March 1, 2019

NOTE TO THE REVIEWER OF:	OMB CLEARANCE 1220-0141 "Cognitive and Psychological Research"
FROM:	Jean Fox and Robin Kaplan Office of Survey Methods Research
SUBJECT:	Submission of Materials for Evaluating Qualifiers in Rating Scales

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 130 hours.

If there are any questions regarding this project, please contact Jean Fox at 202-691-7370.

Background

Like many other research organizations, the Bureau of Labor Statistics (BLS) Behavioral Science Research Center (BSRC) uses survey questions with rating scales to study a variety of issues, including respondent burden, satisfaction and feedback survey data, and usability. These scales typically have five or seven response options, generated by pairing the construct being measured with a series of qualifiers. The scales can be unipolar, where the response options use qualifiers ranging from something like "not at all" to "extremely," or bipolar, where the response options range from something like "very" of one construct to "very" of the opposite construct.

Ideally, the scales would provide true intervals across response options that provide distinctions for which most respondents would find meaningful, to obtain the most valid and reliable data possible. Researchers do this by selecting qualifiers that appear to create equi-distant response options, and then assume they are interval scale for analysis. However, they often don't know for sure.

To help researchers generate response options that more closely resemble true interval scales, previous studies have evaluated the values that participants associate with qualifiers. These studies have used two approaches to determine a value associated with each of a number of qualifiers.

Previous Methods Used to Study Response Scales

One approach is the Method of Successive Intervals, where participants provide ratings for each qualifier based on the "value" that they ascribe to the qualifier (Jones & Thurstone, 1955). For example, "Extremely" might get a high rating while "None at All" might get a low rating. By collecting values from a large sample of participants, we can assign compiled values to each qualifier. With those values, we can identify labels that will generate a scale with equidistant response options. We may also be able to identify cases where qualifiers are very similar to each other and should perhaps be avoided in the same scale (e.g., "somewhat'" and "moderately" may not be noticeably different from each other).

The other approach is the method of paired comparisons, where participants evaluate pairs of qualifiers and select which of the two represents "more" (Guilford, 1954). The pairs will each include two qualifiers that are likely to have similar values associated with them (such as "Somewhat" and "Moderately"). The paired comparison task will help clarify the differences between these similar qualifiers. This differs from the Method of Successive Intervals because instead of rating the quantity that a qualifier represents, participants will make comparative judgments between two pairs of qualifiers to determine which one represents a greater value. This method allows researchers to yield an interval-scale ordering of qualifiers. Because there are many qualifiers, it would be highly burdensome to ask participants to rank order each of

them relative to each other (e.g., Bramely & Oates, 2010; Rounds et al., 1978). Paired comparison tasks reduce burden by asking participants to make judgments between two qualifiers at a time.

Both methods provide different types of information necessary for this research. The Method of Successive Intervals will allow us to understand the "quantity" a qualifier represents on its own, independently of other qualifiers within a scale. The method of paired comparisons will allow us to generate relative values between closely related qualifiers to understand how they compare to one another.

There have been a few studies addressing how to generate optimal interval response scales that are meaningful and distinctive for respondents in the past. However, the majority of these studies are several decades old and do not include many of the qualifiers we often use today (e.g., Saffir, 1937; Jones & Thurstone, 1955; Stilson & Maroney, 1966; Bradburn & Miles, 1979). Thus, a gap in the literature remains in evaluating current qualifiers with a current audience, which is the focus of this research.

Study Goals

The goal of this study is to try to fill this gap in the literature regarding the value of different qualifiers in response scales, and to assess the values that people today assign to qualifiers commonly used in social science research currently. This will help researchers in BLS and other organizations ensure the scales they use contain response options suitable for analysis as interval level data. We will use methods that are similar to those used successfully by previous researchers, with updated qualifiers, a modern group of participants, and currently available research tools. The results will provide important insights into how much each qualifier represents, the distance between qualifiers, and the ability to distinguish values of similar qualifiers. The results of this study will help researchers from BLS and beyond to design more valid, reliable scales in our surveys. This research will focus on the following three commonly used categories of qualifiers:

- Strength/Intensity (e.g., Not at all, Somewhat, Very)
- Frequency (e.g., Never, Sometimes, Often)
- Evaluation (e.g., Bad, Good, Great)

Analysis of Respondent Burden

Finally, at the end of the survey instrument, we will ask a set of questions about respondent burden. We will ask participants to rate how burdensome they found it was to complete the entire survey, as well as related measures including how interesting, effortful, how easy or difficult it was to complete the survey, willingness to participate in a similar survey again in the future, and

perceptions of the survey length. As previous research has shown (e.g., Fricker et al., 2012), these perceptions of burden may have an effect on data quality and important survey outcomes and also being researched in other ongoing studies. These items will used to assess how participants rate the burden of the study, and also to compare how participants rate burden on this survey compared to other surveys where we also assessed burden. Burden ratings will us compare how perceptions of burden are related to objective measures of burden (e.g., length of time spent completing the survey).

Analysis of Gender Identity

In a continued effort to better understand measures of gender identity, BLS staff participate in the Measuring Sexual Orientation and Gender Identity Research Group. In this group, we conduct research to collect feedback on terminology for gender identity and terms used to describe one's gender identity. As such, the demographic section of this protocol contains additional response options and an open-ended question where respondents can indicate their gender identity. Asking the question in this format has been approved for this use in in previous OMB packages.

Methodology

Participants will complete an online survey where they will provide values for a series of qualifiers (i.e., they will be asked to quantify "how much" that qualifier represents; see the Appendix for the full survey instrument and for examples of the comparisons we plan to make). To maximize the generalizability of the results, we will not use any one specific construct for the rating scale task, but ask respondents to provide general ratings of three commonly used qualifiers (i.e., strength/intensity, frequency, and evaluation). To provide some context for participants, we will give participants an example at the beginning of each rating task (e.g., "For example, in thinking about usefulness, you could find a product to be 'the least useful' or 'the most useful.'). After providing values, participants will be presented with a series of qualifiers shown in pairs. They will be asked to do a paired comparison task, where they indicate which of the two qualifiers represents "more" of a particular amount. Finally, they will complete demographic questions and the respondent burden items.

Following previous analysis techniques used by researchers (i.e., the method of successive internals and paired comparisons), we will standardize the mean values respondents assigned to each qualifier to understand the "quantity" a qualifier represents on its own and in relation to one another.

Participants

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., 2015). 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 volunteers in the DC metro area. Further, the results of this study are more concerned with internal validity than the representativeness of any one population. MTurk is also an efficient way to collect information from large numbers of participants.

We will recruit a total of 390 participants from MTurk. 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., Paolacci & Chandler, 2014). 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.

Burden Hours

The survey is expected take approximately 20 minutes to complete for a burden of up to 20 minutes per participant.

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 Hours	Total Burden Hours (Screening + Collection)
Online survey	390	0	0	390	20	130	130
Total Burden							130 hours

Payment

Participants will receive $2.00 for participating in the survey, which is a typical rate for similar MTurk tasks (Paolacci & Chandler, 2014). The estimated maximum total for participant fees is $1100. This includes a commission fee the company requires.

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

Participants will be informed of the OMB number and the voluntary nature of the study with the following statement:

> This voluntary study is being collected by the Bureau of Labor Statistics under OMB No. 1220-0141. This survey will take approximately 20 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 SurveyMonkey and resides on a server outside of the BLS Domain. Your participation is voluntary, and you have the right to stop at any time.

Attachments

Survey instrument (see PDF of SurveyMonkey instrument attached.)

References

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), 351–368. doi: 10.1093/pan/mpr057.

Bradburn, N.M. and Miles, C. (1979). Vague quantifiers. *Public Opinion Quarterly, 43*(1), 92-101.

Bramley, T. and Oates, T. (2010). Rank ordering and paired comparisons – the way Cambridge Assessment is using them in operational and experimental work. Retrieved from: http://www.cambridgeassessment.org.uk/Images/125350-summary-of-rank-ordering-and-paired-comparisons-research.pdf. Downloaded 1/14/2019.

Fricker, S., Kreisler, C., & Tan, L. (2012). An exploration of the application of PLS path modeling approach to creating a summary index of respondent burden. In *JSM Proceedings* (pp. 4141-4155).

Guilford, J.P. (1954). *Psychometric Methods*. New York: McGraw-Hill Book Co.

Jones, L.V. and Thurstone, L.L. (1955). The psychophysics of semantics: An experimental investigation. *The Journal of Applied Psychology, 39* (1), 31- 36.

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), 109–138. doi: 10.1017/XPS.2015.19.

Paolacci, G., and J. Chandler. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Rounds, J.B., Miller, T.W., and Dawis, R.V. (1978). Comparability of multiple rank order and paired comparison methods, *Applied Psychological Measurement, 2*(3), 415-422.

Saffir, M.A. (1937). A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 2(3), p. 179-198.

Stilson, D.W. and Maroney, R.J. (1966). Adverbs as multipliers: Simplification and extension. *The American Journal of Psychology*, *79*(1), p. 82-88.