June 19, 2018

NOTE TO THE REVIEWER OF:	OMB CLEARANCE 1220-0141 "Cognitive and Psychological Research"
FROM:	Robin Kaplan Office of Survey Methods Research
SUBJECT:	Submission of Materials for Subjective Burden Research

Please accept the enclosed materials for approval under the OMB clearance package 1220-0141 "Cognitive and Psychological Research." In accordance with our agreement with OMB, we are submitting a brief description of the study.

The total estimated respondent burden for this study is 426 hours.

If there are any questions regarding this project, please contact Robin Kaplan at 202-691-7383.

I. Introduction

Respondent burden in the federal government is often defined as the time it takes to complete a survey, which can include survey length, number of questions, as well as the time spent reading complex instructions, gathering and entering data, and reviewing it (Paperwork Reduction Act Guide, 2011). These are considered "objective" measures of respondent burden - they exist independently of respondents' perceptions of the survey. But another aspect of burden has to do with participants' subjective appraisals of the survey, such as level of effort, interest, or sensitivity of the questions (e.g., Fricker, Yan, & Tsai, 2014; Bradburn, 1978). Many federal surveys, including the American Community Survey (ACS) (e.g., Holzberg et al., 2018) and the Consumer Expenditure (CE) survey (e.g., Fricker et al., 2014) have been exploring measurement of subjective burden and how it may affect response rates and data quality, but subjective burden remains an understudied topic in federal surveys. In addition, many federal surveys also ask for proxy reports, including the Current Population Survey (CPS), CE, and ACS, where one member of the household reports for other household members, but even less is known about the impact of proxy-reporting on burden, as well as the potential burden imposed by the contact attempts of interviewers and the mailings sent out trying to reach respondents to participate in surveys.

Prior research shows that objective measures of burden are related to response rates (e.g., Bradburn, 1978; Rolstad et al., 2011; Crawford et al., 2001). Subjective measures of burden have been linked to data quality, attrition rates in longitudinal surveys, and feelings of survey fatigue (e.g., Rolstad et al., 2011; Fricker et al., 2014). Thus, both objective and subjective measures of burden may have effects on survey outcomes, but few studies have explored both types of burden in a single study to better understand the unique contributions each may have. As such, the goals of this research are as follows:

- To better understand respondents' experience of objective and subjective levels of burden when reporting for themselves and reporting for others as a proxy.

- How burden relates to data quality, attrition rates, and memory for burden over time. Memory for how burdensome a survey may be a more important predictive factor of future participation in federal surveys than the actual burden experienced at the time of taking the survey (Wirtz et al., 2003), but no research to our knowledge has assessed memory for burden over time in a longitudinal survey and its impact on response rates. With response rates declining (Miller, 2017), it is critical to better understand what causes respondent burden and how to ease respondent burden in the future, an important factor to both the mission of BLS and OMB.

- To explore the impact of respondents' interactions with the survey agent (e.g., number of contacts, frequency of the contact, an interviewer coming to your residence versus calling you, and the number of mailings received. Using vignettes (Alexander & Becker, 1978; Atzmüller & Steiner, 2010), an effective way to assess respondents' reactions to hypothetical survey situations, we will aim to assess how issues of contact may affect burden.

- To explore how survey paradata that reflect objective burden, such as length of time spent completing the survey instrument, reading instructions, and completing tasks on different survey pages affect data quality metrics and interact with subjective reports of burden.

- To assess how expectations of participating in a Low vs. High burden survey affect objective and subjective burden, memory for burden over time, completion and response rates, and overall experience with the surveys.

- In addition, we also have a secondary research question which will examine the qualifiers used in the burden scales in this research. There is consensus in the survey methods field that using 5 to 7 point, fully labeled scales are the best practice (e.g., Krosnick, 1999). However, the vague qualifiers (or labels) that are often used in the scales such as 'a lot' vs. 'somewhat' do not have a strong consensus in the field as to whether these vague qualifiers produce the same distribution of responses (e.g., Al Baghal, 2014, 2014). Despite the lack of consensus, very little research has examined the use of these vague qualifiers, which labels are reliable with one another, and best practices for which labels are recommended. Thus, for the burden items included in Surveys 1 and 2, participants will be randomly assigned to answer using response scales that contain the label 'A lot' vs. 'Somewhat' (See Attachments B and C for the exact items that this experiment pertains to). Due to a lack of consensus in the survey methods literature on which qualifier is best to use, we plan to assess the distributions of our rating scales to determine whether responses differ based upon each of the qualifiers.

II. Methodology

In this research, we will use online data collection with participants recruited from Amazon.com's Mechanical Turk (MTurk; Berinsky et al., 2012; Paolacci & Chandler, 2014). MTurk is an online marketplace where individuals can sign up to participate in short online research tasks for nominal compensation. Although the Mechanical Turk population may not be

representative of the entire U.S. population, studies using MTurk samples obtain similar results to surveys using population-based samples (e.g., Mullinix et al., 2016). Samples obtained from MTurk are more representative of the general population than other convenience samples, such as university students (e.g., Berinsky et al., 2012), or using the OSMR participant database which only contains participants in the DC metro area. Further, the results of this study are more concerned with internal validity than the representativeness of any one population. However, the MTurk panel can be considered a proxy for BLS survey respondents because they consist of households that could be sampled in any BLS household survey. Amazon Mechanical Turk is also an efficient way to collect information from large numbers of participants. It would be difficult to assess response rates with BLS household survey respondents without potentially affecting data quality of ongoing production surveys. Although these participants receive a small monetary incentive, and are actively seeking to participate in surveys, we have found variability in the levels of perceived burden that participants from mTurk experience in our surveys. For instance, some participants are more or less interested in the study topic, put more effort into answering the questions, or found the topic more or less sensitive. These are all factors that contribute to subjective levels of burden that exist independently of the level compensation. As such, the burden estimates we observe in online research can be considered underestimates of what we may see in a production survey.

Finally, because Mechanical Turk allows researchers to conduct two-part studies, we can assess response rates over time using this methodology. We will conduct a two-part longitudinal survey to assess response rates over time (2 waves). Participants will complete the first survey, which will consist of a series of questions drawing on items from BLS and Census surveys, designed to mimic the experience of completing a BLS household survey (see Attachment G). The questions were selected to represent a number of questions that appear on federal surveys, a mix of questions that household respondents would be able to answer for themselves and for other household members, and items that represent a mixture of topics that may differ on important dimensions of burden (e.g., difficulty, sensitivity, interest level, etc). They will then answer a series of questions to assess their subjective burden experienced while completing the survey. One month later, they will be notified that they are invited to complete a second follow-up survey. See the below links for similar work previously conducted at BLS using similar methodologies and samples:

- CIPSEA research
- Respondent burden research
- SOGI Online Proxy Survey

We will add language to the beginning of the survey stressing the importance of responding "don't know" or "prefer not to say" rather than giving an inaccurate answer. This has shown to

help avoid a ceiling effect of participants answering every item that is typical when using online, voluntary, non-probability panels (de Leeuw et al., 2015; Joinson et al., 2008; Kaplan & Edgar, 2018). We will also note that each question is voluntary so that participants are aware they can skip items (Betts, 2016). This mirrors what actual interviewers say to survey respondents in BLS surveys.

As mentioned, we will include a range of survey items that real BLS household respondents might encounter across various surveys, including the CPS and its supplements, CE, the American Time Use Survey (ATUS) and its supplements, and the ACS, which shares many of the same items as the CPS. We expect to get a wider distribution of answers by including these questions, and therefore more power to detect any potential differences in response unit and item rates across items. In doing so, this will more closely approximate the actual length of BLS surveys.

One month after completing the first survey (chosen because the CPS is also a monthly survey), participants will receive a notification via Mechanical Turk inviting them to participate in a follow-up survey (see Attachment D for the full text of the invitation). We will ask participants what they recall about the first survey they completed and how burdensome they recalled it to be. We will also then be able to assess the impact of burden on data quality metrics, including reporting, response rates to the follow-up survey, nonresponse bias, and memory bias for experienced versus recalled burden.

The following experimental conditions will be embedded within the survey:

- Low burden condition: participants will receive a vignette in the form of an advance letter to participate in a federal survey. The letter will state that they have been randomly selected to participate in the survey, they will receive one mailing notifying them of the survey, the survey will be conducted on the phone, it will take ten minutes to complete, and they will only participate one time. This resembles what one-time surveys like the ATUS currently do. See Attachment E for the full text of the Low burden condition advance letter.

- High burden condition: participants will receive a vignette in the form of an advance letter to participate in a federal survey. The letter will state that they have been randomly selected to participate in the survey, they will receive five mailings notifying and reminding them to complete the survey, it will be conducted in person with an interviewer in their household, it will take forty minutes to complete each interview, and they will be asked to complete the survey every month for 8 months. This resembles what surveys such as the CPS, ACS, and CE currently do. See Attachment F for the full text of the High burden condition.

- During Survey 1, participants will be randomly assigned to read either the low or high burden advance letter, at Survey 2, they will be assigned to read the other advance letter. This design will allow us to assess individual variability in level of subjective burden within participants, as well as any potential order effects in subjective burden based on answering a high burden survey first vs. a low burden survey first, as there is mixed evidence in the survey methods literature regarding whether answering a complex, multi-wave survey affects burden outcomes or not (e.g., Rodhouse et al., 2018; McCarthy et al., 2006; Sinibaldi & Karlsson, 2016).

- We will also embed an experiment where half of the participants receive a scale item that states “A little” and the other half “Somewhat” for the subjective burden items (See Attachments B and C for the exact items that this experiment pertains to). Due to a lack of consensus in the survey methods literature on which qualifier is best to use, we plan to assess the distributions of our rating scales to determine whether responses differ based upon each of the qualifiers.

Pretesting

We will conduct a pretesting phase of this research to assess the measures to be used in the final survey instrument. The pretesting phase will be a very short version of the first survey, where participants answer questions that currently appear on federal surveys, were pretested to potentially appear on federal surveys, or are approximately the same question as one that appears on a federal survey to be adapted for an online mode with an mTurk audience. The pretest will be done to ensure that the burden measures being used have high construct validity and are measuring the construct of burden and its multidimensional nature as accurately as possible. Participants will be randomly assigned to either the Low or High burden conditions. They will then answer a set of questions about themselves and as a proxy for up to one additional household member, randomly selected. They will then be asked follow-up questions, first a set of open-ended questions about their experience completing the survey, and then questions about how burdensome they perceived the survey to be. These measures will be used to validate that the Low and High burden conditions elicited the intended level of burden in participants. Afterward, participants will rate a set of words that may or may not have described their experience completing the survey, because there are over 30 words that need to be pretested, we will divide the pre-testing sample into thirds, asking participants to rate a subset of 13 of the words only. We will obtain ratings of how positive or negative participants find each word, and to select which of those 10 words reflect their experience completing the survey. The results from this pretesting will help us select words to include in a

word bank in the final survey instrument, where participants will complete a word bank task asking them to select up to 15 presented words (positive, negative, and neutral) that described their experience completing the survey questions. Pretesting is required to verify the words we categorized as positive, negative, or neutral are perceived as such to participants. These words are based on feedback from focus group participants who were former ACS respondents.

Because this research involves pretesting, interim results from the pretesting phase will be used iteratively to generate the final instrument for Survey 1 and 2. To see the entire Prestesting protocol, see Appendix A.

Survey 1

Results from the Word Bank task in the Pretesting Survey will be used to inform the final design of Survey 1. In the first survey, participants will start by imagining they received a letter in the mail informing them they have been randomly selected to participate in a federal survey, the letter will reflect the information from the Low vs. High burden conditions. Participants will then proceed to complete the actual survey questions, completing a series of questions drawn from the aforementioned federal surveys (see Attachment G). The items in Attachment G will serve as demographic items about the participants in this study – they will not be asked to report their demographics outside of these questions. They will answer these questions for themselves and up to two other randomly selected household members. They will then be asked follow-up questions, first a set of open-ended questions about their experience completing the survey, and then questions about how burdensome they perceived the survey to be. Participants will then complete the previously mentioned word bank task, where they will be asked to select the words that describe their experience completing the survey. We will select the final words to include in the word bank based on the Pretesting results. The words will be shown in random order to avoid order effects. Finally, participants will complete items aimed to assess the level of subjective burden they experienced while answering the survey questions. See Appendix B for the full Survey 1 instrument.

Survey 2

About one month after participating in the first survey, participants who completed Survey 1 will be invited to complete a follow-up survey. Participants who provided their Mechanical Turk ID in Survey 1 will receive the invitation via their Mechanical Turk account explaining that they previously participated in a survey about data collection for BLS, and that they're invited to complete a follow-up survey. The survey will assess their memory for the burden they experienced in the previous survey. Using this design, we will be able to assess what information about the survey participants retained over time. We will repeat the open-ended questions to assess their memory for Survey 1, as well as the word bank task (the same words

in the same order that they saw in Survey 1) to see how memory of subjective burden is affected over time. This will provide insight into what information participants recall about the survey and what aspects of burden were retained in long-term memory. After completing the memory questions, we will again present participants with the vignette advance letters (the one they did not receive in Survey 1) and ask them to rate how burdensome it would be to participate in the survey and answer survey questions about them and their household (see Attachment G). For the sake of time, participants will only answer as a proxy for up to one additional household member. This design will allow us to assess within-participant perceptions of low versus high levels of burden, as well as order effects of subjective burden (i.e., participants' level of subjective burden based on whether they saw a high vs. low burden letter first in Survey 1 or last in Survey 2). At the very end of Survey 2, participants will read a debriefing page explaining in full that the survey was fictional, they were not selected to participate in any official government survey, and will not be contacted again. See Appendix C for the full Survey 2 instrument.

Outcome Variables

By matching the data from both surveys, we will be able to determine the following across the Low versus High burden conditions:

- Correlation between subjective and objective burden measures (the subjective burden measures include the ratings of how burdensome, effortful, sensitive, easy/difficult, invasiveness, and perception of how long the survey was; objective measures include time spent completing the survey & time spent reading instruction pages)
- Overall response rate (overall) and by burden condition
- Item non-response rate (overall and by question) and by burden condition
- Nonresponse bias (e.g., demographic groups that were more or less likely to participate in the second survey) by burden condition type
- Attrition rate within survey one and between survey one and two by burden condition
- Consistency across ratings of burden over time (Surveys 1 and 2)
- What information about the survey participants retained across Surveys 1 and 2
- Within-participation variation of the above effects based on receiving a high vs. low burden vignette, and any order effects of answering a high vs. low burden survey first

We will use the results to better understand the reasons people find surveys burdensome, how they describe their experience completing our surveys, to assess perceptions of burden over time, and to potentially use these insights to ease respondent burden, in conjunction with cognitive interviews, focus groups, and other online research conducted on burden at the Census and BLS. This research will also result in a research paper to contribute to the survey methods literature on subjective burden more broadly (not for publishing estimates).

III. Participants

Up to 870 Amazon Mechanical Turk participants will be recruited. This sample size was determined to sufficiently explore the range of variables of interest, and because we expect a very small effect size as the study manipulations are subtle for online surveys of this nature (e.g., Hill et al., 2016). This sample size also takes into account break-offs, incomplete data, and participants who do not follow the task instructions, similar to other OMB-approved samples used for studies of this nature listed in the introduction.

IV. Burden Hours

The Pretesting Survey is expected to take 10 minutes; Survey 1 is expected to take 20 minutes; and Survey 2 will take about 20 minutes, for a potential burden of up to 10 minutes for participants in the Pretest, and 40 minutes per participant who completes both Surveys 1 and 2 (participants who participated in the Pretesting Survey will not be eligible for Surveys 1 or 2).

Table 1. Estimated Burden Hours

	# of Participants Screened	Minutes per participant for Screening	Total Screening Burden	Maximum number of Participants	Minutes per participant for data collection	Total Collection Burden	Total Burden (Screening + Collection)
Pretesting Survey	170	0	0	170	10	29	29
Survey 1	700	0	0	700	20	233	233
Survey 2	0 (note these are the same participants from Survey 1)	0	0	490* assuming a 30% attrition rate, based on prior OSMR studies; also see Hall et al. (2016).	20	164	164
						Total Burden	426 hours

V. Payment to Participants

Participants will receive $1.00 for participation in the Pretesting Survey, (participants who participated in the Pretesting Survey will not be eligible for Surveys 1 or 2). Participants will receive $2.00 for participating in survey one and $2.00 for participating in survey two, for a potential total of $4.00 if participants complete both surveys, all of which are typical rates for

similar MTurk tasks (Paolacci & Chandler, 2014). The Pretesting Survey will take 10 minutes to complete; Survey 1 should take about 20 minutes and Survey 2 should take about 20 minutes to complete. The estimated maximum total for participant fees is $2,550.

Recruiting of participants will be handled by Amazon Mechanical Turk. Once participants are recruited into the study, they will be given a link to the survey, which is hosted by Qualtrics.com. The data collected as part of this study will be stored on Qualtrics servers.

Participants will be informed of the OMB number and the voluntary nature of the study.

> This voluntary study is being collected by the Bureau of Labor Statistics under OMB No. 1220-0141. This survey will take approximately [20 minutes / 10 minutes] to complete. The BLS cannot guarantee the protection of survey responses and advises against the inclusion of sensitive personal information in any response. This survey is being administered by Qualtrics and resides on a server outside of the BLS Domain. Your participation is voluntary, and you have the right to stop at any time.

1. Attachments

Attachment A: Pretesting Survey instrument

Attachment B: Survey 1 instrument

Attachment C: Survey 2 instrument

Attachment D: Email notification for invitation to complete Survey 2

Attachment E: Low burden advance letter

Attachment F: High burden advance letter

Attachment G: Survey questions

References

Al Baghal, T. (2014). Numeric estimation and response options: an examination of the accuracy of numeric and vague quantifier responses. *Journal of Methods and Measurement in the Social Sciences*, *5*(2), 58-75.

Al Baghal, T. (2014). Is vague valid? The comparative predictive validity of vague quantifiers and numeric response options. In *Survey Research Methods* (Vol. 8, No. 3, pp. 169-179).

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public opinion quarterly*, *42*(1), 93-104.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(3), 128.

Berinsky, A.J., Huber, G.A. and Lenz, G.S. (2012) 'Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk', Political Analysis, 20(3), pp. 351–368. doi: 10.1093/pan/mpr057.

Bradburn, N. (1978, August). Respondent burden. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (Vol. 35, p. 40).

Buchanan, T., Paine, C., Joinson, A. N., & Reips, U. D. (2007). Development of measures of online privacy concern and protection for use on the Internet. *Journal of the American Society for Information Science and Technology*, *58*(2), 157-165.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social science computer review*, *19*(2), 146-162.

de Leeuw, E. D., Hox, J. J., & Boevé, A. (2015). Handling Do-Not-Know Answers Exploring New Approaches in Online and Mixed-Mode Surveys. *Social Science Computer Review*, 0894439315573744.

Fricker, S., Yan, T., & Tsai, S. (2014, May). Response burden: What predicts it and who is burdened out. In *JSM proceedings* (pp. 4568-4577).

Hall, H., N. Lewis, J. Chandler, and P. Ellsworth. Conducting Longitudinal Studies on Amazon's Mechanical Turk: A Meta-analysis with Recommendations. Working Paper, 2016.

Holzberg, J., Katz, J., Morales, G., Davis, M. (2018). Assessing Respondents' Perceptions of Burden in the American Community Survey. Presentation at the Federal Committee on Statistical Methodology Conference.

Joinson, A. N., Paine, C., Buchanan, T., & Reips, U. D. (2008). Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. *Computers in Human Behavior*, *24*(5), 2158-2171.

Kaplan, R.K. and Edgar, J. (2018). Priming confidentiality concerns: How reminders of privacy affect response rates and data quality in online data collection. Paper to be presented at AAPOR May of 2018. Denver, CO.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys the effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847-865.

McCarthy, J.S., Beckler, D.G., Qualey, S.M., (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, 97-112.

Miller, Peter V. 2017. "Is There a Future for Surveys?" *Public Opinion Quarterly* 81: 205-12.

Mullinix, K.J., Leeper, T.J., Druckman, J.N. and Freese, J. (2015) 'The Generalizability of Survey Experiments', *Journal of Experimental Political Science*, 2(2), pp. 109–138. doi: 10.1017/XPS.2015.19.

Paolacci, G., and J. Chandler. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." Current Directions in Psychological Science, vol. 23, no. 3, 2014, pp. 184–188.

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

Rodhouse, J., Wilson, T., & Ridolofo, H. (2018). Response Likelihood to an Establishment Survey with a Simple Questionnaire Following an Establishment Survey with a Complex Questionnaire. Presented at FCSM 2018, Washington, DC.

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101-1108.

Sinibaldi, J., Karlsson, A.O., (2016) "The Effect of Rest Period on Response Likelihood," Journal of Survey Statistic and Methodology, 00, 1-14.

Wirtz, D., Kruger, J., Napa Scollon, C., & Diener, E. (2003). What to do on spring break? The role of predicted, on-line, and remembered experience in future choice. Psychological Science, 14, 520 –524. http://dx .doi.org/10.1111/1467-9280.03455