variance of the values of exports are higher, which together imply much higher optimal cutoffs than those set in practice. In most EU countries, the cutoffs change infrequently over time (see the national exemption thresholds in national currency between 2017–2020 in Eurostat (2021b, Table A.2.2)). For example, in Slovenia the cutoff set by SURS for export flows was EUR 200,000 between 2008 and 2017. When the economy was growing, the accuracy of total exports based on this cutoff was increasing, whereas during the great depression in 2009 accuracy decreased (results not shown).

One of the most striking implications of this approach to setting cutoffs is that during sharp changes in economic activity when the mean and variance change significantly, the cutoffs should be adjusted as well. In a similar way, cutoffs should be allowed to differ across sectors (and possibly industries), to maintain the optimum social benefits of the survey. In particular, manufacturing bore too large a burden in 2013 in our application because of the uniform cutoff. To achieve the optimum cost-benefit trade-off in the future,

the response burden on manufacturing should be relieved by increasing the cutoff and sacrificing some accuracy.

These findings present some operational challenges for a NSI, since the true cutoff for a given year depends on the mean and variance of the survey variable and the estimated costs in that year, and neither is available at the time that the reporting cutoff is set. However, information from the previous year can be used to provide mean and variance estimates to allow forecasting of the cutoffs. When there is a particularly dramatic change in economic conditions, some adjustment to annual cutoffs may be needed.

Reporting costs are not expected to change rapidly, but the collection of data on burden should continue, preferably for all units (to provide the information to make the management of the burden possible).

There is an operational argument for a constant cutoff across all industry sectors, which is based on simplicity. However, it is not *fair* in the sense measured by our cost-benefit trade-off, and using an approach such as the one we propose here seems to present a more equitable sharing of the burden of responding across businesses. How far the idea of varying cutoffs can be extended is an interesting question. Potentially, businesses reporting through declarants will have different costs to those who report directly, and this could be handled by putting these businesses in separate strata with different cutoffs. Similarly, businesses new to the survey could be separated, as they typically have higher costs than businesses which have established procedures. But handling a larger number of strata leading to a large number of different cutoffs according to business characteristics is challenging and may be too complex to implement.

5.2. *Extensions*

The approach used in this article is based on a cost-benefit trade-off, because there are sources available with which the parameters can be estimated. Ideally, however, this approach would be extended to a full social benefits function, which requires information reflecting the social benefits of accuracy. Future research should focus on measuring these benefits, which will require more investment in understanding the uses and users of these data. The concept of measuring the effects of accuracy seems a bit narrow in this context; it applies principally to established uses, while data archives prove that the same data may be used for very different purposes. Progress in this direction is likely to be challenging as illustrated with an assessment of the consequences of population census inaccuracy for apportionment of the U.S. House of Representatives and for allocation of federal funds (Spencer and Seeskin 2015). For an overview of the public good of statistics see Office for Statistics Regulation (2020).

In a similar way, the cost functions could be extended to include other cost components, for example the cost to the NSI of running the survey. This would present a second cost variable, which would need to be included through its effect on social welfare (for example noting that additional costs for the survey have an opportunity cost for other statistical outputs, or even for other government functions).

Aside from the social benefits function, future research should focus on more general and possibly more realistic distributions for key variables of interest, such as the

generalized Gamma distribution, which includes the lognormal distribution as a special case, but also allows more flexibility in its shape.

To improve the estimates of the cost function parameters, more attention could be paid to obtaining information on reporting costs of smaller units, below the cutoff values. Although the smallest businesses are not required to report, some of them do, so it is in principle possible to collect this information. Similarly, obtaining larger responding samples would allow for more accurate estimation. Extending the estimation to the parts of the business population which report through a declarant may also modify the cost parameters and therefore the cutoffs; it is an open question whether these differences are best reflected in separate cutoffs or pooled in calculating a common cutoff.

Our exposition and application are focussed on the benefits of the accuracy of the aggregate values. The impact on more detailed levels (of either variables or classifications) is not considered. Some additional complications arise in dealing with multiple components rather than an aggregate value. However, we reflect the process in Intrastat (and other similar surveys), where the reporting cutoff is set on a single, aggregate variable, with no explicit control of lower level breakdowns. In principle, of course, the currently used Intrastat cutoffs may be set artificially low to protect against poor accuracy of estimates for breakdowns.

We examined a situation where the cutoff variable is defined on past administrative information about the variable of interest, assuming that it is a strong predictor. We may also consider the situation where the cutoff variable is only a proxy for the variable measured in the survey, which will weaken the relationship. Baillargeon and Rivest (2009) discuss the design of stratified samples when the stratification variable differs from the response variable, and this approach could be applied in our context; we leave this for future research.

A natural extension is to make an estimate for the excluded part of the population, E, based on auxiliary information (which may include the variable used to define the cutoff). This should reduce the impact of the cutoff approach on the quality of estimates, compared with the strategy of neglecting the cutoff used in this article. This will link better to the use of estimates other than of totals (such as means and ratios), which implicitly include such compensation.

5.3. *Conclusion*

We have demonstrated that it is possible to set cutoffs in cutoff sampling in a way which optimises a simple cost-benefit trade-off. This objective framework offers a new direction for managing burden in official statistics with this type of design, and gives a framework for evaluating the fairness of the burden in existing surveys. The availability of information on costs of reporting is an essential component of this approach, and these data should be routinely collected.

6. Appendix

6.1. *Cutoff-Determination for Pareto Distributed Variables*

In Subsection 3.2 of the main text we assumed that the key variable of interest is distributed according to a lognormal distribution. While this assumption may be relevant

for many variables like businesses' revenues, it may not be the best choice for variables like employment, which do not feature a bell-shaped distribution. Instead, a more appropriate distribution may be the Pareto (Axtell 2001; Gabaix 2009). In this Appendix we provide analytical results for the optimal reporting cutoff assuming the variable of interest is Pareto distributed.

Let us assume a probability density function $f(V) = \alpha \frac{V_{min}^{\alpha}}{V^{\alpha+1}}$. We assume $\alpha > 1$, which ensures the existence of a finite mean for V. (Alternatively, we could impose an upper bound on the domain of the reporting variable, which would allow calculation of means also for $\alpha = 1$.) For this functional form, the measure of accuracy is a power function of the reporting cutoff, algebraically, $A(\underline{V}) = (\underline{V}/V_{min})^{1-\alpha}$. Assuming the benefits function in Equation (7), the marginal benefit of changing the reporting cutoff is $\mathrm{d}B/\mathrm{d}\underline{V} = \lambda_0\lambda_1(1-\alpha)\underline{V}^{(1-\alpha)\lambda_1-1}V_{min}^{(\alpha-1)\lambda_1}$, which is negative for the assumed parameters ($\alpha > 1$, and $\lambda_0, \lambda_1 > 0$). Hence a decrease in the reporting cutoff increases the benefits pertaining to accuracy and vice versa.

Assuming a log-linear reporting cost function yields a closed-form solution to the net social benefit maximization problem. Assuming such a cost function, again with two parameters, $c(V) = c_0 V^{c_1}$, where $c_0 > 0$, $c_1 < \alpha$ and $1 + c_1 - \alpha > 0$, the functional form for the aggregate reporting cost is $C(\underline{V}) = c_0 V_{min}^{\alpha} \int_{\underline{V}}^{\infty} V^{c_1-\alpha}\mathrm{d}V$, which can be further expressed as $C(\underline{V}) = -\frac{c_0 V_{min}^{\alpha}}{1+c_1-\alpha}\underline{V}^{1+c_1-\alpha}$. The marginal cost of changing the reporting cutoff is $\mathrm{d}C(\underline{V})/\mathrm{d}\underline{V} = -c_0 V_{min}^{\alpha}\underline{V}^{c_1-\alpha}$. The analytical solution to the first-order condition $\frac{\partial NSB}{\partial \underline{V}} = 0$, yields

$$\underline{V} = \left(\frac{\lambda_0\lambda_1(\alpha-1)}{c_0 V_{min}^{(\alpha-1)\lambda_1-\alpha}}\right)^{\frac{1}{1+c_1-\alpha-\lambda_1(1-\alpha)}}. \tag{12}$$

This expression only yields an economically sensible solution when the following condition on the model parameters holds: $1 + c_1 - \alpha - \lambda_1(1-\alpha) < 0$. Assuming this condition does hold, an increase in the preference parameters λ_0 and λ_1 – which increase valuation of accuracy – leads to a lower reporting cutoff. Moreover, an increase in the parameters capturing the costs of reporting predicts a higher reporting cutoff. For $\alpha > 1$ and $0 < \lambda_1 < 1$, the above inequality holds when the variable costs of reporting increase sufficiently slowly, that is, c_1 must be sufficiently close to 0.

6.2. Additional Descriptive Statistics

Table 8. Correlation coefficients for the key variables from response burden survey.

	Reporting time	Value	Countries	Products	Partners	Records
			Exports			
Reporting time	1.000					
Value	0.382	1.000				
Countries	0.328	0.549	1.000			
Products	0.431	0.402	0.488	1.000		
Partners	0.386	0.525	0.741	0.513	1.000	
Records	0.470	0.676	0.695	0.764	0.678	1.000
			Imports			
Reporting time	1.000					
Value	0.377	1.000				
Countries	0.367	0.525	1.000			
Products	0.517	0.530	0.584	1.000		
Partners	0.443	0.543	0.733	0.597	1.000	
Records	0.506	0.652	0.561	0.880	0.543	1.000

Source: Survey on current response burden in Intrastat, May 2014 (SURS 2014).
Note: Variables are log transformed.

Table 9. Estimates of cost function with an extended set of regressors.

	Exports	Imports
Value (log)	0.098**	0.180**
	0.027	0.030
Records (log)	0.132**	0.132**
	0.050	0.033
Products (log)	0.169**	0.206**
	0.043	0.038
Partners (log)	0.136**	0.262**
	0.043	0.034
Markets (log)	-0.105	-0.074
	0.059	0.046
Experience (years)	-0.015	-0.030
	0.013	0.008
$D_{both\ flows}$	0.133	-0.077
	0.095	0.055
Const.	-1.318	-0.842
	0.303	0.211
Adj. R^2	0.252	0.311
Observations	852	1,835

Source: Own estimates based on data from Survey on current response burden in Intrastat, May 2014 (SURS 2014).
Notes: ***,**,* indicate statistical significance level at 0.1,1 and 5%.

7. References

Abowd, J.M., and I.M. Schmutte. 2019. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices." *American Economic Review* 109(1): 171–202. DOI: https://doi.org/10.1257/aer.20170627.

Adler, M.D. 2019. *Measuring Social Welfare: An Introduction.* Oxford: Oxford University Press.

Adler, M.D., and E.A. Posner. 1999. "Rethinking Cost-Benefit Analysis." *The Yale Law Journal* 109(2): 165–247. DOI: https://doi.org/10.2307/797489.

AJPES. 2014. "Annual Balance Sheets and Income Statements Data, 2013." [data collection]. Available at: https://www.ajpes.si/Letna_porocila/Splosno. (accessed January 2021).

Axtell, R. 2001. "Zipf Distribution of U.S. Firm Sizes." *Science* 293: 1818–1820. DOI: https://www.ajpes.si/Letna_porocila/Splosno.

Baillargeon, S., and L.-P. Rivest. 2009. "A general algorithm for univariate stratification." *International Statistical Review* 77(3): 331–344. DOI: https://doi.org/10.1111/j.1751-5823.2009.00093.x.

Bavdaž, M., D. Giesen, S. Korenjak Černe, T. Löfgren, and V. Raymond-Blaess. 2015. "Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes." *Journal of Official Statistics* 31(4): 559–588. DOI: https://doi.org/10.1515/jos-2015-0035.

Bee M., R. Benedetti, E. Espa, and F. Piersimoni. 2011. "Cut-Off Approach To The Design Of Longitudinal Business Surveys," In Proceedings of the Section on Government Statistics, Joint Statistical Meetings, July 30–August 4, 2011: 2488–2500, Alexandria, VA: American Statistical Association.

Bee, M., and S. Schiavo. 2018. "Powerless: gains from trade when firm productivity is not Pareto distributed." *Review of World Economics* 154(1): 15–45. DOI: https://doi.org/10.1007/s10290-017-0295-z.

Benedetti, R., M. Bee, and G. Espa. 2010. "A framework for cut-off sampling in business survey design." *Journal of Official Statistics* 26(4): 651–671. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/20200206/jos_cut-off_benedetti-mfl.pdf. (accessed January 2021).

Bikhchandani, S., and J.W. Mamer. 2013. "Decreasing marginal value of information under symmetric loss." *Decision Analysis* 10(3): 245–256. DOI: https://doi.org/10.1287/deca.2013.0274.

Bradburn, N. 1978. "Respondent burden." In Proceedings of the Survey Research Methods Section, Joint Statistical Meetings. August 14–17, 1978: 35–40, Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf. (accessed July 2021).

Cabral, L.M.B., and J. Mata. 2003. "On the Evolution of the Firm Size Distribution: Facts and Theory." *American Economic Review* 93(4): 1075–1090. DOI: https://doi.org/10.1257/000282803769206205.

Chaudron, R., and K. Carlier. 2015. "The advantages of random sampling versus cutting-of-the-tail: the application of a stratified sample design for the collection of data on

special financial institutions in the Netherlands." *IFC Bulletin* 39: 37. Available at: https://EconPapers.repec.Org/RePEc:bis:bisifc:39-37. (accessed July 2021).

Christensen, L.R., and W. Green. 1976. "Economies of Scale in U.S. Electric Power." *Journal of Political Economy* 84(4): 655–676. Available at: https://econpapers.repec.org/RePEc:ucp:jpolec:v:84:y:1976:i:4:p:655-76. (accessed January 2021).

Commission Implementing Regulation (EU) 2020/1197 of 30 July 2020 laying down technical specifications and arrangements pursuant to Regulation (EU) 2019/2152 of the European Parliament and of the Council on European business statistics repealing 10 legal acts in the field of business statistics (Text with EEA relevance) (2020). *Official Journal* L271, 18.8.2020: 1–170. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32020R1197. (accessed July 2021).

Dalén, J. 2005. "Sampling Issues in Business Surveys." *PHARE 2002 Multi Beneficiary Statistics Programme (Lot 1).* Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/20-Quality-in-statistics-sampling-issues-in-business-statistics-2005.pdf. (accessed July 2021).

Eurostat. 2007a. *Impact of a change in threshold on the number of PSIs*. Doc QA/G3/05-07/10 (SIM A/10).

Eurostat. 2007b. *Impact of a change in threshold on data quality at detailed level*. Doc QA/G3/05-07/11 (SIM A/11).

Eurostat. 2018. *European Statistics Code of Practice.* Luxembourg: Publications Office of the European Union. Available at: https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf. (accessed January 2021).

Eurostat. 2020a. *ICT usage in enterprises (isoc_e)*. Available at: https://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm. (accessed January 2021).

Eurostat. 2020b. *Labour costs survey 2008, 2012 and 2016–NACE Rev. 2 activity (lcs_r2).* Available at: https://ec.europa.eu/eurostat/cache/metadata/en/lcs_r2_esms.htm. (accessed January 2021).

Eurostat. 2020c. *Results of the community innovation survey (CIS2-CIS7)*. Available at: https://ec.europa.eu/eurostat/cache/metadata/en/inn_cis2_esms.htm. (accessed January 2021).

Eurostat. 2020d. *Statistics on the production of manufactured goods (prom)*. Available at: https://ec.europa.eu/eurostat/cache/metadata/en/prom_esms.htm. (accessed January 2021).

Eurostat. 2021a. *European business statistics compilers' manual for international trade in goods statistics.* Luxembourg: Publications Office of the European Union. Available at: https://ec.europa.eu/eurostat/documents/3859598/12399711/KS-GQ-21-004-EN-N.pdf. (accessed June 2021).

Eurostat. 2021b. *National requirements for the Intrastat system*. Luxembourg: Publications Office of the European Union. Available at: https://ec.europa.eu/eurostat/documents/3859598/12889594/KS-GQ-21-011-EN-N.pdf. (accessed June 2021).

Fernandes, A.M., P.J. Klenow, S. Meleshchuk, M.D. Pierola, and A. Rodriguez-Clare. 2018. *The Intensive Margin In Trade*. World Bank Policy Research Working Paper. DOI: https://doi.org/10.1596/1813-9450-8625.

Gabaix, X. 2009. "Power Laws in Economics and Finance." *Annual Review of Economics* 1: 255–294. DOI: https://doi.org/10.1146/annurev.economics.050708.142940.

Giesen, D., M. Vella, C.F. Brady, P. Brown, D. Ravindra, and A. Vaasen-Otten. 2018. "Response Burden Management for Establishment Surveys at Four National Statistical Institutes." *Journal of Official Statistics* 34(2): 397–418. DOI: https://doi.org/10.2478/jos-2018-0018.

Guadarrama, M., I. Molina, and Y. Tillé. 2020. "Small area estimation methods under cut-off sampling." *Survey Methodology* 46(1): 51–75. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X202000100004. (accessed July 2021).

Haziza, D., G. Chauvet, and J.C. Deville. 2010. "Sampling and estimation in the presence of cut-off sampling." *Australian & New Zealand Journal of Statistics* 52(3): 303–319. DOI: https://doi.org/10.1111/j.1467-842X.2010.00584.x.

Head, K., T. Mayer, and M. Thoenig. 2014. "Welfare and Trade Without Pareto." *American Economic Review: Papers & Proceedings* 104(5): 310–316. DOI: https://doi.org/10.1257/aer.104.5.310.

Knaub, J.R. 2007. *Cutoff Sampling and Inference*. InterStat. DOI: https://doi.org/10.1111/j.1467-842X.2010.00584.x. (accessed January 2021).

Lavallée, P., and M.A. Hidiroglou. 1988. "On the stratification of skewed populations." *Survey Methodology* 14: 33–43. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198800114602. (accessed July 2021).

Layard, R., and S. Gleister. 1994. "Introduction." In *Cost-Benefit Analysis*, edited by R. Layard, and S. Gleister: 1–56, Cambridge (UK): Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511521942.001.

Lee, S.E., and K.-I. Shin. 2016. "The cut-off point based on underlying distribution and cost function." *Journal of Applied Statistics* 43(6): 1061–1073. DOI: http://dx.doi.org/10.1080/02664763.2015.1089222.

Melitz, M.J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1695–1725. DOI: https://doi.org/10.1111/1468-0262.00467.

Nerlove, M. 1963. "Returns to scale in electricity supply". In *Measurement in Economics-Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld,* by C.F. Christ, M. Friedman, L.A. Goodman, Z. Griliches, A.C. Harberger, Y. Mundlak, M. Nerlove, D. Patinkin, L.G. Telser, and H. Theil: 167–198, Stanford (Calif.): Stanford University Press.

Office for Statistics Regulation. 2020. *The Public Good of Statistics: What we know so far*. London: Office for Statistics Regulation. Available at: https://osr.statisticsauthority.gov.uk/publication/the-public-good-of-statistics-what-we-know-so-far/. (accessed January 2021).

Perše, S. 2016. *Analiza stroskov in koristi moznih razvojnih scenarijev sistema Intrastat v Sloveniji (Cost-Benefit Analysis of Intrastat Development Options in Slovenia).* MSc dissertation. Ljubljana: University of Ljubljana, School of Economics and Business. Available at: http://www.cek.ef.uni-lj.si/magister/perse5068.pdf. (accessed January 2021)

Prest, A.R., and R. Turvey. 1965. "Cost-Benefit Analysis: A Survey." *The Economic Journal* 75(300): 683–735. DOI: http://dx.doi.org/10.2307/2229670.

Regulation (EC) No 638/2004 of the European Parliament and of the Council of 31 March 2004 on Community statistics relating to the trading of goods between Member States and repealing Council Regulation (EEC) No 3330/91 (2004). *Official Journal* L102,

7.4.2004: 1–8. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32004R0638 (accessed July 2021).

Regulation (EC) No 222/2009 of the European Parliament and of the Council of 11 March 2009 amending Regulation (EC) No 638/2004 on Community statistics relating to the trading of goods between Member States (2004). *Official Journal* L87, 31.3.2009: 160–163. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009R0222 (accessed July 2021).

Samuelson, P.A. 1954. "The Pure Theory of Public Expenditure." *Review of Economics and Statistics* 36: 386–389. DOI: https://doi.org/10.2307/1925895.

Smith, P.A., and W. Yung. 2023. "Introduction to sampling and estimation for business surveys." In *Advances in Business Statistics, Methods and Data Collection* edited by G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, K.J. Thompson, and A. van Delden: chap. 27. Hoboken: Wiley.

Spencer, B.D. 1985. "Optimal Data Quality." *Journal of the American Statistical Association* 80(391): 564–573. DOI: https://doi.org/10.1080/01621459.1985.10478155.

Spencer, B.D., and L.E. Moses. 1990. "Needed Data Expenditure for an Ambiguous Decision Problem." *Journal of the American Statistical Association* 85(412): 1099–1104. DOI: https://doi.org/10.1080/01621459.1990.10474981.

Spencer, B.D., and Z.H. Seeskin. 2015. "Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds." In *JSM Proceedings, Government Statistics Section*, August 8–13: 3061–3075. Alexandria, VA: American Statistical Association. Available at: https://www.norc.org/PDFs/Publications/SeeskinZ_Effects%20of%20Census%20Accuracy_2015.pdf. (accessed January 2021).

Stern, N.H. 1976. "On the specification of models of optimum income taxation." *Journal of Public Economics* 6(1–2): 123–162. DOI: https://doi.org/10.1016/0047-2727(76)90044-X.

SURS. 2014. *Survey on current response burden in Intrastat, 2013.* [data collection].

SURS and FURS. 2014. *Intra-EU and Extra-EU Trade Flows, 2013*. [data collection]. Available at: https://www.stat.si/StatWeb/en/DataCollection/Intrastat (accessed January 2021).

Trottini, M., and S.E. Fienberg. 2002. "Modelling User Uncertainty for Disclosure Risk and Data Utility." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5): 511–528. DOI: https://doi.org/10.1142/S0218488502001612.

Yorgason, D., B. Bridgman, Y. Cheng, A.H. Dorfman, J. Lent, Y.K. Liu, J. Miranda, and S. Rumburg. 2011. "Cutoff Sampling in Federal Surveys: an Inter-Agency Review." In *Proceedings of the American Statistical Association, Section on Government Statistics:* 76–90. Available at: https://www.bls.gov/osmr/research-papers/2011/pdf/st110050.pdf. (accessed June 2021).

Received January 2021
Revised July 2021
Accepted September 2021

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1235–1251, http://dx.doi.org/10.2478/JOS-2022-0052

A User-Driven Method for Using Research Products to Empirically Assess Item Importance in National Surveys

Ai Rene Ong[1], *Robert Schultz*[1], *Sofi Sinozich*[2], *Jennifer Sinibaldi*[3], *Brady T West*[1], *James Wagner*[1], *and John Finamore*[3]

Large-scale, nationally representative surveys serve many vital functions, but these surveys can be long and burdensome for respondents. Cutting survey length can help to reduce respondent burden and may improve data quality but removing items from these surveys is not a trivial matter. We propose a method to empirically assess item importance and associated burden in national surveys and guide this decision-making process using different research products produced from such surveys. This method is demonstrated using the Survey of Doctorate Recipients (SDR), a biennial survey administered to individuals with a science, engineering, and health doctorate. We used three main sources of information on the SDR variables: a bibliography of documents using the SDR data as a measure of item use and importance, SDR data table download statistics from the Scientists and Engineers Statistical Data System as an additional measure of item use, and web timing paradata and break-off rates as a measure of burden. Putting this information together, we identified 35 unused items (17% of the survey) and found that the most burdensome items are highly important. We conclude with general recommendations for those hoping to employ similar methodologies in the future.

Key words: Respondent burden; public use data; survey length.

1. Introduction

Large-scale, nationally representative, federally sponsored surveys in the U.S. provide essential statistics for the general population and subpopulations of interest. Some examples of such statistics include unemployment rates (e.g., U.S. Bureau of Labor Statistics 2021), health behaviors (e.g., Schoenborn et al. 2004), and measures of food security (e.g., Coleman-Jensen et al. 2020). These data are important to both government and academic research communities, which use the data to guide policy decisions and conduct secondary analysis. However, with increases in survey nonresponse in all modes of survey data collection (De Heer and De Leeuw 2002; Williams and Brick 2017; Luiten et al. 2020) and recent emphases on alternative data sources, justifying the expense and effort associated with these surveys as well as the burden that they place on the U.S. public

[1] Institute for Social Research, University of Michigan, 426 Thompson St, Ann Arbor, MI 48104 U.S.A. Emails: aireneo@umich.edu, schultro@umich.edu, bwest@umich.edu, and bwest@umich.edu
[2] University of Maryland, College Park, MD 20742, U.S.A. Email: sofi.sinozich@gmail.com
[3] The National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation, 2415 Eisenhower Avenue, Alexandria, Virginia 22314, U.S.A. Emails: jsinibal@nsf.gov and jfinamor@nsf.gov
Acknowledgments: This research was supported by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (Contract #49100420C0020).

requires a close examination of the utility of the data that they gather (see Foundations for Evidence-Based Policymaking Act of 2018).

Close examination of the utility of the data with input from the research communities will help decision-making when modifying the survey, for example, by reducing the survey length. Reducing survey length may also improve data quality (e.g., Deutskens et al. 2004; Galesic and Bosnjak 2009; Peytchev and Peytcheva 2017). However, reducing the length of a questionnaire can be difficult. It is often the case that each question has at least some stakeholders who deem the question as important for official statistics, key indicators, or secondary data analysis. Understanding the relative importance of each question to the user community and prioritizing the survey questions creating those data is a good place to start. Frequency of use is one measure of that importance. To our knowledge, a systematic approach to prioritizing survey questions based on their frequency of use and importance to the user community has not yet been documented.

In this study, we describe and evaluate a method for reducing respondent burden in national surveys by examining how frequently survey items are used in research products, presenting a specific case study based on the Survey of Doctorate Recipients.

2. Method

2.1. Data

The Survey of Doctorate Recipients (SDR) is a longitudinal survey with a fixed panel design, conducted by the National Center for Science and Engineering Statistics (NCSES) biennially since 1973 with individuals under the age of 76 who have obtained a science, engineering, and health doctorate from the NCSES (NCSES, a). The sampling frame for the SDR is the Survey of Earned Doctorates (SED), an annual census of all individuals receiving a research doctorate from an accredited U.S. institution. A new sample of recent doctorate recipients from the previous two academic years is added every wave, and the sample is carried forward until they age out of eligibility.

Prior to 2003, the SDR targeted only individuals residing in the U.S., but it has since expanded to include those recipients of a doctorate from a U.S. institution residing outside of the U.S. Initially, individuals residing outside of the U.S. completed the International Survey of Doctorate Recipients (ISDR), but the ISDR has since been folded into the SDR. The current SDR is administered using three modes: mail, web, or telephone interview. In 2019, a total of 80,882 PhDs responded to the SDR, out of which 75,547 (93%) completed on the web. Reluctant respondents who did not respond to the full instrument in any of the three modes were presented with the critical items only (CIO) instrument at the end of the data collection period, which presents only a subset of the questions deemed to be most critical by SDR managers. Of these 75,547 web respondents, 70,770 (94%) completed the full instrument and the remainder completed the CIO instrument.

The SDR covers topics such as most recent employment, past employment, other work-related experiences (e.g., additional training), recent educational experiences, and demographic information. As the SDR sample is pulled from the SED, some questions asked in the SED that are time-invariant are not asked again in the SDR (e.g., race). The full instrument takes a median time of 18 minutes to complete. There are 79 questions on

the SDR (see the online Supplementary Material for the list of questions), and some of them have "select all that apply" categorical response options that each represent a unique variable in the final data set (e.g., "What were your reasons for not working during the week of February 1, 2019?" had eight response options, each of them being represented in the data as a binary (1 = this is a reason, 0 = not a reason) variable). Taking this into account, we considered every "select all that apply" categorical response option as a separate item, resulting in 202 unique items.

For clarity moving forward, "questions" refers to the survey questions in the SDR (e.g., "What were your reasons for not working during the week of February 1, 2019?"), while "items" refers to variables in the final dataset – that is, either a question in the SDR with a single response option (e.g., "What was the title of the last job you held prior to the week of February 1, 2019?") or a single response option to a "select all that apply" categorical question (e.g., "What were your reasons for taking this postdoc?: Additional training in the Ph.D. field"). These "select all that apply" questions will be referred to as SA questions for the rest of this article.

To measure item importance, we created a data set at the item level. We then used two sources of information to populate this dataset: (1) a bibliography of studies using the SDR data (refer to the online Supplementary Material Bibliography) provided by NCSES, and (2) the SDR website that allows users to download summary data. The bibliography was developed by compiling results from daily alerts on Google Scholar and SCOPUS for articles that mentioned SDR. The bibliography includes written materials such as legislatively mandated Congressional Reports from NCSES, peer-reviewed journal articles, book chapters, newspaper and magazine articles, web blogs, issue briefs, dissertations, working papers, conference proceedings, presentation slides, Federal Register documents, newsletters, and unpublished working papers from the years 1992 to 2020. We reviewed each document and determined which variables were used.

We also used download statistics about the data tables from the Scientists and Engineers Statistical Data System (SESTAT Data Tool: see NCSES (NCSES, b). The SESTAT Data Tool has information from three demographic surveys conducted by NCSES: The National Survey of College Graduates, The National Survey of Recent Graduates, and the Survey of Doctorate Recipients. We were able to acquire download statistics for the SDR in 2017, ISDR in 2010, and ISDR in 2013. These download statistics are a simple count of the number of times a variable from these surveys has been requested from the website. These data are not publicly available and had to be specially requested from NCSES. These download statistics do not include downloads from data repositories other than the SESTAT Data Tool, and we acknowledge this as a limitation to the download statistics we were able to acquire.

We also acknowledge that item usage is just one of many possible measures of importance. For example, some items or groups of items have more impact on policy than others. While this might more accurately measure importance, it is also more subjective and difficult to code reliably. In addition to item use by document, we also record the number of citations each of those documents received as another easily-coded dimension of importance. We then rely on these two simple, easy-to-calculate measures as an efficient and useful method of assessment, particularly for identifying items with no usage.

As a measure of burden, we used the timing paradata and break-off rates from the SDR administered in 2019. We focus on the timing data from the web survey as the vast majority of completed interviews were conducted in this mode. However, combining question timings across modes may be necessary for surveys with lower web response. The timing data are measured in seconds for each page of the web survey. Most of the survey was presented question by question, so this is a fair approximation of the item-level burden. It is not possible to disaggregate the time spent on each item for questions with multiple sub-questions displayed on the same page. Therefore, for the burden measures, these items are assigned an average time (this is calculated using total time for the page divided by the number of items – see the Analysis section for additional details). As for the questions that required respondents to choose their job category (see A7 and A21 in the online Supplementary Material SDR Questions), an average of the average time across the screens was used instead of the sum across screens. Only the 70,770 respondents who started the full questionnaire were used for the analysis.

As with measures of importance, burden could include several dimensions (Bradburn 1978) that our measure may be only partially capturing. For example, some questions may be more stressful to answer. However, rigorous coding of the stress induced by a question is more subjective and difficult and may be better assessed with a different method. We use break-off rates but acknowledge that they may capture both exhaustion with the survey as a whole as well as the burden of specific items, making this difficult to disentangle. Despite being an imperfect measure of subjective burden, question timing does correspond with the federal government's definition of burden (The Paperwork Reduction Act 1995) and breaking off on a particular item has been shown to correlate with the difficulty of the survey question (Peytchev 2009).

2.2. *Coding Procedures*

A total of 105 documents included in the bibliography described above were coded for their use of the SDR data (see online Supplementary Material Bibliography). For the documents where the specific SDR survey question could not be determined, all questions related to the construct were coded as used. Some papers used variables computed from several survey questions. In this case, each item used in the creation of a variable was credited in the coding. As the sample is from the SED, some variables mentioned in these papers were pulled from the SED instead of the SDR (e.g., race, gender). Since these questions were not asked in the SDR, they were not coded. One-off questions that were not part of the regular SDR questions asked every wave were not coded as well (e.g., participation in teamwork, asked in SDR 2006).

As for the frequency of data table downloads, the tables sometimes refer to computed variables that were created based on more than one variable in the SDR. Using the same logic with the coding described above, computed variables were included in the counts for each of the variables used to create it.

As previously mentioned, we coded the number of citations as an additional measure of importance to account for situations where a variable was used very rarely but resulted in extremely influential research products. Some articles have an outsized influence on their field. Therefore, even if the questions themselves were not used very much, if they were used in such high-impact articles, this will still capture another dimension of the importance of these questions. This is a simple tally of the number of citations in Google

Scholar; for example, a document that has been cited ten times will be coded as "10". For the documents that had no citations, this was coded as "0". Using this simple tally, we calculated the average influence of the documents by dividing the total number of citations for those documents that analyzed those questions with the frequency of usage in the coded documents. For example, if a question was used by two documents, and the total citation count for the two documents that had used it was 200, then the mean citation per document is 200/2, which will be interpreted as 100 citations per research product.

These codes were also disaggregated for different stakeholders. We considered Congress, academics, and others (non-academic, non-Congress) to be three different types of stakeholders for these data. The tally of items used in the Congressional reports represents the importance of the items to Congress. The tally of items used in peer-reviewed articles represents the importance of the items to the academic community. All other types of documents (e.g., web blogs) that were coded represent the importance of the items to stakeholders other than Congress and the academic community.

As for the timing paradata, we top-coded outlier durations to the 95th percentile for each item, as these outliers could be due to the respondents timing out or leaving the screen for an extended period of time for other reasons unrelated to the question burden. As previously mentioned, the timing paradata and break-off rates are for each web page, and there may be multiple items per page. The average time taken to answer these items was calculated as an average of the time spent on the page across the items on the page. For example, if the time taken on the page is 60 seconds, and there are four items on that page, we consider each item to take 15 seconds to answer.

2.3. Analysis

Our analysis is entirely descriptive, presenting frequencies of the various metrics of influence and burden that we coded. For burden, we plot time per question and (separately) break-off to identify burdensome questions. When making conclusions about influence, we pay particular attention to variables with no or low use in the publications analyzed. We also report the average influence, based on the number of additional citations generated by articles using the item, and compare this to the burden. Finally, we combine these measures and examine differences across stakeholder communities to make a statement about the overall value of each item.

3. Results

Looking first at the timing paradata to understand burden, the median time taken per page overall is 23.92 seconds (min = 3.94 seconds, max = 96.35 seconds), and the median time taken per item is 4.36 seconds (min = 1.37 seconds, max = 96.35 seconds). The item that took the longest average time to answer, A20, was an open-ended question asking the respondents to describe the duties and responsibilities of their last job. A total of 13% (n = 9,295) of the respondents ever broke off across all the pages when responding to the SDR; this includes break-offs on the instruction pages. The highest break-off rate, 5%, is on item A30 which is a forced choice question with fourteen options asking about the work activities at the respondent's principal job. Figure 1 illustrates the average time taken and the break-off rate for each item across all the respondents.

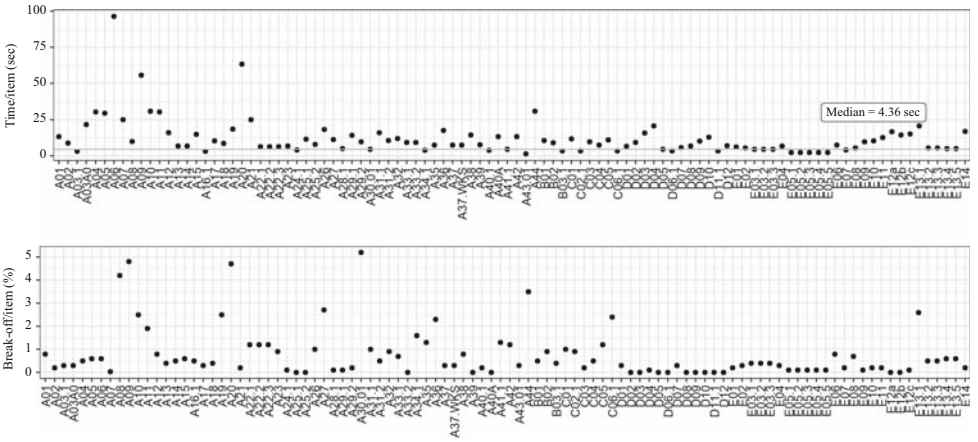


Fig. 1. Average time taken (sec) and break-off rate (%) on each page, with a horizontal grey line indicating the median average time.

Regarding item importance, we first examine which items were not used at all (see Appendix, Subsection 5.2, for an overview of usage and the average citation ratios for all the items). A total of 17% (n = 35) of the SDR items had not been used at all among the documents coded, nor had they been downloaded from the website. The items are spread across the different sections; most of them are from the SA (select all) questions in Section C and D pertaining to extra training (C2), furthering education (D6), and taking courses (D11) (see Appendix, Subsection 5.1). Taking into account the burden of these 35 items, the total average time taken to answer these items was 3.28 minutes. They also had a low break-off rate, with the highest being only 1.0% (n = 95).

Figure 2 illustrates the break-off rate and associated time taken for each item that was never used in the materials we analyzed. The SA questions were displayed on the same page and therefore have the same average time taken and break-off rate. These SA questions are represented by the first item in the series (e.g., C02.1 to C02.7 were represented by C02.1). Most of the SA questions, that is, C02.1–C02.7, D06.1–D06.9, and D11.1–D11.9, were answered relatively quickly, being under the median average time across all items. However, a few items stood out as potentially more burdensome than the rest. Item A33.2, which asked the respondents for the number of people they supervise indirectly, and E14, the earliest age of experiencing any functional difficulties (e.g., seeing, hearing, walking), took a much longer average time to answer than the rest of the unused items.

Besides looking at the unused questions, we can examine the questions that were only used once across all the research products. Of all the items (n = 29) were 14% only used once, though some of them were used in the most influential articles. For example, the article that used the variable "the second most important reason for taking a postdoc position" was cited 180 times (see Corley and Sabharwal 2007). Table 1 indicates the items that were used once across all the research products with no citations. This means that these items were only used in an analysis once or had a single data table download request. The total mean time taken to answer the questions that contain these items is 1.31 minutes.

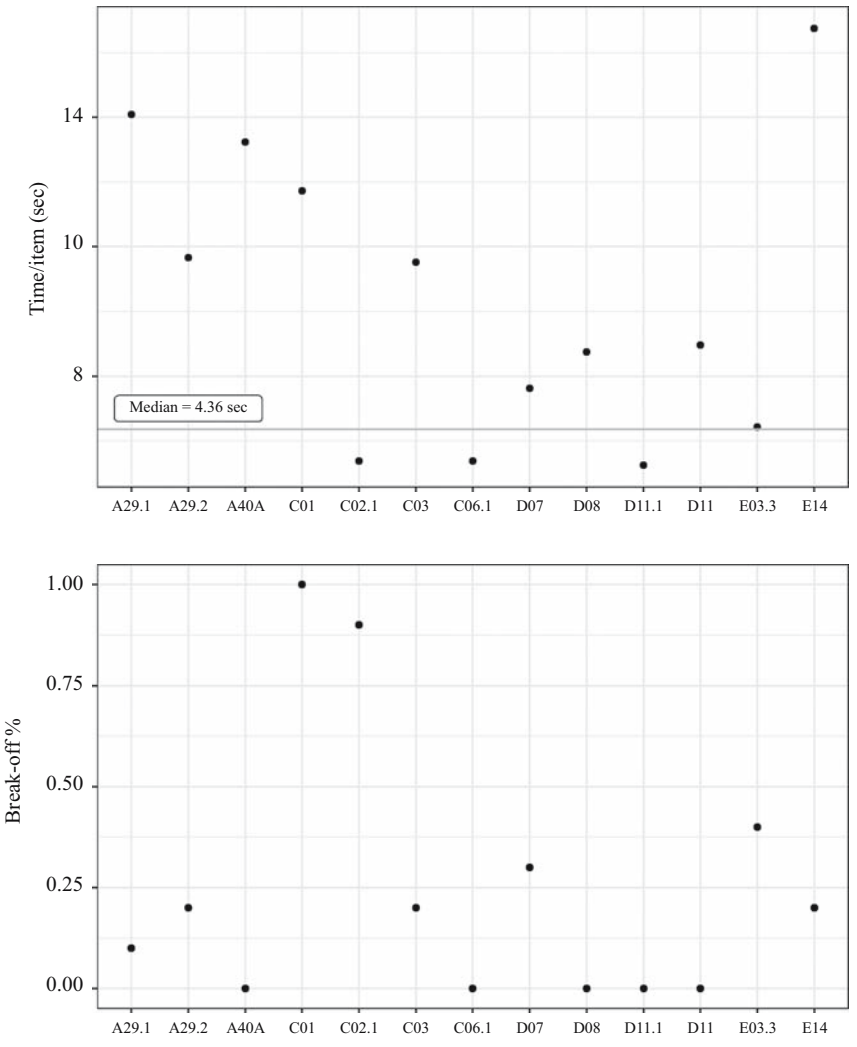


Fig. 2. Time taken and break-off rates for items that did not appear in the literature.

In Table 1, the SA question regarding important factors for consideration when it comes to the respondent's job (C06) is the most burdensome, taking 29.7 seconds to answer the whole grid with a break-off rate of 2.4% (n = 222). That SA question consisted of nine items, but four items (i.e., salary, opportunities for advancement, intellectual challenge, and contributions to society) were not displayed on the table as they were used once more than the rest. Overall, these items were not used very often, and even when they were used, the research products using these questions were rarely cited.

We now turn our attention to the other most burdensome items in the SDR. We start by looking at items taking 20 seconds or longer on average (the 95th percentile for the

Table 1. Question burden for questions that were used once across all research products (with no citations).

Item	Question topic	Time (sec)	Break-off (%)
A22.2	*Job requires technical expertise of a bachelor's degree or higher in the social sciences	6.18	1.20
A22.3	*Job requires technical expertise of a bachelor's degree or higher in some other field	6.18	1.20
C06.2	*Important to job: Benefits	3.30	2.40
C06.3	*Important to job: Job security	3.30	2.40
C06.4	*Important to job: Job location	3.30	2.40
C06.7	*Important to job: Level of responsibility	3.30	2.40
C06.8	*Important to job: Independence	3.30	2.40
D02	Type of degree earned (additional degree)	9.06	0.00
D03	Primary field of study of degree (additional degree)	15.79	0.00

* These were displayed on the same page, and therefore share the same average time and break-off rate

average time taken to answer an item, which is 20.74 seconds). There were 12 such items (presented in Table 2) which took a total average time of 7.64 minutes to answer. The mean citation ratio across these 12 items is 44.28. Section A has the most "high burden" questions and is the largest section of the SDR. Section A covers questions about current and recent employment and collects details about the respondent's principal job. The items with the highest burden were either frequently used or were used in articles that were generally well-cited, as evident by the citation ratios in Table 2. These items are the most important items in SDR to NCSES (personal communication).

The two items that took the longest time to answer, A06 and A20, are open-ended items asking respondents to describe their duties and responsibilities in their current job or last job held. Respondents spent at least a minute to answer them. Though burdensome, they were also of great interest to the academic stakeholders. NCSES uses the responses to correctly classify the respondent's job in the taxonomy of occupations in tandem with the responses provided in A07 and A21, which ask respondents to self-classify their job category. These four items were important to the academic community and in preparing Congressional reports. For these reasons, the high burden of A06 and A20 seems justified. The break-off rate of 4.7% for A20, however, is quite high when compared with the break-off rates for the other items in the SDR.

Looking at differences in item importance across stakeholder groups, though E13.1 was not used by academic stakeholders, it appears in Congressionally-mandated reports. If the usage counts were examined without considering these different stakeholders, its importance in the SDR would not be clear. Disaggregating item usage by stakeholders enables a clearer assessment of item importance. However, it must be mentioned that item E13.1 was displayed on the same page as the instructions to this SA question. This would have artificially inflated the time on the page for all the respondents, as they would have had to read the instructions before answering. By comparison, E13.2 to E13.5 took much less time on average (about ten seconds each).

Table 2. Item importance for the most burdensome questions by stakeholders.

Item	Question topic	Frequency of use (all documents)	Citation ratio	Frequency of use (academic documents)	Frequency of use (congressional reports)	Frequency of use (others documents)	Data table download	Time (sec)	Break-off (%)
A04	Date last worked for pay or profit	3	12.00	2	0	1	3	30.37	0.50
A05	Title of last job held	3	17.67	3	0	0	0	29.4	0.60
A06	Duties and responsibilities in the last job	2	26.50	2	0	0	0	96.35	0.60
A07	Job category for last job	3	16.00	2	0	1	0	25.00	0.03
A09	Principal employer in the week of February 1, 2019	20	53.10	17	0	3	24	55.80	4.80
A10	No. of people who worked for principal employer	6	72.17	3	0	3	0	30.26	1.90
A13	Description of principal employer	4	70.00	2	0	2	0	30.81	2.50
A20	Duties and responsibilities in the job	14	61.07	11	0	3	0	63.37	4.70
A21	Job category code for current job	27	72.96	17	2	8	33	25.00	0.20
A44	Total earned income in 2018	12	109.83	8	0	4	6	30.70	3.50
D04	Month and year degree awarded	2	2.00	1	1	0	0	20.74	0.10
E13.1	Functional limitation – Seeing	2	18.00	0	2	0	9	20.62	2.60

4. Discussion

The case study presented in this article quantifies concepts of importance and burden based on how respondents and data users interact with survey questions. Through simple descriptive analyses of item use, citations related to items, timing paradata, and break-off rates, we were able to identify potential questions and items for survey redesign. We identified a few items not used by any of the stakeholders. Many of these items pertain to post-PhD training, furthering education, and taking courses. NCSES could consider dropping some of these questions or simplifying them, especially since none of these items are on the critical items only (CIO) version of the SDR. The lesser-used items with no citations were about the importance of different job factors (e.g., benefits, job security), type of degree, and primary field of study of the respondent's new degree. As mentioned earlier, the "important factors to a job" question (C06) is an SA question, with some of the items within the question being used more than others. An example of a recommendation here would be to collapse the less-used categories.

When we examined usage across the most burdensome items, we noted that some (i.e., duties and responsibilities on the job, principal employer) are also items that are important to stakeholders of the SDR. Item A20 is a good example. Although A20 has a relatively high break-off rate of 4.7%, it is a very important item for the academic community. It is also a contributor to the quality of item A21, which is important across all stakeholders. Given this, it is important to preserve this question and the example illustrates the importance of combining burden and importance measures.

Solutions for important but burdensome questions would be to revise the questions to motivate respondents and reduce break-offs. For example, A20 could include wording such as "Your answer to this question is very important in order to correctly classify your job category." Another recommendation would be to administer these questions selectively, for example, to ask A21 of respondents who have changed jobs in the two years from the last survey. For respondents who stayed in the same job, SDR could ask to confirm that their duties are still the same. Changes such as these could be tested experimentally with a small sample first to ascertain whether they do make a difference for break-off rates before committing to the revision for the next SDR.

We also encourage users of our method to be sensitive to the time limitations of the data. Although post-PhD training does not appear at all in the SDR research products, it might be that these items are important for understanding an emerging topic. That is, post-PhD training might be of particular interest currently. If so, it might be that our measures of usage lag behind current developments as new publications and other documents may soon emerge based upon these items. Consistent monitoring of changing trends will help shed light on emerging interest areas.

Though we would not want to dictate how users of this method prioritize the different metrics of item usage and burden, a concrete example of how one might use this method might be to first prioritize the removal of items that have low usage (e.g., one use or less). In this case, the 35 unused items and 29 items used only once across all the research products are candidates for removal. One could then check if any of these items are used in important (or mandated) reporting and retain all such items. Following the importance assessment, a user of this method might check the burden associated with the remaining items and weigh

the utility of these items against the burden associated with them. If they are burdensome but have seen heavy usage among stakeholders, one might consider moving these items so that they will be displayed early in the survey when the respondents are not as fatigued. We acknowledge that the specific choices that might be made to modify a questionnaire are highly context dependent. We suggest these simply to demonstrate how they might be used.

We highlight that our quantitative measures of utility should not be the only factor in making design decisions. In this analysis, we chose to integrate importance to various stakeholder groups to curtail short-sighted redesign suggestions. For example, our analysis determined that the questions on the type of degree earned and primary field of study for post-PhD degrees are not very important. However, the answers to these questions are used by NCSES to update information on the most recent degree in the SESTAT database, and data users might be using them to filter their analysis samples. These items are extremely important to multiple stakeholders, including the administrators of the survey, despite being missed in the coding of these research products.

Our method is heavily dependent on having good data on variable use and the availability of timing and break-off paradata. In the illustrative example above, our analysis was restricted to the respondents who completed SDR in the web mode. We note that the question burden might be different in other modes. However, given that most respondents responded in the web mode, this measurement of burden would apply to a majority of the SDR panel members, but this may be different for other surveys. Though it is easy to obtain timing and break-off paradata for web surveys, it is more challenging (and maybe impossible) for other modes of survey administration. Also, combining timing data across modes might be complicated since each mode functions differently (e.g., whether it is interviewer-administered vs. self-administered). If timing paradata are not available or accurate, other measures of burden might need to be considered for other survey modes.

As for having good data on variable use, we had access to a bibliography of articles mentioning the use of the SDR, the data table requests from the SESTAT Data Tool, and the Congressional reports. Given that all these data sources were readily available to our research team, this method was straightforward to implement. In the absence of such data, the survey researcher implementing this method will have to conduct a systematic search in Google Scholar or in one of the existing databases (e.g., JSTOR, Web of Science). We would recommend more generally that large survey programs dedicate resources to maintaining these types of bibliographies to assist with the type of study presented here.

This method also requires manual coding of the frequency of use of the variables in a survey. For longer surveys with a wider scope for article inclusion, this might be a burdensome task. One way to mitigate this problem is to focus on specific sections of a survey and code those questions. If one has some intuition as to which questions should be considered for removal, this is a way to systematically confirm that. It would also be possible to draw a random sample of documents to be coded.

Coding the documents was challenging when the documents themselves were not explicit about the SDR variables used in their analysis. We erred on the side of caution and coded more generally when we encountered such documents. For example, "family-related variables" were mentioned in Moguérou (2002). It is unclear what "family-related variables" encompasses, and the only family-related variable mentioned was "number of

children." To err on the side of caution, marital status was coded as used as well. This meant that the overall SDR question usage could be an overestimation.

With these caveats in mind, our proposed method can also be used to inform a modular design, which is a survey design that splits up a survey into several modules to be administered at different times to the same respondents. A modular design might use information about use, utility, and burden to plan which set of questions to include in each module. The literature on an optimum design for modular surveys (e.g., the number of modules and the number of questions in a module) is still developing (e.g., West et al. 2015; Toepoel and Lugtig 2018; Andreadis and Kartsounidou 2020; Peytchev et al. 2020). Using our approach, which has the advantage of generalizability across surveys and easy interpretability, as an input to the process of designing modules would be one productive avenue for research. Statistical considerations aside, one can envision a modular design where the questions are split into modules based on their importance and usage (e.g., most well-used or highly-cited items in the first module). Or perhaps, the most burdensome items could be split across modules so that they appear earlier in each module when respondents are less fatigued.

Further research on this topic could include a more robust method of measuring burden. Timing and break-offs are imperfect measures of burden, and burden could encompass many other things such as respondent effort, discomfort, and stress (Bradburn 1978). Incorporating these other measures of burden could paint a different picture of what questions are burdensome. Furthermore, we did not account for possible differences between new panel members and existing panel members. It is possible that new panel members might find the questions more burdensome compared to existing panel members, and therefore the strategies to reduce burden will have to be tailored based on different segments of the sample. Besides that, further research could also include methods for assigning differential weights to uses of variables. For our analysis, we relied on simple counts of how often variables were used. We could, for example, assign higher weights to the usage of a variable in the SDR survey for Congressionally-mandated reports. Since Congressionally-mandated variables are included in federally-sponsored surveys, it is unlikely that these variables can be removed. Future work could also assign greater weight to variables that are used in more recent publications to account for the changing needs of data users.

5. Appendix

5.1. Burden Associated with Unused Questions

Item	Question topic	Avg. time (sec)	Break-off (%)
A29.1	Most important reason for working in an area outside the field of your first U.S. doctoral degree	14.1	0.1
A29.2	Second most important reason for working in an area outside the field of your first U.S. doctoral degree	9.7	0.2

5.1. Continued.

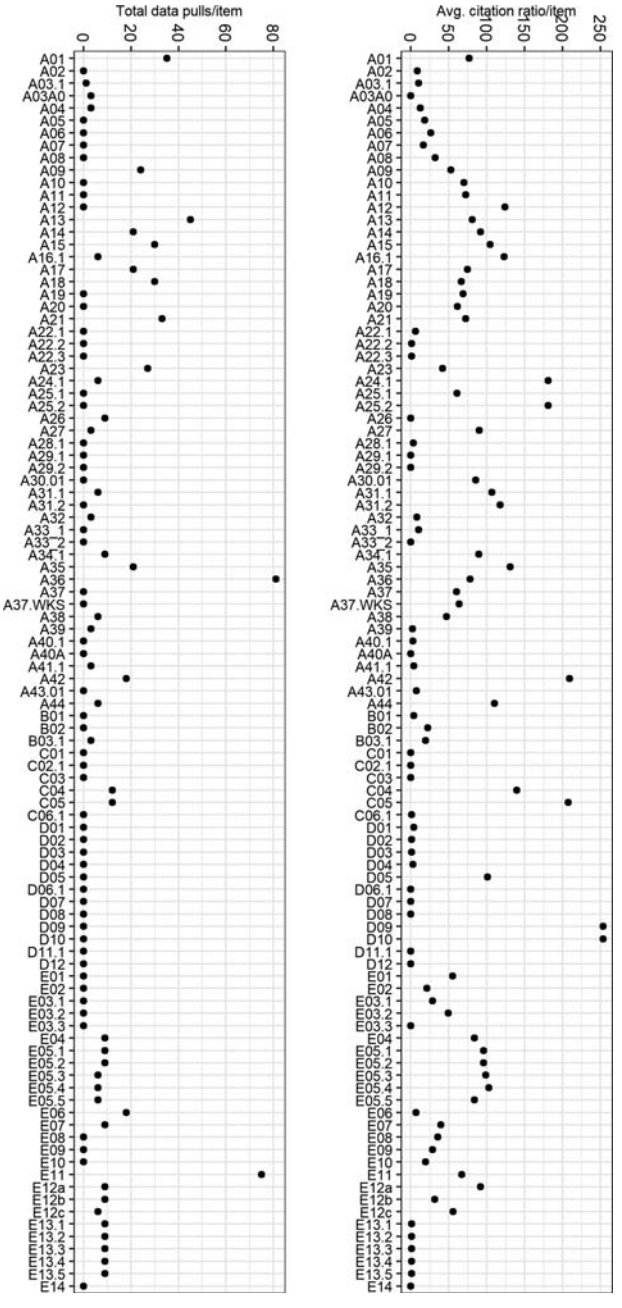
Item	Question topic	Avg. time (sec)	Break-off (%)
A33.2	No. of people supervised indirectly	18.2	0
A40A	Year retired	13.2	0
C01	Attend any work-related training	11.7	1.0
C02.1	*Reasons for training: Improve skill/knowledge	3.4	0.9
C02.2	*Reasons for training: Increase opportunities	3.4	0.9
C02.3	*Reasons for training: Licensure/certification	3.4	0.9
C02.4	*Reasons for training: Change to a different field	3.4	0.9
C02.5	*Reasons for training: Expected by employer	3.4	0.9
C02.6	*Reasons for training: Other	3.4	0.9
C02.7	*Reasons for training: Personal interest	3.4	0.9
C03	Most important reason for taking training	9.5	0.2
D06.1	*Reasons for most recent degree: Further education before career	3.4	0
D06.2	*Reasons for most recent degree: Prepare for graduate school	3.4	0
D06.3	*Reasons for most recent degree: Change field	3.4	0
D06.4	*Reasons for most recent degree: Gain skills	3.4	0
D06.5	*Reasons for most recent degree: Licensure/certification	3.4	0
D06.6	*Reasons for most recent degree: Promotion/salary	3.4	0
D06.7	*Reasons for most recent degree: Required by employer	3.4	0
D06.8	*Reasons for most recent degree: Personal interest	3.4	0
D06.9	*Reasons for most recent degree: Other	3.4	0
D07	Enrolled in a college/taking courses on the week of Feb 1, 2019	5.6	0.3
D08	Enrolled full-time/part-time/not enrolled but taking courses	6.8	0
D11.1	*Reasons for taking course: Further education before career	3.3	0
D11.2	*Reasons for taking course: Prepare for graduate school	3.3	0
D11.3	*Reasons for taking course: Change field	3.3	0

5.1. Continued.

Item	Question topic	Avg. time (sec)	Break-off (%)
D11.4	*Reasons for taking course: Gain skills	3.3	0
D11.5	*Reasons for taking course: Licensure/certification	3.3	0
D11.6	*Reasons for taking course: Promotion/salary	3.3	0
D11.8	*Reasons for taking course: Personal interest	3.3	0
D11.9	*Reasons for taking course: Other	3.3	0
D12	School-related costs paid for by an employer	7.0	0
E03.3	Partner's duties on job require technical expertise of a bachelor's degree and above	4.4	0.4
E14	Earliest age of difficulty	16.8	0.2

∗ these were displayed on the same page, and therefore share the same break-off rate; the time on these pages is averaged across the number of items

5.2. Overview of Usage and the Average Citation Ratios for all the Items



6. References

Andreadis, I., and E., Kartsounidou. 2020. "The impact of splitting a long online questionnaire on data quality." *Survey Research Methods* 14 (1): 31–42. DOI: https://doi.org/10.18148/srm/2020.v14i1.7294.

Bradburn, N. 1978. "Respondent Burden." In Proceedings of the Survey Research Methods Section of the American Statistical Association, American Statistical Association Alexandria, VA. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf (accessed December 2020).

Coleman-Jensen, Al., M.P. Rabbitt, P. Matthew, C.G. Gregory, and A. Singh. 2020. "Household Food Security in the United States in 2019". ERR-275, U.S. Department of Agriculture, Economic Research Service. Available at: https://www.ers.usda.gov/publications/pub-details/?pubid=99281 (accessed December 2020).

Corley, E.A., and M. Sabharwal. 2007. "Foreign-born academic scientists and engineers: producing more and getting less than their US-born peers?" *Research in Higher Education* 48 (8): 909–940. DOI: https://doi.org/10.1007/s11162-007-9055-6.

De Heer, W., and E. de Leeuw. 2002. "Trends in household survey nonresponse: A longitudinal and international comparison." *Survey nonresponse* 41: 41–54. New York: Wiley.

Deutskens, E., K. de Ruyter, M. Wetzels, and P. Oosterveld. 2004. "Response rate and response quality of internet-based surveys: An experimental study." *Marketing Letters* 15(1): 21–36. Available at: https://www.jstor.org/stable/40216512 (accessed December 2020).

Foundations for Evidence-Based Policymaking Act of 2018. Available at: https://www.congress.gov/bill/115th-congress/house-bill/4174 (accessed January 2021).

Galesic, M., and M. Bosnjak. 2009. "Effects of questionnaire length on participation and indicators of response quality in a web survey." *Public Opinion Quarterly* 73(2): 349–360. DOI: https://doi.org/ 10.1093/poq/nfp031.

Luiten, A., J. Hox, and E. de Leeuw. 2020. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys", *Journal of Official Statistics* 36(3): 469–487. DOI: https://doi.org/10.2478/jos-2020-0025.

Moguérou, P. 2002. "Job satisfaction among US Ph. D. graduates: the effects of gender and employment sector." *Labor and Demography* EconWPA. Available at: https://econpapers.repec.org/paper/wpawuwpla/0204002.htm (accessed December 2020).

NCSES (Center for Science and Engineering Statistics), a. "Survey of Doctorate Recipients." Available at: https://www.nsf.gov/statistics/srvydoctoratework/ (accessed December 2020).

NCSES (Center for Science and Engineering Statistics), b. "Scientists and Engineers Statistical Data System Data Tool." Available at: https://ncsesdata.nsf.gov/sestat/ (accessed December 2020).

Peytchev, A. 2009. "Survey break-off." *Public Opinion Quarterly* 73(1): 74–97. DOI: https://doi.org/ 10.1093/poq/nfp014.

Peytchev, A., and E. Peytcheva. 2017. "Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach." *Survey Research Methods* 11(4): 361–368. DOI: 10.18148/srm/2017.v11i4.7145.

Peytchev, A., E. Peytcheva, J.G. Conzelmann, A. Wilson, and J. Wine. 2020. "Modular survey design: Experimental manipulation of survey length and monetary incentive structure." *Journal of Survey Statistics and Methodology* 8(2): 370–384. DOI: 10.1093/jssam/smz006.

Schoenborn, C.A., P.F. Adams, P M. Barnes, J.L. Vickerie, and J S. Schiller. 2004. "Health behaviors of adults; United States, 1999–2001." Available at: https://stacks.cdc.gov/view/cdc/7007 (accessed December 2020).

The Paperwork Reduction Act. 1995. 44 U.S.C. chap. 1 § 101; 44 U.S.C. chap. 35, subchap. I § 3501 et seq. Available at: https://uscode.house.gov/view.xhtml?path = /prelim@title44/chapter35&edition = prelim (accessed December 2020).

Toepoel, V., and P. Lugtig. 2018 "Modularization in an Era of Mobile Web: Investigating the Effects of Cutting a Survey Into Smaller Pieces on Data Quality." *Social Science Computer Review* 40 (1). DOI: https://doi.org/10.1177/0894439318784882.

U.S. Bureau of Labor Statistics. 2021. *The employment situation – December 2020*. Available at: https://www.bls.gov/news.release/pdf/empsit.pdf (accessed Jan 2021).

West, B.T., D. Ghimire, and W. Axinn. 2015. "Evaluating a Modular Design Approach to Collecting Survey Data Using Text Messages." *Survey Research Methods* 9 (2): 111–123. DOI: https://doi.org/ 10.18148/srm/2015.v9i2.6135.

Williams, D, and J.M. Brick. 2018. "Trends in US face-to-face household survey nonresponse and level of effort." *Journal of Survey Statistics and Methodology* 6 (2) 2: 186–211. DOI: https://doi.org/ 10.1093/jssam/smx019.

Received March 2021
Revised July 2021
Accepted January 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1253–1257, http://dx.doi.org/10.2478/JOS-2022-0053

Editorial Collaborators

The editors wish to thank the following referees and guest editors of theme issues who have generously given their time and skills to the Journal of Official Statistics during the period 1 October 2021 to 30 September 2022. An asterisk indicates that the referee served more than once during the period.

Adiguzel, Feray, Erasmus University Rotterdam, Rotterdam, the Netherlands
Angelico, Cristina, Bank of Italy, Rome, Italy
Arias-Salazar, Free University of Berlin, Berlin, Germany
Arora, Sanjay, Ernst and Young LLP, Washington D.C., U.S.A.
Ashmead, Robert, Ohio Colleges of Medicine, Columbus, Ohio, U.S.A.
Astill, Gregory, U.S. Department of Agriculture, Kansas City, Missouri, U.S.A.
Awan, Jordan, Purdue University, West Lafayette, U.S.A.
Bacchini, Fabio, Italian National Institute of Statistics, Rome, Italy*
Bakar, Shuvo, University of Sydney, Sydney, Australia
Baker, Kellan, Johns Hopkins University, Baltimore, Maryland, U.S.A.
Balk, Bert, Amersfoort, the Netherlands*
Bauder, Donald, U.S. Census Bureau, Washington D.C., U.S.A.
Beatty, Paul, U.S. Census Bureau, Washington D.C., U.S.A.*
Bell, William, U.S. Census Bureau, Washington D.C., U.S.A.*
Benassi, Federico, Italian National Institute of Statistics, Rome, Italy*
Benedetti, Ilaria, University of Tuscia, Viterbo, Italy
Beręsewicz, Maciej, Poznań University of Economics and Business, Wielkopolska, Poland*
Berg, Emily, Iowa State University of Science and Technology, Ames, Iowa, U.S.A.*
Berger, Yves, University of Southampton, Southampton, UK*
Bethlehem, Jelke, Leiden University, Hazerswoude-Rijndijk, the Netherlands
Bivand, Roger, Norwegian School of Economics, Bergen, Norway
Bollinger, Christopher, University of Kentucky, Lexington, Kentucky U.S.A.
Bonnéry, Daniel, University of Cambridge, Cambridge, UK
Boyle, John, ICF International Inc., Rockville, Maryland, U.S.A.*
Brick, Michael, Westat, Rockville, Maryland, U.S.A.*
Bryant, John, Bayesian Demography Limited, Russley, Christchurch, New Zealand
Cantor, David, Westat, Rockville, Maryland, U.S.A.
Carstensen, Johann, German Centre for Higher Education Research and Science Studies, Hannover, Germany*
Chaudhuri, Sanjay, National University of Singapore, Singapore.*
Chen, Lu, National Institute of Statistical Sciences, Washington D.C., U.S.A.*
Choi, InKyung, UNECE, Geneva, Switzerland
Christoph, Bernhard, Institute for Employment Research, Nuremberg, Germany

Hilton, Jason, University of Southampton, Southampton, UK
Holzberg, Jessica, U.S. Census Bureau, Washington D.C., U.S.A.
Honchar, Oksana, Australian Bureau of Statistics, Sydney, Australia
Inoue, Toshikatsu, Hitotsubashi Daigaku – Institute of Economic Research, Tokyo, Japan
Janick, Ryan, U.S. Census Bureau, Washington D.C., U.S.A.
Japec, Lilli, Statistics Sweden, Solna, Sweden
Jauslin, Raphael, University of Neuchâtel, Neuchâtel, Switzerland
Jin, Haomiao, University of Southern California, Los Angeles, California, U.S.A.
Kaplan, Robin, U.S. Bureau of Labor Statistics, Washington D.C., U.S.A.
Kapteyn, Arie, University of Southern California, Los Angeles, Califonia, U.S.A.
Karlberg, Forough, Statisics Luxembourg, Luxembourg
Keller, Andrew, U.S. Census Bureau, Washington D.C., U.S.A.
Kenett, Ron, KPA, Raanana, Israel
Kennedy, Courtney, Pew Research Center, Washington D.C., U.S.A.*
Kern, Christoph, University of Mannheim, Mannheim, Germany
Kim, Brian, University of Maryland, College Park, Maryland, U.S.A.
Kim, Jae-kwang, Iowa State University, Ames, Iowa, U.S.A.
Knottnerus, Paul, Statistics Netherlands, the Netherlands
Krieg, Sabine, Statistics Netherlands, Heerlen, the Netherlands
Lamboray, Claude, Eurostat, Luxembourg, Luxembourg
Lanzieri, Giampaolo, Eurostat, Luxembourg, Luxembourg
Larsen, Michael, George Washington University, Rockville, Maryland, U.S.A.
Laureti, Tiziana, University of Tuscia, Viterbo, Italy
LeClere, Felicia, University of Chicago, Chicago, Illinois, U.S.A.*
Legleye, Stephane, Inserm CESP, Villejuif, France
Lehtonen, Risto, University of Helsinki, Helsinki, Finland*
Liseo, Brunero, University of Rome, Rome, Italy
Little, Roderick, University of Michigan, Ann Arbor, U.S.A.*
Lockhart, Richard, Simon Fraser University, Burnaby, British Columbia, Canada*
Loong, Bronwyn, Australian National University Canberra, Australia
Luna-Hernandez, Angela, University of Southampton, Southampton, UK
Malmros, Jens, Stockholm University, Stockholm, Sweden*
Marchetti, Stefano, University of Pisa, Pisa, Italy
Mariyah, Siti, STIS Polytechnic of Statistics, Jarkarta, Indonesia
Maślankowski, Jacek, University of Gdańsk, Sopot, Poland
Mavletova, Aigul, National Research University Higher School of Economics, Mocow, Russia*
McClain, Colleen, University of Michigan, Ann Arbor, Michigan, U.S.A.
McConway, Kevin, Open University, Milton Keynes, UK*
McGonagle, Katherine, University of Michigan, Ann Arbor, Michigan, U.S.A.
Mehrhoff, Jens, Deutsche Bundesbank, Frankfurt am Main, Germany.
Merfeld, Joshua, KDI School of Public Policy and Management, Sejong City, Republic of Korea*
Meyer, Ilan, University of California, Los Angeles, California, U.S.A.
Mosaferi, Sepideh, Iowa State University, Ames, Iowa, U.S.A.

Van der Laan, Jan, Statistics Netherlands, The Hague, the Netherlands
Van Riper, David, Minnesota Population Center, Minneapolis, Minnesota, U.S.A.
Von Auer, Ludwig, University of Trier, Trier, Germany*
Wakefield, Jon, University of Washington, Seattle, Washington, U.S.A.*
Westling, Sara, Statistics Sweden, Örebro, Sweden
White, Gentry, Queensland University of Technology, Brisbane, Australia
Williams, Douglas, U.S. Bureau of Labor Statistics, Washington D.C., U.S.A.
Williams, Matthew, RTI International, Research Triangle Park, North Carolina, U.S.A.
Wu, Changbao, University of Waterloo, Waterloo, Ontario, Canada
Würz, Nora, Free University of Berlin, Berlin, Germany*
You, Yong, Statistics Canada, Ottawa, Canada
Zabala, Felipa, Statistics New Zealand, Wellington, New Zealand
Zamanzade, Ehsan, University of Isfahan, Isfahan, Iran
Zevi, Giordano, bank of Italy, Rome, Italy
Zhang, Chan, University of Michigan, Ann Arbor, Michigan, U.S.A.
Zhang, Li-Chun, University of Southampton, Southampton, UK
Zhang, Mark (Xichuan), Australian Bureau of Statistics, Belconnen, Australia*
Zult, Daan, Statistics Netherlands, The Hague, the Netherlands

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1259–1262, http://dx.doi.org/10.2478/jos-2022-0054

Index to Volume 38, 2022

Contents of Volume 38, Numbers 1–4

Author Index

Book Reviews