

Use of Amazon MTurk Online Marketplace for Questionnaire Testing and Experimental Analysis of Survey Features

Stephanie Fowler, PhD, MPH

Gordon Willis, PhD

Richard Moser, PhD

Rebecca Ferrer, PhD

David Berrigan, PhD, MPH

Division of Cancer Control and Population Sciences
National Cancer Institute
9069 Medical Center Drive
Rockville, MD 20850

Introduction

Many data elements in the Federal Statistical System are obtained via standardized survey questions. It is well established that such questions, as well as groups of questions addressing more complex psychological and behavioral constructs, may be improved through the use of reliability and validity studies, cognitive testing, and analyses of the psychometric properties of alternative versions. Additionally, a growing body of evidence has identified contextual factors that influence responses to survey questions, such as question order and mode of administration. Further, examining viable platforms for evaluation of Federal surveys, in addition to data collection, is a vital endeavor given our rapidly changing world of data collection and movement toward online and Internet-based studies.

However, testing of Federal survey questions, especially those in health surveys, have sometimes been limited because of cost and difficulties in obtaining samples for studies. Despite the efforts by some researchers to test for question order effects in large, complex surveys focusing on surveillance (McClain et al. 2012), testing and evaluating of questions in Federally fielded surveys via experiments is minimal. In order to inform future iterations of surveys, make sound decisions on potential survey redesigns, and to ensure data quality, it is important to conduct split-sample survey experiments (Behr et al., 2013; Blair, 2011; DeMaio & Willson, 2011; Krosnick, 2011). Furthermore, examining question reliability, through test-retest reliability studies, requires repeated measurement over time, and cross-sectional Federal surveys are not set up according to a longitudinal or panel design. Finally, although cognitive testing is often viewed as the gold standard for determining item function and interpretability (Willis, 2005), it can be limited in its application because of its high costs and respondent burden. Nimble, cost effective methods that allow broad geographic and demographic samples for testing surveys and survey items could complement existing approaches.

The Internet affords us the opportunity to potentially conduct survey experiments, to follow respondents over time, and to evaluate how questions are interpreted via an online variant of cognitive testing, in the form of web probing (Behr et al., 2013; Murphy et al., 2013). Given the importance of evaluating the functioning of health and other surveys, including those implemented by Federal Agencies in a quick and cost-effective manner, the present investigation explored the utility of an online platform – Amazon’s Mechanical Turk - that allows for the inclusion of split-sample surveys, longitudinal designs, and web probing. The present investigation was threefold: a) conduct an experimental study of question order effects, b) evaluate the test-retest reliability of questions using a longitudinal design, c) and collect and examine free text responses to structured cognitive probes (i.e., web probes, Behr et al., 2013). The survey questions evaluated in the present investigation are also currently fielded in the 2015 National Health Interview Survey (NHIS) Cancer Control Supplement (CCS) which is sponsored by the National Center for Health Statistics and fielded by the United States Census Bureau.

Methods

We used Amazon's Mechanical Turk (and Mechanical Turk Prime) as a source of respondents for an Internet-based survey experiment that examined question order effects, obtained repeated measures for test-retest reliability, and explored the utility of a web-based platform for implementing structured cognitive probes assessing question interpretation. Mechanical Turk is a community of internet users who login to the platform to complete tasks ("crowdsourcing"). Run by Amazon.com, the online labor market, Mechanical Turk, allows requestors to post tasks requiring human intelligence, and in turn, respondents are compensated small amounts of money per task. Although these tasks often involve transcribing audio into written text or product evaluation and test as examples, Mechanical Turk is increasingly being leveraged to recruit participants for research studies. The evaluated survey questions concerned perceptions of the environment for walking, and the frequency and duration of walking for both transportation and leisure. These questions have already undergone preliminary cognitive testing, and are fielded in the 2015 NHIS CCS. The cognitive probes we implemented in the online platform were based on those administered in person during pretesting of the survey module.

Prior to data collection, clearances from the Office of Human Subjects Research Protections (OHSRP) and Office of Management and Budget (OMB) were obtained. We recruited $n = 1,447$ respondents at Time 1 using Mechanical Turk (MTurk). At Time 1, respondents were randomly assigned to either *first* (a) answer a set of questions on walking for transportation and leisure during the last 7 days, or *first* (b) answer 9 questions about their perceptions of the walking environment for which they used a yes/no response format. Respondents in both conditions then answered the remaining questions (e.g., perceptions second, if walking for transportation/leisure was asked first). Note that, in the fielded 2015 NHIS CCS, there was only one order of questions: respondents first answered a set of questions on walking for transportation and leisure followed by questions on perceptions of the walking environment. At Time 1, respondents also answered a set of basic demographic questions. Approximately 4 weeks later, at Time 2, we invited back the same respondents from Time 1 to repeat the study. We obtained $n = 960$ (66%) of the original respondents at Time 2 using the Mechanical Turk Prime platform, which allows researchers to repeat a task with the same workers through matching based on MTurk worker ID. The finding that only 66% of Time 1 respondents repeated at Time 2 was due in part to an unanticipated maintenance shutdown of the Mechanical Turk Prime platform on day 3 of Time 2 data collection. At Time 2, respondents completed the study in the same order to which they were assigned at Time 1. In addition to answering the questions about their perceptions and walking behaviors, respondents were also administered 4 cognitive probes at the end of the questionnaire. The response format for probes was open-ended, and respondents were given the chance to explain their interpretations concerning 4 of the tested survey questions, specifically: (a) what they were thinking when answering questions about infrastructure, (b) what walkability meant to them, (c) who they were thinking about when answering questions about walkability, and d) and if they were answering based on what they did do versus could do (Please see Table 1). We used Qualtrics at both times as the web-based administration software into which the questionnaire, and probes, were programmed.

Table 1. Cognitive probes administered at Time 2.

When answering the question, "Where you live, are there roads, sidewalks, paths or trails where you can walk?" Please say more about what you were thinking when answering this question?
We asked some questions about whether there are places in your neighborhood that you can walk to. Please say more about how you decided whether or not you can walk there?
When answering the questions about places in your neighborhood that you can walk to, were you thinking of places that YOU might walk to, or places that OTHERS might walk to, or both?
When answering the questions about places in your neighborhood that you can walk to, were you thinking about what you actually DO, or about what you COULD do if you wanted to?

Results

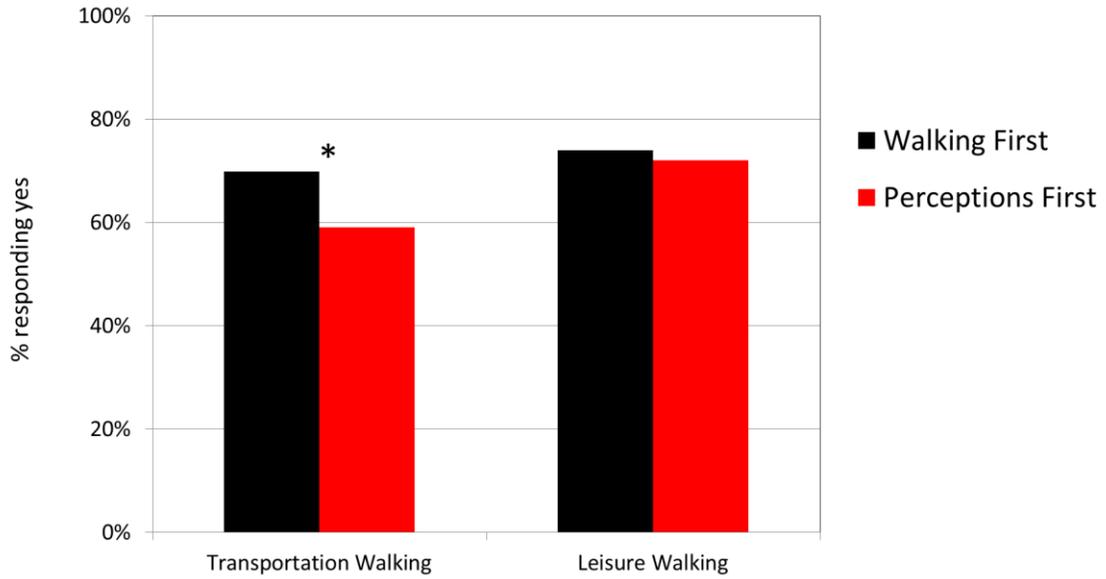
Table 2 organizes the distribution of demographic information for our MTurk sample and the U.S. population. The majority of the sample was more male (56%), more White (83%), and younger ($M_{age} = 31.8$, $SD_{age} = 10.6$) relative to the same characteristics of the U.S. adult population. In addition, approximately 50% of the MTurk sample ($N = 1,447$) had a college degree or higher.

Table 2. Demographic data for Time 1 MTurk respondents.

N = 1,447		MTurk Sample	U.S.
Variable		%	%
Sex			
	Female	43.8	50.8
	Male	56.2	49.2
Race			
	White	83.2	78.3
	Black	6.6	13.1
	All other races	7.5	8.6
	Don't know	2.7	
Ethnicity			
	Hispanic	7.2	16.9
Education			
	Less than Bachelor's degree	50.5	69.8
	Bachelor's degree or higher	49.5	31.2

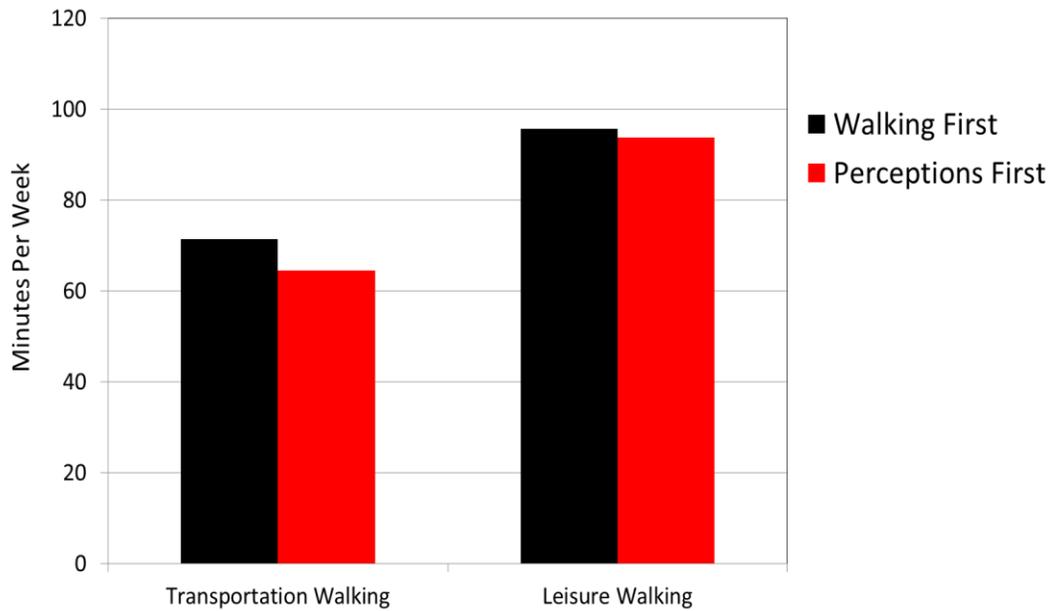
Question Order Effects. Participants either first answered questions about walking for transportation and leisure, or they first answered questions about their perceptions of the walking environment. To assess the presence of order effects between the two conditions on prevalence and duration of walking, we conducted logistic (for prevalence) and linear (for duration) regression analyses. Order condition served as the main predictor variable with age, sex, race, ethnicity, and education as covariates in the model. The only significant effect to emerge was on prevalence of transportation walking, $OR = .604$, 95% CI (.49, .75), $p < .001$. Respondents first completing questions on perceptions of the walking environment, as opposed to first completing questions on walking for transportation and leisure, had significantly lower odds of transportation walking prevalence.

Table 3. Prevalence of transportation walking and leisure walking as a function of order condition.



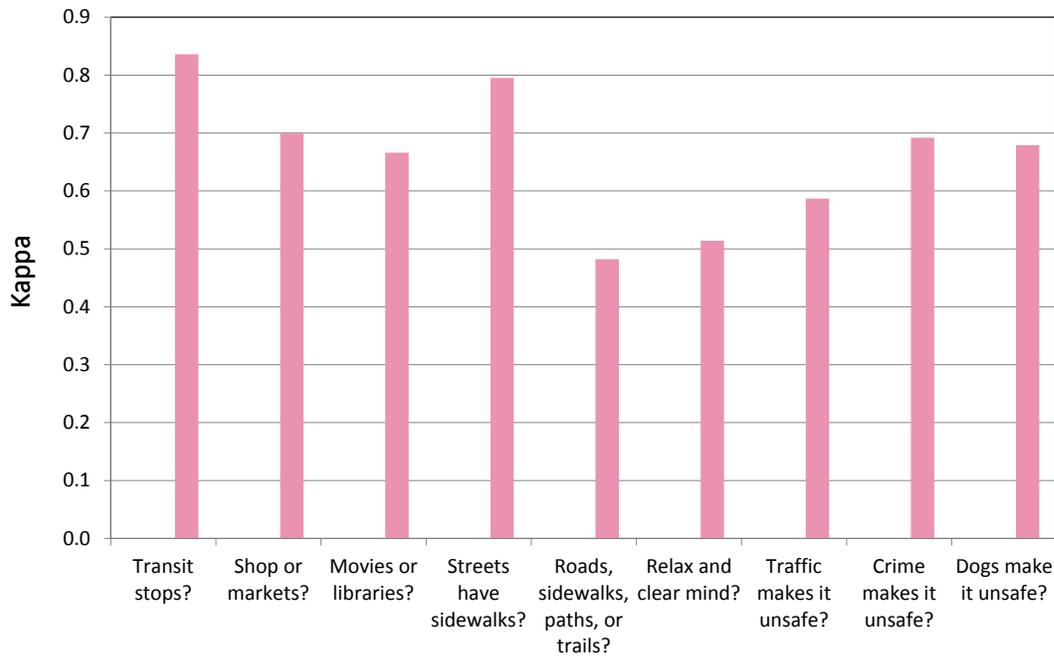
*OR = .604, CI 95% (.485, .751), $p < .001$
Walking behaviors first as reference group; Yes coded 1; no coded 0
Binary logistic regressions include age, sex, race, and education as covariates

Table 4. Duration of transportation walking and leisure walking as a function of order condition.



Test-Retest Reliability. Test-retest reliability was examined by computing a Kappa coefficient comparing the yes/no responses to each of the 9 items on perceptions of the walking environment at Time 1 and Time 2. Higher Kappa values indicate greater consistency between responses at Time 1 and Time 2. Kappa coefficients ranged from .48-.84. Most exhibited moderate to high consistency with most exhibiting a Kappa value of .65 or higher. Table 5 illustrates the variability in Kappa values.

Table 5. Item test-retest reliability of the 9-item set of questions on perceptions of the walking environment.



Web Probing. Nine-hundred fifty-five respondents (99% of N = 960) answered all 4 cognitive probes at Time 2. Below are example web probes from the study followed by a small sample of responses from MTurk.

Web Probe 3: *...tell me more about what you were thinking when answering the question... "Where you live, are there roads, sidewalks, paths or trails where you can walk?"*

Example 1: "There are sidewalks alongside almost every road, but no nice walking paths or anything around. And the regular sidewalks are not always kept in good condition or kept clear, so while I technically can walk on them, I sometimes would be discouraged from doing so."

Example 2: "I was thinking of the streets and sidewalks outside of my home."

Example 3: "I was thinking about walking paths and trails near my house that I can take the dog for a walk. I was thinking about the walk last week and how enjoyable it was."

Web Probe 4: We asked several questions about walking... “Please say more about how you decided whether or not you could walk there?”

Example 1: “I would consider places unwalkable if there are busy, high speed roads and it would take more than 10 minutes to walk there.”

Example 2: “I deem a place within walking distance if I can walk there within 30 minutes”

Example 3: “I live in a hilly area so I was thinking about the fact that it's uphill to get to downtown (ha, up to get down), plus there's no sidewalk and some blind corners.”

Our qualitative assessment of responses to the probes were particularly informative in the diagnosis of the relatively low reliability obtained for the question asking respondents if there are roads, sides, paths, or trails they could walk to. Many respondents indicated thinking about more than one of the infrastructure types listed in the question, but not all 4. Because it was effectively a multi-barreled question, and we see variability in anchoring on several features of the question, it stands to reason that there may have been intra-individual differences in what respondents chose to anchor on between Time 1 and Time 2 responses, such that they may have been thinking of different elements of the question at these two points, accounting for the lower test-retest reliability observed.

Discussion

The goal of the study was to evaluate question order of a set of questions in an already fielded Federal survey. In addition to examining if placement of questions influenced responses, we also examined test-retest reliability for one of the newly developed item sets, specifically examining perceptions of the walking environment. Finally, we included open-ended web probes to assess item function.

One significant order effect was found, but overall these were minimal in extent. Of the 4 measures of walking (transportation prevalence, leisure prevalence, transportation duration, leisure duration), only prevalence of transportation walking was affected by question order. Given that the order of questions in the 2015 NHIS CCS is walking questions first followed by perceptions questions, this lends credence to the choice of ordering currently in the field (as a reversal would not be expected to markedly affect the results obtained). Further, information on ordering effects – or rather, the lack of such effects – can be used to inform decisions to be made within the Division of Cancer Control and Population Sciences at the National Cancer Institute concerning future redesign of the NHIS CCS.

Most of the 9-item set of questions about perceptions of the walking environment exhibited moderate to high test-retest reliability. Use of the Mechanical Turk and Mechanical Turk Prime platforms allowed for repeated measurement. In addition, most MTurkers responded to all 4 web probes with free text that was substantive and rich in information. The qualitative analysis helped determine item function and trouble shoot an item that exhibited low test-retest reliability. In the current case, the relatively low reliability could be plausibly explained by MTurkers' responses to a cognitive probe question asking them specifically about the item that exhibited low reliability. We conclude that the use of a mixed-methods approach, involving both quantitative data (test-retest reliability), and qualitative data (web probes) is helpful in understanding item function.

In general, we found that the responses to web probes were rich and informative, which was somewhat surprising, given the low incentive amount compensated to respondents, and the fact that open-ended questions often fail to provide useful information for survey researchers. Perhaps the probes tapped into issues that were interesting to respondents, or that allowed them to expound upon their opinions and reactions to the survey questions in a way they found to be attractive. Emerging research shows that workers report a greater amount of intrinsic motivation (e.g., genuine interest in task) versus extrinsic motivation (e.g., doing it for the compensation) for completing different tasks (Buhrmester et al., 2011).

Finally, similar to other studies in the literature (Goodman et al., 2013), we found that Amazon's Mechanical Turk appears to be a viable platform for questionnaire testing and evaluation involving split-sample designs, repeated measurements, and free-text responding to web probes. Furthermore, inclusion of Mechanical Turk Prime allowed us to follow our respondents over a 4 week period so that we could obtain repeated measures for estimates of test-retest reliability. In addition to its utility for questionnaire testing and evaluation, MTurk and other online platforms should be evaluated as mechanisms for non-probability-sample based field data collection for our Federal surveillance systems, especially given that these systems are moving toward online format.

References

- Behr, D., Bandilla, W., Kaczmerik, L., & Braun, M. (2013). Cognitive probes in web surveys: On the effect of different text box size and probing exposure on response quality. *Social Science Computer Review*, 32, 524-533.
- Blair, J. *Response 1 to Krosnick's Chapter: Experiments for Evaluating Survey Questions*. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality*, pp. 239-251. Hoboken: NJ: Wiley, 2011.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, 6(1), 3-5.
- DeMaio, T., & Willson, S. *Response 2 to Krosnick's Chapter: Experiments for Evaluating Survey Questions*. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality*, pp. 253-262. Hoboken: NJ: Wiley, 2011.
- Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3), 213-224.
- Krosnick, J. *Experiments for Evaluating Survey Questions*. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality*, pp. 215-238. Hoboken: NJ: Wiley, 2011.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1), 1-23.
- McClain, J.J., Grant, D., Willis, G., & Berrigan, D. (2012). Effects of temporal domain on self-reported walking behaviors in the California Health Interview Survey. *Journal of Physical Activity & Health*, 9, 344-351.
- Murphy, J., Keating, M., & Edgar, J. (November, 2013). *Crowdsourcing in the Cognitive Interviewing Process*. Paper presented at the annual Federal Committee on Statistical Methodology Research Conference, Washington DC.
- Willis, G. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Thousand Oaks, CA: Sage, 2005