# Multi-mode question pretesting: Using traditional cognitive interviews and online testing as complementary methods

**Robin L. Kaplan, U.S. Bureau of Labor Statistics, BLS, USA**
**Jennifer Edgar, U.S. Bureau of Labor Statistics, BLS, USA**

8.12.2020

**How to cite this article:** Kaplan R. L. & Edgar  J. (2020). Multi-mode question pretesting: Using traditional cognitive interviews and online testing as complementary methods in *Survey Methods: Insights from the Field, Special issue: 'Advancements in Online and Mobile Survey Methods'.* Retrieved from https://surveyinsights.org/?p=14659
**DOI:**10.13094/SMIF-2020-00016

## *Abstract*

Questionnaire development, evaluation, and pretesting research is critical for ensuring that survey questions, materials, and data collection procedures produce the highest quality data possible. Interviewer-administered cognitive interviews is a common pretesting method used to collect rich, qualitative data. As technology has advanced, researchers can conduct similar research online in self-administered modes (Behr 2016), allowing for pretesting with larger samples. Each approach has strengths and limitations that researchers can leverage to address their pretesting goals. This research presents a multi-study, iterative project using traditional and online pretesting to evaluate new confidentiality language. Study 1 used traditional cognitive interviews to collect information on respondents' qualitative reactions to, and comprehension of, the new language, but was limited by a small sample of prior survey respondents. Study 2 used online testing to help verify the previous findings with a larger sample, but was limited to hypothetical respondent behaviour. Study 3 used online testing over two waves of data collection to evaluate actual behaviour over time and expanded on the previous two studies by using an experimental design. We discuss the utility of using multiple pretesting methods to complement each other, providing research findings that would not be possible when using one alone.

## *Keywords*

cognitive interviewing, item nonresponse, multi-method studies, survey methods, survey pretesting, web pretesting

## *Acknowledgement*

The first author would like to dedicate this paper to the memory of Dr. Jennifer Edgar, a wonderful collaborator, leader, and all-around exceptional person. This work would not have been possible without her guidance and leadership of the Statistical Community of Practice and Engagement (SCOPE) Confidentiality Pledge Revisions Subcommittee. The authors would also like to thank Morgan Earp for her assistance with initial analyses of the data, and Erica Yu and William Mockovak for their helpful feedback on this paper.

Disclaimer: This report is released to inform interested parties of research and to encourage discussion of work in progress. Any views expressed on the methodological issues are those of the authors and not necessarily those of the U.S. Bureau of Labor Statistics.

## *Copyright*

## Introduction

Ensuring that survey questions are consistently understood, and that respondents are able and willing to answer them, is of critical importance for reducing measurement error and enhancing data quality.  Survey researchers have multiple question pretesting methods available (Groves et al., 2009), each with their own strengths and limitations (D'Ardenne & Collins, 2019). Traditionally, pretesting studies have been conducted using interviewer-administered cognitive testing. But with the increased use of web-data collection, pretesting in self-administered, online modes has emerged (Behr, 2016). However, few pretesting studies have used both traditional interviewer-administered cognitive testing and online, self-administered testing as complementary approaches to address a single research question. This research describes a series of studies using both approaches, demonstrating how web-based pretesting can complement and expand on traditional pretesting.

## Questionnaire Pretesting

Survey organizations often use traditional, interviewer-administered cognitive testing for questionnaire pretesting (Willis, 2005), collecting rich, qualitative data to gain insight into the response process and potential questionnaire problems. Recent technological advances have led to new and innovative, online testing approaches to obtain similar data to traditional cognitive interviews via self-administered web surveys (Behr, 2016), while Internet panels have made access to respondents readily available (Fowler & Willis, 2019). Similar to the advantages gained when surveys moved from interviewer-administered to online, self-administered modes, online testing allows for questionnaire evaluation and pretesting with less cost and effort and more geographically diverse samples (Scanlon, 2019). Online testing also allows for analysis of paradata, which can provide additional insight into the response process (McCalin et al., 2018), and larger sample sizes required for experiments that are not feasible with traditional, interviewer-administered cognitive interviews.

A number of considerations must be taken into account when designing pretesting research.  People who participate in traditional cognitive interviews are often more cooperative than online panel respondents (Ross et al., 2010). Further, online panels may vary in their data quality (e.g., Hillyguys et al.. 2014; Matthijsse et al., 2015; Toepoel et al., 2008; Kennedy et al., 2016; Cornesse et al., 2020), but they tend to be more geographically diverse than samples recruited for traditional cognitive interviews (Berinsky et al., 2012; Scanlon, 2019).  Both generally rely on convenience samples, which are appropriate for pretesting efforts seeking to gain insight into the respondents' thought processes rather than trying to predict survey outcomes or produce general-population estimates.

A number of methodological choices also need to be considered when selecting a pretesting mode. In traditional cognitive interviewing, researchers have the flexibility to use unscripted probes to ask specific follow-up questions about the response processes. Online testing limits researchers to scripted probes that can be programmed into the instrument in advance and responses are sometimes shorter in online testing (Meitinger & Behr, 2016; Lenzner & Neuert, 2017).  These factors must be carefully considered when designing studies to best leverage the strengths of each method (Kaplan et al., 2019; McCarthy, 2019).

## Motivation

The goal of this research was to use multiple pretesting modes (traditional cognitive interviews and online testing) in an iterative, multi-stage research design. Each study added value to the prior, while addressing limitations due to the mode or design, leading to a well-supported conclusion that would not have been possible had only one method been used.

In 2015, the Federal Cybersecurity Enhancement Act was passed, requiring U.S. Federal government agencies to include language about cybersecurity screening of survey responses. Before implementation, BLS needed to understand respondents' reactions to and comprehension of the new confidentiality language and its potential impact on survey response. Although confidentiality language is technical and written in legalistic language, respondents do expect to see such disclaimers before participating in U.S. Federal surveys (Landreth et al., 2008). Additionally, this research was of critical importance, as confidentiality language can sometimes raise suspicion about responding to a survey (Singer et al., 1995), increase item and unit nonresponse (Reamer, 1979; Keusch et al., 2019), and overall trust in federal statistics (Childs et al., 2019).

Study 1 used traditional, interviewer-administered cognitive interviews to assess qualitative reactions to, and comprehension of the pledge language. This approach is particularly effective to gauge baseline reactions of how former respondents perceived the pledge language. Study 2 used self-administered, online testing to validate and expand on the findings of Study 1 with a larger sample, examining completion rate and item nonresponse by pledge type, but could only assess hypothetical respondent behavior. Study 3 used an experimental design to validate and add value to the previous studies by assessing the impact of different pledge versions on participation over time. Results were combined to draw a well-supported conclusion about the potential impact of the new language. We discuss how using multiple pretesting modes and the addition of online testing can provide a fuller and more complete evaluation in survey pretesting.

# Study 1

## Overview

Study 1 used traditional, interviewer-administered cognitive interviews to collect qualitative data from former BLS survey respondents. This study was exploratory, without formal hypotheses, and instead aimed to understand former respondents' recollection of how BLS keeps data confidential, and to gauge their reactions to, and comprehension of the original and revised pledge language (see Appendix A for the complete language).

## Methods

A total of 23 former BLS survey respondents participated in semi-structured cognitive interviews via telephone lasting an average of 30 minutes, administered by two trained interviewers conducted over one month. Respondents were contacted from a list of prior BLS respondents who had recently responded to a BLS survey, and were asked to participate in a debriefing interview. Demographic or other quotas were not used, but efforts were made to get a mix of respondents who participated in various BLS surveys. Respondents were asked what they recalled about their previous participation in BLS surveys, the confidentiality assurances given, and whether they had any concerns about providing their information. Afterward, they were provided with the original pledge language verbally and visually (via email) followed by the revised pledge language. Concurrent probes were used to gauge respondents' initial reactions and comprehension of the language, whether the language would affect their decision to participate in a BLS survey, and any concerns with the language.

## Results

### *Reaction to pledge language*

The open-ended nature of traditional cognitive testing allowed respondents to express their reactions to the pledge language. Most respondents (21 of 23) had no concerns about either the original or revised pledges, stating "it sounds fine" or "it sounds like the standard language that always accompanies surveys or websites." When asked for their reactions to the revised pledge, the majority of respondents had a neutral or positive reaction. Many commented that additional cybersecurity protection was a good thing, that the data would be more secure. The few respondents who expressed concerns mentioned how broad and vague they found the language to be: "It's a pretty broad statement, you could do just about anything with it."

### *Comprehension of pledge language*

Most respondents understood the pledges correctly, as assessed using a series of questions designed to evaluate participants' understanding of the data protection being offered (see Appendix B). All respondents correctly understood that their survey responses could be combined with survey answers from other respondents to create summary statistics and most understood that their data could not be shared with private companies.

### *Recollection of pledge language*

Respondents were asked an open-ended question about what they recalled about their previous participation in BLS surveys. Most did not recall much about their previous participation. About half recalled the general topic of the survey, but none recalled the name of the survey.  Most admitted not remembering what they were told about confidentiality, with only a few noting they thought the interviewer said something about keeping the data confidential.

### *Impact of pledge language*

All participants indicated that they did not have any concerns about how BLS would keep their data confidential at the time they provided data. When asked about whether the pledge language provided would affect their decision to participate in a future BLS survey, about half felt it would not affect their decision either way,  and half felt it would have a positive impact on the decision to participate (e.g., “I’d feel more comfortable giving my data with the added security”).

### Discussion

Using traditional cognitive interviews allowed for the collection of qualitative data revealing that most respondents correctly understood and did not have serious concerns about the confidentiality language. Use of open-ended probes provided additional insight into the potential impact of the pledge language on future response to BLS surveys.  However, since these respondents had not only previously responded to a BLS survey, but also agreed to this study, these findings may indicate an overall comfort with BLS rather than providing insight into the effects of the pledge language.

## Study 2

### Overview

Study 2 used online testing to expand on the findings from Study 1 with a larger sample who had not previously participated in a BLS survey. The larger sample and online mode allowed for experiments to assess paradata, such as differences between pledge versions on response to survey items and completion rates. For instance, if one pledge version was associated with lower completion rates and higher item nonresponse, it could be indicative of heightened concerns.  Based on Study 1, it was hypothesized that participants would display similar reactions to and comprehension by pledge version, but since respondents may lack familiarity with BLS in general, they may have more concerns about confidentiality overall.

### Method

Study 2 adapted and expanded Study 1 to an online, self-administered survey. Participants were recruited from Amazon’s Mechanical Turk (MTurk) panel in the U.S. and compensated $0.75 for their participation. MTurk was used because it is a robust source of recruiting participants for evaluation and pretesting research quickly and efficiently (Mullinix et al., 2016), is comparable to other convenience samples typically used for pretesting and to quota samples limited to participants in a particular geographic region (Berinsky et al., 2012; Scanlon, 2019). Although panels such as MTurk may overrepresent some demographic groups (Fahimi et al., 2018; Chandler & Shapiro, 2016), MTurk participants have been shown to put substantial effort into their responses and provide high quality data, though they only receive small token incentives (Anson, 2018). MTurk is also unique in that available tasks are typically categorization, coding, and transcribing, reducing the probability that MTurk workers are “professional survey respondents” (e.g., Hillygus et al., 2014).

Participants completed questions about their prior survey experiences and were randomly assigned to either the original or revised pledge. Then they completed survey questions that appear on typical surveys (e.g., education, income) to explore the potential impact of the pledge on survey response, with the option to skip questions or select ‘Prefer not to say.’ Afterward, participants completed debriefing questions similar to those from Study 1.

## Results

### *Participants*

A total of 1,128 (Mean age = 36 years; 78.0% White; 63.1% with a college degree or higher) MTurk participants started the study; 264 participants partially completed the study (less than 80% of the survey), and 864 completed the study. The survey took an average of 14 minutes to complete.

### *Reaction to pledge language*

Like Study 1, a majority (78.8%) of participants indicated they would have no concerns with how BLS would keep their information confidential, regardless of pledge version ($p > .05$).

### *Comprehension of pledge language*

Using a series of closed-ended questions mirroring those from Study 1, participants’ understanding of the data protection being offered by each pledge was assessed using a series of questions (similar to those in Appendix B). We found no differences between pledge versions on beliefs about who could access information participants provided on a BLS survey (all $p > .05$). Participants who read the original pledge (46.0%) were more likely than those who read the revised pledge (38.3%) to incorrectly think that their data from a BLS survey could be given to other statistical agencies for policy making, (χ2, = 6.43, df = 2, $p = .04$). No other differences were found (all $p > .05$). Thus, most participants understood the intended meaning of both pledges.

### *Recollection of pledge language*

To mirror the Study 1 question about what prior BLS respondents recalled about their previous survey response, participants in Study 2 completed a free recall task where they listed as many words or phrases from the pledge as they could remember. On average, participants reported 5.8 words and a median of 4 words, with no difference by pledge version ($p > .05$). Of the words reported, 62.0% were actually included in the pledge.

Participants then completed a recognition task where they read a list of 15 words and marked whether each word was part of the pledge they read. Some of the words were in both pledges (e.g., “Consent”) and others were not (e.g., “Risk”). Overall, participants correctly recognized an average of 75.0% of the words from the original pledge and 43.0% of the words from the revised pledge, with no difference in correct or incorrect recognition by pledge version (all $p > .05$).

### *Impact of pledge language*

A majority of participants (59.8%) across both pledge versions indicated that how BLS kept their data confidential would play a role in their decision to participate in BLS surveys. On average, participants spent 31.4 seconds on the page with the pledge language, with no difference between pledge versions ($p > .05$) despite the revised pledge being one sentence longer.

No effect of pledge version on survey completions was observed ($p >.05$), with the overwhelming majority of participants who started the survey completing it, regardless of pledge version. Item nonresponse rates were extremely low; over 98% of participants answered each item on the survey, with no difference by pledge version ($p > .05$). This was even the case for income questions (about 1% in this study), which often have very high item nonresponse rates of up to 20% (Bollinger et al., 2013).

## Discussion

Using an online platform, Study 2 confirmed findings from Study 1. In both studies, participants understood the pledge correctly and did not have strong concerns about the pledge. Study 2 added value via a larger sample size and collection of paradata. The former allowed for statistical comparison of the two pledge versions, and the latter allowed us to confirm that the pledge version did not impact survey completion, item nonresponse, or memory for the pledge language. However, both studies only collected data at one point in time and could not assess the actual impact of pledge language on survey participation over time. This is crucial, as many BLS and other surveys ask respondents to complete multiple waves of a survey. A third study was designed to address these limitations.

# Study 3

## Overview

Study 3 used an online design with similar research questions as Study 2, but expanded to a longitudinal design to assess the impact of pledge language and survey participation over time. Additionally, given that the item nonresponse rate in Study 2 was lower than typically observed in federal surveys (Bollinger et al., 2013), language was added to the instrument to encourage participants to provide honest responses and use ‘Don’t know’ or ‘Prefer not to answer’ options rather than give incorrect responses (de Leeuw et al., 2015). The instrument also asked for participants’ commitment to provide quality data (Betts, 2016). Study 3 also assessed whether priming, or reminding participants about confidentiality concerns prior to or after reading the pledge, would affect survey outcomes (Joinson et al., 2008; Singer et al., 2005). Together, these interventions would increase the ability to detect any differences in survey outcomes by pledge version.

## Method

Study 3 followed the Study 2 protocol closely, asking about prior survey experience and demographic questions. Participants were recruited from MTurk and compensated $1.50 per survey in a longitudinal design with two surveys administered one week apart. In the first survey, participants were randomly assigned to read either the original or revised pledge. Half of the participants completed a confidentiality concern questionnaire designed to remind them of prior confidentiality concerns (see Buchanan et al., 2007). The priming questionnaire was randomly assigned to be located either before or after they read the pledge. One week later, the same participants were recontacted to participate in a second survey. They were given a brief reminder of the first survey, then asked to complete the same recollection and comprehension tasks used in Study 2.

## Results

A total of 610 MTurk participants (Mean age = 38.4 years; 51.7% male; 88.4% White; 66.8% with a college degree or higher) from the U.S. completed the first survey and 463 of those completed the second survey (Mean age = 39.3 years; 52.0% male), for a 76.0% retention rate between the two surveys. Survey 1 took 10.5 minutes and Survey 2 took 7.3 minutes on average.

### *Reaction to pledge language*

Like Studies 1 and 2, the majority of participants (65.7%) indicated that they had no or very little concern about the confidentiality pledge, while 69.2% of participants indicated that the pledge would play a role in their decision to respond to a BLS survey, with no differences based on priming location or pledge version (all $p > .05$).

### *Comprehension of pledge language*

Like Studies 1 and 2, participants generally understood the pledge language. However, pledge comprehension differed based on priming location. Participants reminded about confidentiality concerns prior to reading the pledge (30.0%) were less likely than those reminded about confidentiality concerns after reading the pledge (40.0%) to incorrectly think that data from a BLS survey could be given to other statistical agencies for policy making, F(3, 410) = 4.53, $p = .03$.

## *Recollection of pledge language*

In the free recall task, participants reported an average of 2.8 words with a median of 3 words, with no difference by pledge version (*p* > .05). Participants recalled an average of 19.8% of words they viewed in the pledge, with no difference by pledge version or priming location (all *p* > .05). In the recognition task, participants correctly recognized an average of 58.8% of the words from either pledge, with no differences by pledge version (all *p* >.05). However, controlling for pledge version, participants primed about confidentiality before reading the pledge correctly recognized a greater proportion of pledge words (62.1%) than those primed at the end (55.6%), *t*(417) = 2.62, *p* = .01. Thus, participants' memory of the pledge language declined over the week, while those primed about confidentiality prior to reading the pledge correctly recognized more of the pledge language.

## *Impact of pledge language*

Participants spent an average of 27 versus 35 seconds reading the original pledge and revised pledge, respectively (*p* > .05). There was no interaction or main effect of priming location and pledge version on overall item nonresponse rates (all *p* > .05). Figure 1 shows the item nonresponse rates (i.e., "don't know", "prefer not to say", or skips) for individual survey items. Compared to 1% overall item nonresponse in Study 2, overall item nonresponse was 11% in Study 3, suggesting that respondents may have provided better quality data (i.e., opting to not answer a question rather than guess or providing most honest responses). No interaction or main effects of priming location and pledge language were found on retention between surveys (all *p* > .05).

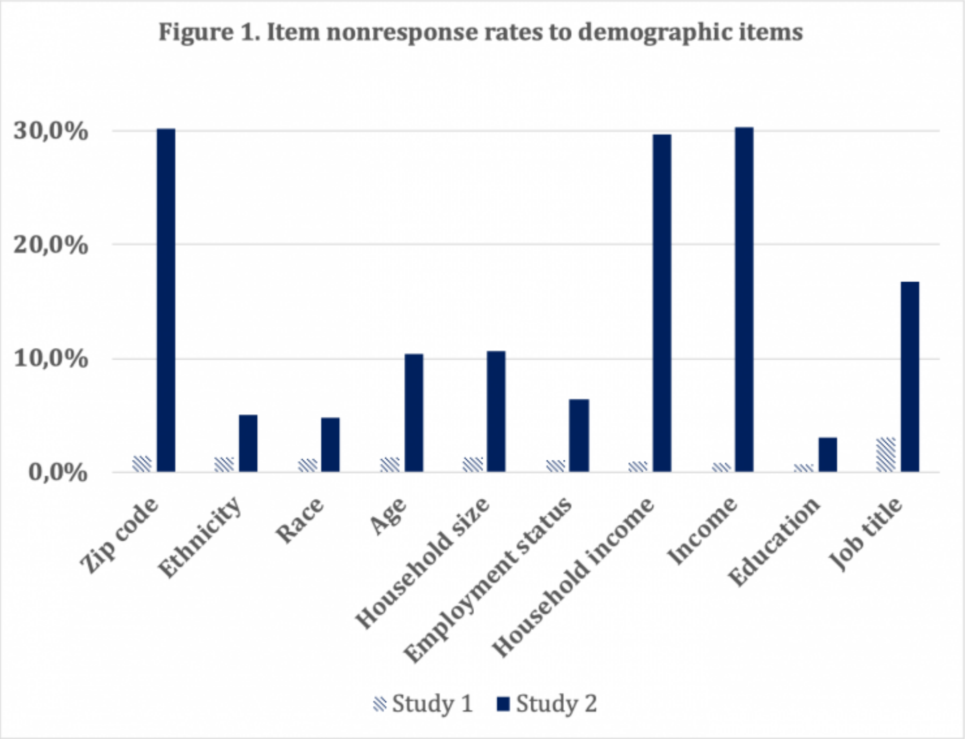
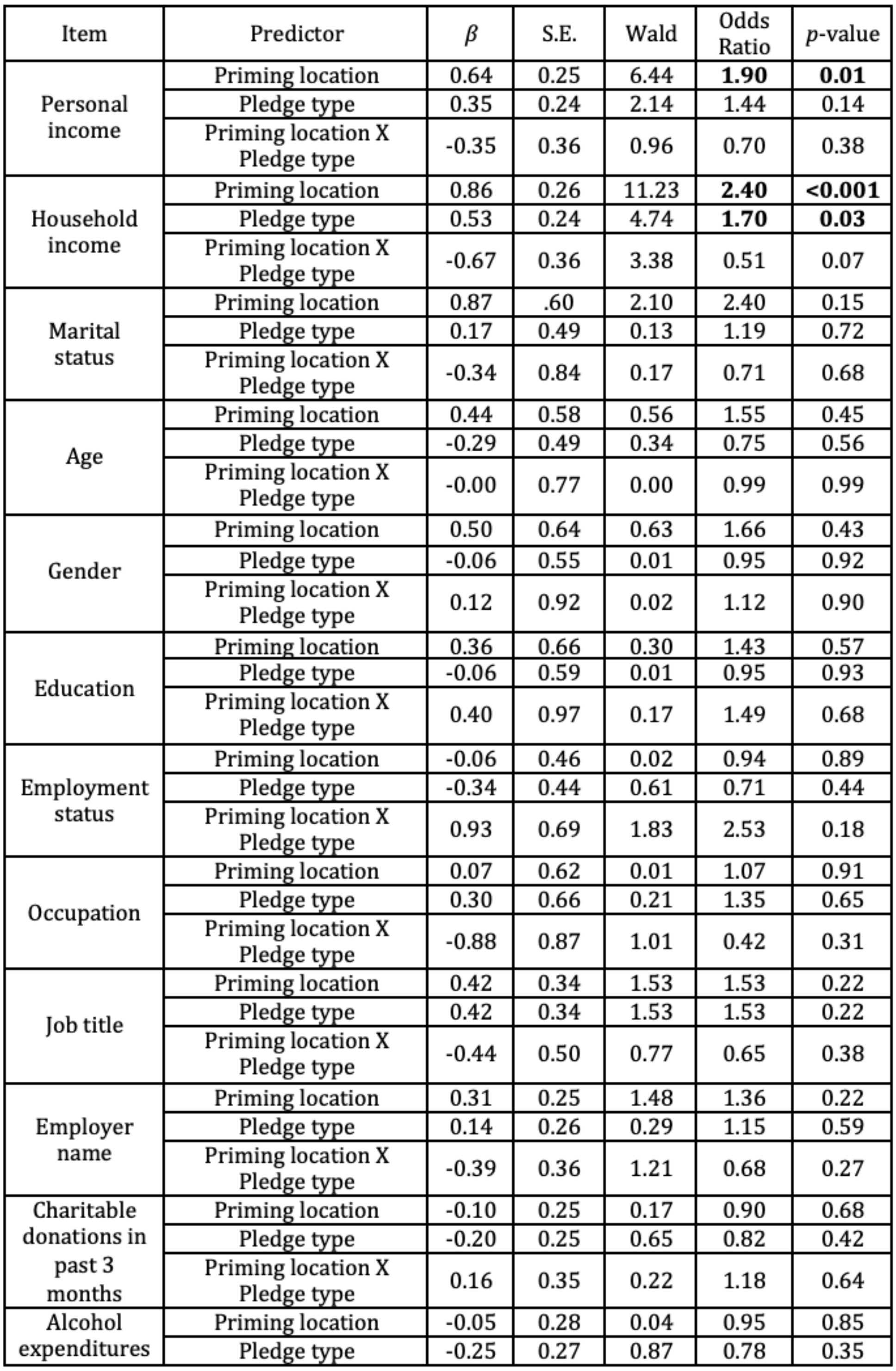
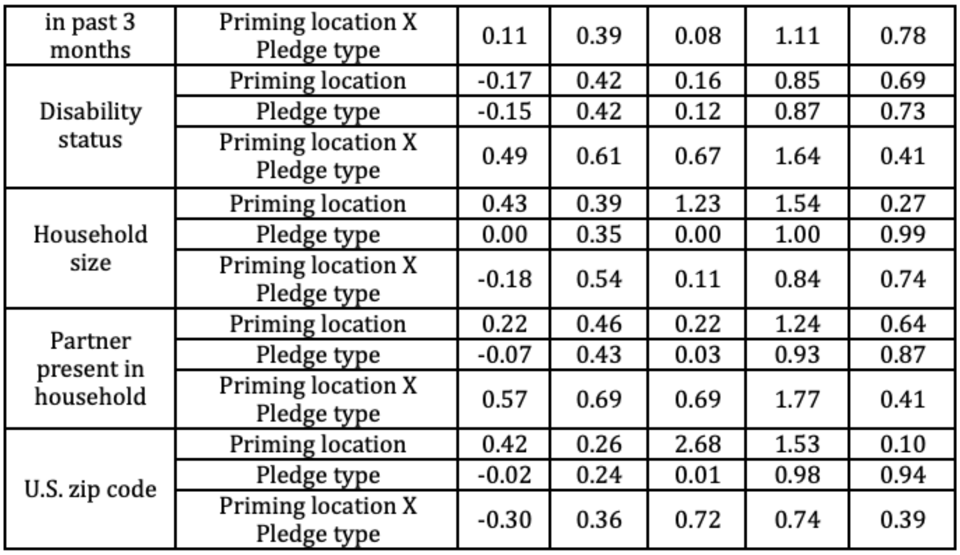


Table 1 displays results of a series of logistic regressions to determine if priming location and pledge version affected the odds of responding to demographic questions. Participants primed before reading the pledge had a 1.9 lower odds of responding to personal income and a 2.4 lower odds of responding to household income. Participants who read the revised pledge had a 1.7 lower odds of responding to household income. No interaction between priming location and pledge type was found and no other items were significant (all *ps* > .05) These findings are consistent with the idea that reminding respondents about confidentiality, and longer confidentiality pledges, can increase item nonresponse to sensitive questions (Singer et al., 1995).

**Table 1. Impact of Priming Location and Pledge Version on Odds of Responding to Demographic Questions. Priming location (0 = before pledge, 1 = after pledge), Pledge type (0 = original, 1 = revised)**



| Item | Predictor | $\beta$ | S.E. | Wald | Odds Ratio | *p*-value |
|---|---|---|---|---|---|---|
| Personal income | Priming location | 0.64 | 0.25 | 6.44 | **1.90** | **0.01** |
| | Pledge type | 0.35 | 0.24 | 2.14 | 1.44 | 0.14 |
| | Priming location X Pledge type | -0.35 | 0.36 | 0.96 | 0.70 | 0.38 |
| Household income | Priming location | 0.86 | 0.26 | 11.23 | **2.40** | **<0.001** |
| | Pledge type | 0.53 | 0.24 | 4.74 | **1.70** | **0.03** |
| | Priming location X Pledge type | -0.67 | 0.36 | 3.38 | 0.51 | 0.07 |
| Marital status | Priming location | 0.87 | .60 | 2.10 | 2.40 | 0.15 |
| | Pledge type | 0.17 | 0.49 | 0.13 | 1.19 | 0.72 |
| | Priming location X Pledge type | -0.34 | 0.84 | 0.17 | 0.71 | 0.68 |
| Age | Priming location | 0.44 | 0.58 | 0.56 | 1.55 | 0.45 |
| | Pledge type | -0.29 | 0.49 | 0.34 | 0.75 | 0.56 |
| | Priming location X Pledge type | -0.00 | 0.77 | 0.00 | 0.99 | 0.99 |
| Gender | Priming location | 0.50 | 0.64 | 0.63 | 1.66 | 0.43 |
| | Pledge type | -0.06 | 0.55 | 0.01 | 0.95 | 0.92 |
| | Priming location X Pledge type | 0.12 | 0.92 | 0.02 | 1.12 | 0.90 |
| Education | Priming location | 0.36 | 0.66 | 0.30 | 1.43 | 0.57 |
| | Pledge type | -0.06 | 0.59 | 0.01 | 0.95 | 0.93 |
| | Priming location X Pledge type | 0.40 | 0.97 | 0.17 | 1.49 | 0.68 |
| Employment status | Priming location | -0.06 | 0.46 | 0.02 | 0.94 | 0.89 |
| | Pledge type | -0.34 | 0.44 | 0.61 | 0.71 | 0.44 |
| | Priming location X Pledge type | 0.93 | 0.69 | 1.83 | 2.53 | 0.18 |
| Occupation | Priming location | 0.07 | 0.62 | 0.01 | 1.07 | 0.91 |
| | Pledge type | 0.30 | 0.66 | 0.21 | 1.35 | 0.65 |
| | Priming location X Pledge type | -0.88 | 0.87 | 1.01 | 0.42 | 0.31 |
| Job title | Priming location | 0.42 | 0.34 | 1.53 | 1.53 | 0.22 |
| | Pledge type | 0.42 | 0.34 | 1.53 | 1.53 | 0.22 |
| | Priming location X Pledge type | -0.44 | 0.50 | 0.77 | 0.65 | 0.38 |
| Employer name | Priming location | 0.31 | 0.25 | 1.48 | 1.36 | 0.22 |
| | Pledge type | 0.14 | 0.26 | 0.29 | 1.15 | 0.59 |
| | Priming location X Pledge type | -0.39 | 0.36 | 1.21 | 0.68 | 0.27 |
| Charitable donations in past 3 months | Priming location | -0.10 | 0.25 | 0.17 | 0.90 | 0.68 |
| | Pledge type | -0.20 | 0.25 | 0.65 | 0.82 | 0.42 |
| | Priming location X Pledge type | 0.16 | 0.35 | 0.22 | 1.18 | 0.64 |
| Alcohol expenditures | Priming location | -0.05 | 0.28 | 0.04 | 0.95 | 0.85 |
| | Pledge type | -0.25 | 0.27 | 0.87 | 0.78 | 0.35 |



| | | | | | | |
|---|---|---|---|---|---|---|
| in past 3 months | Priming location X Pledge type | 0.11 | 0.39 | 0.08 | 1.11 | 0.78 |
| Disability status | Priming location | -0.17 | 0.42 | 0.16 | 0.85 | 0.69 |
| | Pledge type | -0.15 | 0.42 | 0.12 | 0.87 | 0.73 |
| | Priming location X Pledge type | 0.49 | 0.61 | 0.67 | 1.64 | 0.41 |
| Household size | Priming location | 0.43 | 0.39 | 1.23 | 1.54 | 0.27 |
| | Pledge type | 0.00 | 0.35 | 0.00 | 1.00 | 0.99 |
| | Priming location X Pledge type | -0.18 | 0.54 | 0.11 | 0.84 | 0.74 |
| Partner present in household | Priming location | 0.22 | 0.46 | 0.22 | 1.24 | 0.64 |
| | Pledge type | -0.07 | 0.43 | 0.03 | 0.93 | 0.87 |
| | Priming location X Pledge type | 0.57 | 0.69 | 0.69 | 1.77 | 0.41 |
| U.S. zip code | Priming location | 0.42 | 0.26 | 2.68 | 1.53 | 0.10 |
| | Pledge type | -0.02 | 0.24 | 0.01 | 0.98 | 0.94 |
| | Priming location X Pledge type | -0.30 | 0.36 | 0.72 | 0.74 | 0.39 |

# Discussion

Study 3 built on the previous studies by examining survey response over time, allowing for an examination of whether pledge version affected survey attrition. It also added value to the prior studies by including analysis of the impact of pledge version and confidentiality concerns on participants' comprehension and recollection of the pledges, and the ability to detect differences in item nonresponse rates to sensitive questions. This provided valuable insight into the impact of the pledge version on response.

## General Discussion

Researchers have a variety of approaches available for conducting questionnaire development and pretesting, including traditional, interviewer-administered cognitive interviews and self-administered, online testing.  Several studies have compared these two pretesting modes with varying conclusions as to whether they should be considered interchangeable or complementary (Edgar et al., 2016; Behr, 2016). This research supports the approach of using both as complementary.  By using both methods sequentially and designing studies to leverage the strengths of each while addressing the limitations of the prior study, we were able to draw conclusions that would not have been possible if only one method had been used.  Table 2 outlines how the three studies built upon each other and added value to provide a more complete picture in evaluating the confidentiality language than would have been possible with a single mode alone.

Starting with traditional, interviewer-administered cognitive interviews, it was possible to explore BLS respondents' reactions, comprehension, recollection, and potential impact of the confidentiality pledge language through open-ended probes. This study provided useful baseline-level results, showing that respondents had few concerns with BLS confidentiality language, but was limited in its small sample size, highly cooperative respondents, and measurement of only hypothetical respondent behaviour.

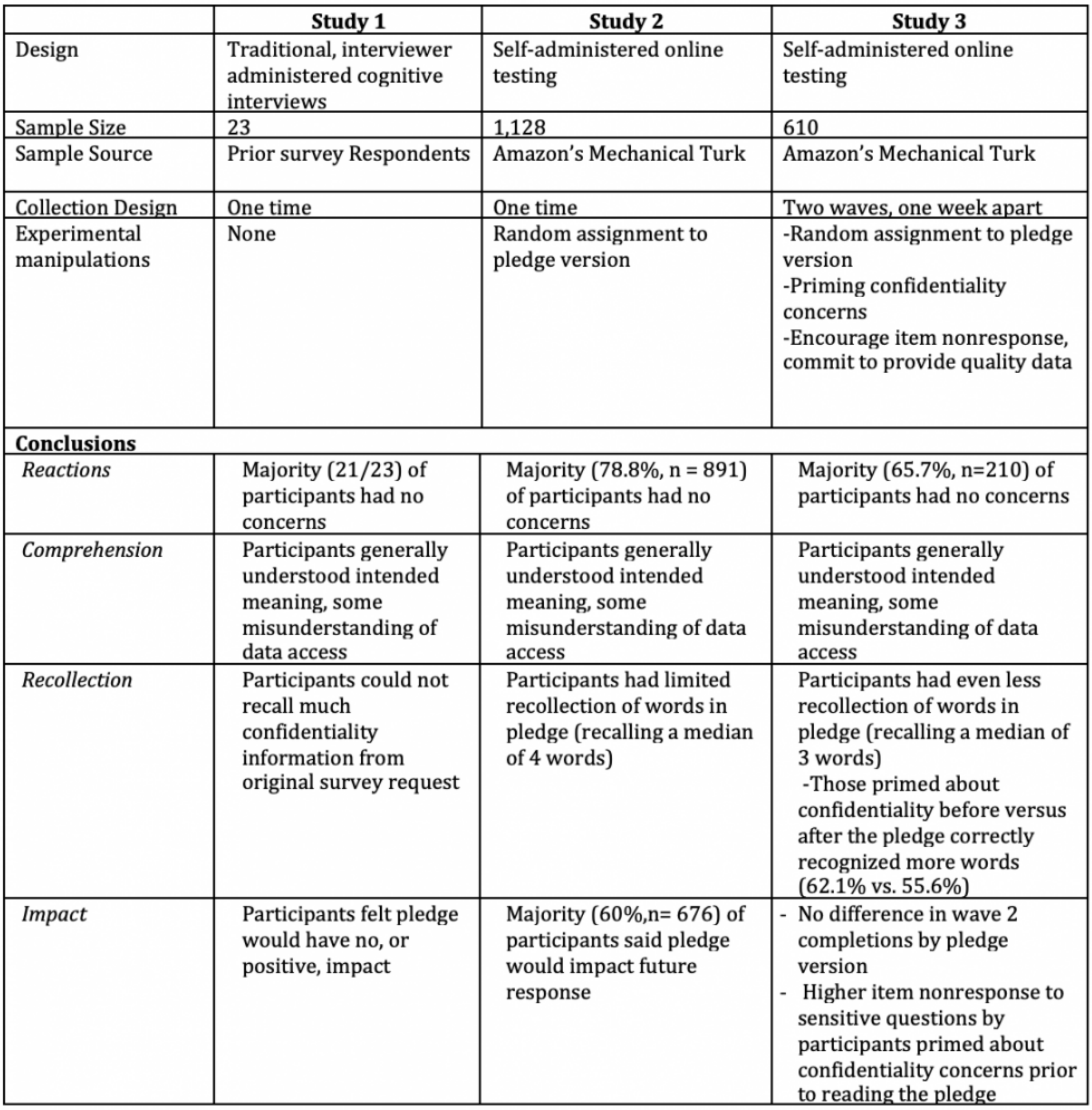
To address these limitations, a second study was conducted online to capture information from a larger number of participants who had no prior relationship with BLS surveys. The larger sample size allowed for statistical evaluation of the two pledge versions to determine its impact on survey completion rates, while also collecting information about participant reactions, comprehension, recollection, and paradata including time spent reading the pledges.  The study confirmed findings obtained from former BLS survey respondents, and added new insights about the impact of pledge wording on survey completion. However, participants were still more cooperative than typical survey respondents, as evidenced by low item nonresponse.  Additionally, since the first two studies were one-time collections, no information was available about the impact of pledge wording on participation in subsequent collection waves.

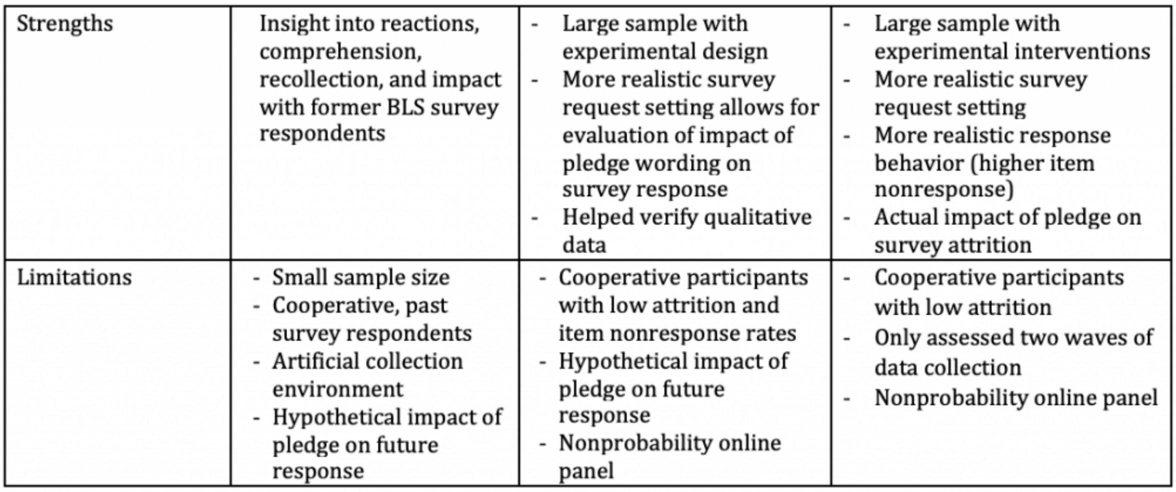
To address these limitations, a third study was designed using similar methods to the previous ones to assess the effect of pledge version on reactions, comprehension, recollection, and impact. The study added value by using a longitudinal design to assess actual survey response over time, and added interventions to improve data quality (eliciting more honest responses) and the impact of heightening confidentiality concerns on response. These interventions made it possible to detect the impact of pledge version on survey attrition and item nonresponse rates to sensitive items that was not possible in the previous studies. The longitudinal design allowed for analysis of actual participation rates in a subsequent survey rather than predicted participation.

One limitation of this research was that each study used different recruitment strategies, samples (both prior BLS respondents and from an online, nonprobability panel), and instructions/interventions. Thus, the conclusions may not be fully comparable, and should be interpreted as complementary. Further, the results are not likely generalizable to a real world environment. This limitation is common amongst most pretesting research, and would ideally be addressed through a large-scale field test to assess actual respondent reactions and response rates using each pledge version. Additionally, this research assessed attrition over only two data collection waves, and the findings may not generalize to longitudinal surveys with more waves. Resource constraints prevented a field test in this study, but is recommended in the future.

In sum, using multiple pretesting modes iteratively allowed for a fuller evaluation of the new BLS confidentiality pledge wording. The addition of online testing to complement traditional methods expanded knowledge of the impact of the pledge language on response. By leveraging the strengths of each method, and carefully designing studies to address limitations, we were able to gain insight into the impact of the pledge wording in a way that would not have been possible had only one study or mode been used.

**Table 2. Summary of Design Results, and Conclusions across Studies**





| | Study 1 | Study 2 | Study 3 |
|---|---|---|---|
| Design | Traditional, interviewer administered cognitive interviews | Self-administered online testing | Self-administered online testing |
| Sample Size | 23 | 1,128 | 610 |
| Sample Source | Prior survey Respondents | Amazon's Mechanical Turk | Amazon's Mechanical Turk |
| Collection Design | One time | One time | Two waves, one week apart |
| Experimental manipulations | None | Random assignment to pledge version | -Random assignment to pledge version<br>-Priming confidentiality concerns<br>-Encourage item nonresponse, commit to provide quality data |
| **Conclusions** | | | |
| *Reactions* | Majority (21/23) of participants had no concerns | Majority (78.8%, n = 891) of participants had no concerns | Majority (65.7%, n=210) of participants had no concerns |
| *Comprehension* | Participants generally understood intended meaning, some misunderstanding of data access | Participants generally understood intended meaning, some misunderstanding of data access | Participants generally understood intended meaning, some misunderstanding of data access |
| *Recollection* | Participants could not recall much confidentiality information from original survey request | Participants had limited recollection of words in pledge (recalling a median of 4 words) | Participants had even less recollection of words in pledge (recalling a median of 3 words)<br>-Those primed about confidentiality before versus after the pledge correctly recognized more words (62.1% vs. 55.6%) |
| *Impact* | Participants felt pledge would have no, or positive, impact | Majority (60%,n= 676) of participants said pledge would impact future response | - No difference in wave 2 completions by pledge version<br>- Higher item nonresponse to sensitive questions by participants primed about confidentiality concerns prior to reading the pledge |
| Strengths | Insight into reactions, comprehension, recollection, and impact with former BLS survey respondents | - Large sample with experimental design<br>- More realistic survey request setting allows for evaluation of impact of pledge wording on survey response<br>- Helped verify qualitative data | - Large sample with experimental interventions<br>- More realistic survey request setting<br>- More realistic response behavior (higher item nonresponse)<br>- Actual impact of pledge on survey attrition |
| Limitations | - Small sample size<br>- Cooperative, past survey respondents<br>- Artificial collection environment<br>- Hypothetical impact of pledge on future response | - Cooperative participants with low attrition and item nonresponse rates<br>- Hypothetical impact of pledge on future response<br>- Nonprobability online panel | - Cooperative participants with low attrition<br>- Only assessed two waves of data collection<br>- Nonprobability online panel |

## Appendix A – Confidentiality pledge language

### Original pledge:

The Bureau of Labor Statistics, its employees, agents, and partner statistical agencies, will use the information you provide for statistical purposes only and will hold the information in confidence to the full extent permitted by law. In accordance with the Confidential Information Protection and

Statistical Efficiency Act of 2002 (Title 5 of Public Law 107-347) and other applicable Federal laws, your responses will not be disclosed in identifiable form without your informed consent.

**Revised pledge:**

[Original pledge with the following sentence added at the end]: Per the Cybersecurity Enhancement Act of 2015, Federal information systems are protected from malicious activities through cybersecurity screening of transmitted data.

## Appendix B – Comprehension questions from Study 1 (Similar questions used in Studies 2-3)

I'm going to read a list of groups, tell me if they'd be able to see the information you provided.

1. Bureau of Labor Statistics (BLS)
2. Department of Labor (DOL)
3. Internal Review Service (IRS)
4. Social Security Administration (SSA)
5. Department of Homeland Security (DHS)
6. The Census Bureau
7. All federal government agencies
8. Congress
9. Other
10. Anyone else who would be able to access your data under this language?

I'm going to read a list of activities, tell me if you think they'd be allowed under this language?

1. It could be combined with other survey answers to create summary statistics
2. It could be published exactly as you provided it
3. It could be given to other statistical agencies
4. It could be given to the IRS
5. It could be given to other federal government agencies
6. It could be given to private companies
7. What else could be done under this language?

## *References*

1. Anson, I.G. (2018). Taking the Time? Explaining effortful participation among low-cost online survey participants. *Research and Politics, 5*(3), 1-8.
2. Behr, D. (2016). Cross-cultural web probing and how it can enhance equivalence in cross-cultural studies. Paper presented at the International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2), Miami, FL, November 9–13.
3. Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines). DOI: 10.15465/gesis-sg_en_023
4. Berinsky, A.J., Huber, G.A., & 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.
5. Betts, P.M. (2016). Asking About Sexual Orientation on the 2021 Census in England and Wales: Research into Public Acceptability, Question Design, and Administration in Online and Paper Modes. Paper presented at the International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2), Miami, FL, November 9–13.
6. Bollinger, C., Hirsch, B. T., Hokayem, C. M., & Ziliak, J. P. (2014). Trouble in the Tails? Earnings Nonresponse and Response Bias across the Distribution Using Matched Household and Administrative Data. In *Annual Meetings of the Society of Labor Economists, Arlington, VA*.

7. 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.
8. Childs, J., Clark Fobia, A., King, R., & Morales, G. (2019). Trust and Credibility in the US Federal Statistical System. *Survey Methods: Insights from the Field*. Retrieved from https://surveyinsights.org/?p=10663.
9. Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly*, *73*(1), 32-55.
10. Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4-36.
11. d'Ardenne, J., & Collins, D. (2019). Combining Multiple Question Evaluation Methods: What Does It Mean When the Data Appear to Conflict? *Advances in Questionnaire Design, Development, Evaluation and Testing*, 91-115.
12. 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*, 34(1), 116-132. doi: 0894439315573744.
13. Edgar, J., Murphy, J., & Keating, M. (2016). Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions. *SAGE Open*, *6*(4), DOI: 10.1177/2158244016671770.
14. Fowler, S., & Willis, G. B. (2019). The practice of cognitive interviewing through web probing. *Advances in Questionnaire Design, Development, Evaluation and Testing*, 451-469.
15. Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly*, *73*(2), 349-360.
16. Groves, R.M., Fowler, F.J., Couper, M.P. et al. (2009). *Survey Methodology*, 2e. Hoboken, NJ: Wiley.
17. Hillygus, D.S., Jackson, N., & Young, M. (2014). Professional Respondents in Non-Probability Online Panels. In Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 219-237). Chichester, UK: Wiley.
18. 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.
19. Kaplan, R. L., Kopp, B., & Phipps, P. (2019). Contrasting Stylized Questions of Sleep with Diary Measures from the American Time Use Survey. *Advances in Questionnaire Design, Development, Evaluation and Testing*, 671-693.
20. Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). Evaluating online nonprobability surveys. *Pew Research Center. Available at: https://www.pewresearch.org/wp-content/uploads/2016/04/Nonprobability-report-May-2016-FINAL.pdf (accessed September 2016).*
21. Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public opinion quarterly*, *83*(S1), 210-235.
22. 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.
23. Landreth, Ashley, Eleanor Gerber, and Theresa DeMaio. 2008. "Report of Cognitive Testing of Privacy and Confidentiality-Related Statements in Respondent Materials for the 2010 Decennial: Results from Cognitive Interview Pretesting with Volunteer Respondents." US Census Bureau. Statistical Research Division Report Series (Survey Methodology# 2008-4). On-line, available: www.census.gov/srd/papers/pdf/rsm2008-04.
24. Lenzner, T., & Neuert, C. E. (2017). Pretesting Survey Questions Via Web Probing–Does it Produce Similar Results to Face-to-Face Cognitive Interviewing? *Survey Practice*, *10*(4).
25. Matthijsse, S. M., de Leeuw, E. D., & Hox, J. (2015). Internet Panels, Professional Respondents, and Data Quality. *Methodology, 11*(3), 81-88.
26. McCarthy, J.S. (2019). Planning Your Multimethod Questionnaire Testing Bento Box: Complementary Methods for a Well-Balanced Test. In *Advances in Questionnaire Design, Development, Evaluation and Testing*, 723-745.
27. McClain, C. A., Couper, M. P., Hupp, A. L., Keusch, F., Peterson, G., Piskorowski, A. D., & West, B. T. (2019). A typology of web survey paradata for assessing total survey error. *Social Science Computer Review*, *37*(2), 196-213.

28. Moore, J. C., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, *16*(4), 331.
29. Mullinix, K.J., Leeper, T.J., Druckman, J.N., & Freese, J. (2015). The Generalizability of Survey Experiments. *Journal of Experimental Political Science*, 2(2), pp. 109–138. doi: 10.1017/XPS.2015.19.
30. Reamer, F. G. (1979). Protecting research subjects and unintended consequences: The effect of guarantees of confidentiality. *Public Opinion Quarterly*, *43*(4), 497-506.
31. Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who Are the Crowdworkers?: Shifting demographics in Mechanical Turk. *Conference on Human Factors in Computing Systems – Proceedings*, 2863-2872.
32. Scanlon, P. J. (2019). The Effects of Embedding Closed-ended Cognitive Probes in a Web Survey on Survey Response. *Field methods*, *31*(4), 328-343.
33. Singer, E., Von Thurn, D. R., & Miller, E. R. (1995). Confidentiality assurances and response: A quantitative review of the experimental literature. *Public Opinion Quarterly*, *59*(1), 66-77.
34. Toepoel, V., Das, M., & van Soest, A. (2008). Effects of Design in Web Surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly, 72*(5), 985–1007.
35. Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, *133*(5), 859-883.
36. Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.