

2018

The Effects of Mismatches Between Survey Question Stems and Response Options on Data Quality and Responses


Jolene Smyth

University of Nebraska-Lincoln, jsmyth2@unl.edu

Kristen Olson

University of Nebraska-Lincoln, kolson5@unl.edu

Follow this and additional works at: <http://digitalcommons.unl.edu/sociologyfacpub>

 Part of the [Family, Life Course, and Society Commons](#), [Quantitative, Qualitative, Comparative, and Historical Methodologies Commons](#), [Social Psychology and Interaction Commons](#), and the [Social Statistics Commons](#)

Smyth, Jolene and Olson, Kristen, "The Effects of Mismatches Between Survey Question Stems and Response Options on Data Quality and Responses" (2018). *Sociology Department, Faculty Publications*. 544.

<http://digitalcommons.unl.edu/sociologyfacpub/544>

This Article is brought to you for free and open access by the Sociology, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Sociology Department, Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Published in *Journal of Survey Statistics and Methodology* (2018), 32pp.

doi 10.1093/jssam/smy005

Copyright © 2018 Jolene D. Smyth and Kristen Olson. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. Used by permission.

The Effects of Mismatches Between Survey Question Stems and Response Options on Data Quality and Responses

Jolene D. Smyth and Kristen Olson

Department of Sociology, University of Nebraska-Lincoln.

Corresponding author – Jolene D. Smyth, Department of Sociology, University of Nebraska-Lincoln,
724 Oldfather Hall, Lincoln, NE 68588-0324; email jsmyth2@unl.edu

Abstract

Several questionnaire design texts emphasize a dual role of question wording: the wording needs to express what is being measured and tell respondents how to answer. Researchers tend to focus heavily on the first of these goals, but sometimes overlook the second, resulting in question wording that does not match the response options provided (i.e., mismatches). Common examples are yes/no questions with ordinal or nominal response options, open-ended questions with closed-ended response options, and check-all-that apply questions with forced-choice response options. A slightly different type of mismatch utilizes a question stem that can be read as asking for two different types of answers with no indication of which type should be provided. In this paper, we report the results of twenty-two experimental comparisons of data quality indicators (i.e., item nonresponse and response time) and response distributions across matched and mismatched versions of questions from a postal mail survey and a telephone survey. We find that mismatched items generally have lower data quality than matched items and that substantive results differ significantly across matched and mismatched designs, especially in the telephone survey. The results suggest that researchers should be wary of mismatches and should strive for holistic design.

Keywords: data quality, mismatches, questionnaire design, question wording, response time

1. Introduction

Several questionnaire design texts emphasize that questions need to simultaneously communicate the concept being measured and the answer format that adequately answers the question (e.g., Fowler 1995; Dillman, Smyth, and Christian 2014). Researchers tend to focus heavily on the first of these goals (the concept), but sometimes overlook the second (the response task), resulting in a situation in which the task communicated by the question wording does not match the task required by the response options (ROs). We use the term “mismatches” to refer to questions where the question wording does not effectively communicate what type of answer constitutes an adequate answer given the provided ROs.¹ While one can hypothesize that mismatches interrupt the response process and thus affect data quality and responses (Smit, Dijkstra, and van der Zouwen 1997), their effects have remained largely unexplored in the empirical literature. This lack of empirical evidence raises several issues. One unexplored issue is whether mismatches in existing surveys have affected estimates and data quality. Additionally, practitioners lack evidence to substantiate their advice to avoid mismatches or to determine whether they should even continue to give this advice. These issues motivated twenty-two experimental comparisons from two US national surveys designed to examine the effects of mismatches on response distributions and data quality. We report the results of those experiments in this paper.

2. Background

Creating a mismatched survey question is generally considered to be poor survey design (Dillman et al. 2014), yet it happens in practice. Figure 1 shows samples of such mismatches from existing surveys (see supplementary materials for additional examples). This is a limited set of examples to illustrate that mismatches occur; we do not know with what frequency they occur across existing surveys or how often they occur in pre-production versions of questionnaires, although they are a problem frequently identified during expert review stages of questionnaire development (e.g., problem 7 b in the Question Appraisal System is to assess a mismatch between the question and answer categories, Willis and Lessler 1999).

1. Ongena and Dijkstra (2010) refer to “mismatched answers” as respondent answers in interviewer-administered surveys that do not fit the provided ROs. This use of the term “mismatched” is slightly different than how we are using it. While mismatched answers are the most common problematic respondent verbal behavior in their study, the focus of their study was not whether a match or mismatch between the question stem and ROs produced this common respondent behavior.

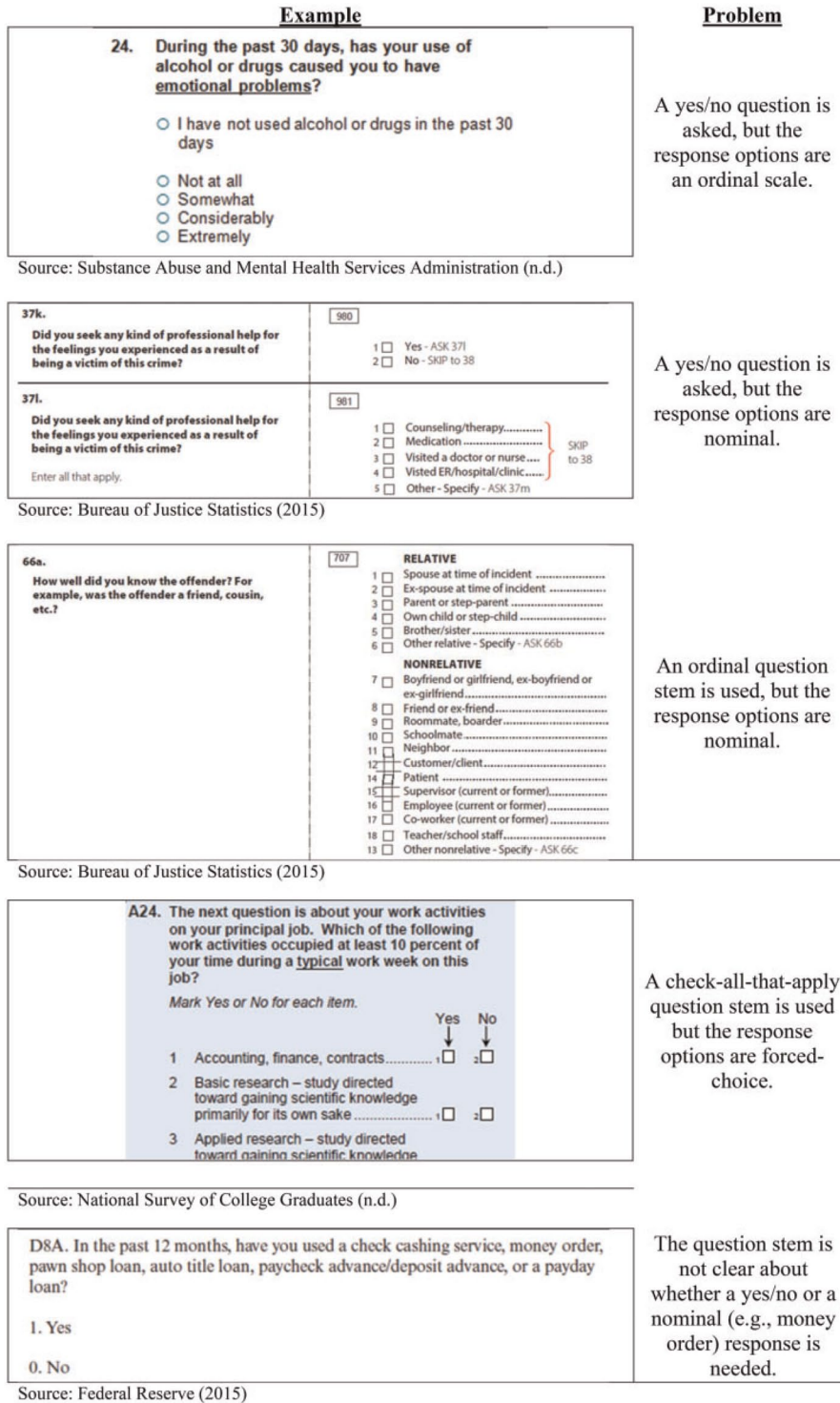


Figure 1. Examples of Mismatched Questions from Existing Surveys.

To our knowledge, only one study has experimentally assessed the impact of mismatches on data quality. In a web survey of college students, Dillman et al. (2014) experimentally varied the question stems paired with forced-choice ROs to be matched or mismatched with check-all question stems. The mismatched version resulted in higher item nonresponse and differences in response distributions compared to the matched version. However, the study is limited to a few items asked of a specialized sample that was homogenous in age and education.

Several researchers have also examined mismatches in observational studies, in particular looking at how abstract or difficult questions, including those with mismatched question stems, affect interviewer/respondent interactions (van der Zouwen 2000; van der Zouwen and Dijkstra 2002; Holbrook, Cho, and Johnson 2006). For example, van der Zouwen and Dijkstra (2002) found that having an “inadequate range of response alternatives” (i.e., mismatches and poorly designed ordinal scales) was highly associated (correlation coefficient = 0.71) with deviations from the paradigmatic interviewer/respondent interaction in a question/answer sequence (i.e., asking the question exactly as worded followed by an adequate answer [Schaeffer and Maynard 1996]). In another observational study, Olson and Smyth (2015) found no difference in response time between telephone survey questions with matched and mismatched stems and ROs. Because these studies are observational and mismatches are sometimes grouped with other question characteristics, we cannot determine the effects of mismatches alone.

Table 1 shows the full wordings of the questions examined in this paper. In each, the matched question stem accurately reflects the type of ROs provided, and the mismatched stem does not. Five types of mismatches are examined: ordinal ROs with yes/no question stems, nominal ROs with yes/no question stems, nominal ROs with ordinal question stems, forced-choice ROs with check-all-that-apply question stems, and a special case in which a nominal question stem can be answered with either general or specific information. Some questions were administered in a mail survey (“National Health, Wellbeing, and Perspectives Study,” or NHWPS), and others were administered in a telephone survey (“Work and Leisure Today II,” or WLT2).

Theoretically, mismatches should increase response difficulty. When answering survey questions, respondents must perceive the question, comprehend it, retrieve relevant information, formulate a judgment, and then report (Jenkins and Dillman 1997; Tourangeau, Rips, and Rasinski 2000). A question stem that does not match the ROs may increase comprehension difficulty. Additionally, respondents may only discover the mismatch when they try to map their answer onto the ROs and see that it does not fit (e.g., “yes” on an ordinal scale). They then need to undertake extra cognitive processing to revise their answer—or decide to not answer at all.

Table 1. Question Wording for Matched and Mismatched Treatments by Type of Mismatch**Ordinal Response Scales with Matched or Yes/No Question Stems****NHWPS – Mail Survey**

Q31 Match: In the past 12 months, how many times did you do each of the following?

Mismatch: In the past 12 months, did you do each of the following?

- A. Threaten to hit or hurt another person
- B. Push or shove another person
- C. Slap, hit, or kick another person

- Never
- Once
- Twice
- 3–4 Times
- 5 or More Times

Q32 Match: How often do you experience each of the following?

Mismatch: Do you experience each of the following?

- A. I feel safe where I live
- B. I avoid places in my town where I do not feel safe
- C. I worry about becoming a victim of a crime
- D. I worry about someone I care for becoming a victim of a crime
- E. I worry about identity theft

- Never
- Rarely
- Sometimes
- Often
- Always

Q46 Match: How concerned are you with threats to personal privacy in America today?

Mismatch: Are you concerned about threats to personal privacy in America today?

- Not At All Concerned
- A Little Concerned
- Somewhat Concerned
- Very Concerned

WLT2 – Telephone Survey

Q14 Match: How concerned are you about threats to personal privacy in America today? [READ LIST]

Mismatch: Are you concerned about threats to personal privacy in America today? [READ LIST]

- Not At All Concerned
- A Little Concerned
- Somewhat Concerned
- Very Concerned

Table 1. Question Wording for Matched and Mismatched Treatments by Type of Mismatch (continued)

Q10A Match: To what extent do you or others help define the objectives of your job?

Mismatch: Do you help define the objectives of your job?

- I mostly define the objectives
- Others mostly define the objectives
- About equal

Q10B Match: To what extent do you or others have control over the scheduling of your work?

Mismatch: Do you have control over the scheduling of your work?

- I mostly schedule my work
- Others mostly schedule my work
- About equal

Q10C Match: To what extent do you or others decide how you get your job done?

Mismatch: Do you decide how you get your job done?

- I mostly decide
- Others mostly decide
- About equal

Nominal Response Options with Matched or Yes/No Question Stem

NHWPS – Mail Survey

Q63 Match: Is your home . . .

Mismatch: Do you or someone in your household own your home?

- Owned by you or someone in this household
- Owned by you or someone in this household free and clear (without a mortgage or loan)
- Rented
- Occupied without payment of rent

Nominal Response Options with Matched or Ordinal Question Stems

NHWPS – Mail Survey

Q44 Match: Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?

Mismatch: Please describe how much you think you can trust other people.

- Most people can be trusted
- You cannot be too careful in dealing with people

Table 1. Question Wording for Matched and Mismatched Treatments by Type of Mismatch (continued)

Q45 Match: In general, would you say that you tend to be suspicious of other people or open to other people?

Mismatch: How suspicious of other people are you?

Suspicious of other people

Open to other people

WLT2 – Telephone Survey

Q12 Match: Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?

Mismatch: Generally speaking, how much do you think you can trust other people?

Most people can be trusted

You cannot be too careful in dealing with people

Q13 Match: In general, would you say that you tend to be suspicious of other people or open to other people?

Mismatch: In general, how suspicious of other people are you?

Suspicious of other people

Open to other people

Forced-Choice Response Options with Matched or Check-All Question Stems

NHWPS – Mail Survey

Q24 Match: Please indicate whether or not you have done each of the following during the past 12 months.

Mismatch: Which of the following have you done during the past 12 months?

You took a class or finished school.

You took professional development training for work.

You learned a new language or improved your language skills.

You tried to eat healthier.

You tried to exercise more.

You set up a household budget.

Table 1. Question Wording for Matched and Mismatched Treatments by Type of Mismatch (continued)

Q35 Match: Do you think that it is okay or not okay to carry a concealed weapon into each of the following locations?

Mismatch: In which of the following places do you think it is okay to carry a concealed weapon?

Stores
 Schools
 Restaurants
 Banks
 Public Parks
 College Campuses
 Daycare Centers
 Bars
 Concerts
 Airplanes

Q36 Match: Please indicate whether or not you think the Federal government should do each of the following.

Mismatch: Which of the following do you think the Federal government should do?

Legalize marijuana for medical use
 Legalize marijuana for personal use
 Legalize same-sex marriage
 Allow same-sex couples to adopt children
 Legalize carrying concealed firearms
 Legalize the sale of alcohol on Sundays

Nominal Question with Unclear Response Task

NHWPS – Mail Survey

Q30 Now for a different topic . . . Do you have a car, truck, or other vehicle?

V1: Yes/No	V2: Detailed
Yes	Car
No	Truck
Don't Know	Other vehicle
Refuse	No vehicle
	Don't know
	Refuse

We expect the increased confusion and cognitive burden caused by mismatches to manifest in lower data quality and differences in substantive responses. First, across all types of mismatches, we hypothesize that the mismatched versions of the questions will yield higher item nonresponse (H1) due to respondents skipping the item or refusing to answer because of the added burden of responding from the added comprehension and mapping problems. Second, we hypothesize that the mismatched versions of the questions will take longer to answer than the matched versions (H2) because of the extra, time-consuming cognitive processing.

With respect to substantive answers, we hypothesize that response distributions will differ between the versions with matched versus mismatched designs (H3), although the direction and type (i.e., bias or variance) of difference depend on the specific question content and type of mismatch.

The first type of mismatches shown in Table 1 are questions with ordinal ROs in which the matched question stem communicated the concept of “how much” (e.g., matched: “how many times,” “how often,” “to what extent”) and the mismatched question stem communicated that an answer of “yes” or “no” was needed. Many of these items used a unipolar scale in which a “no” answer maps to only one of the ordinal ROs (e.g., “never,” “not at all concerned”), and a “yes” answer maps to the rest (Figure 2). For these items, our specific hypotheses depend on whether we expect most respondents to have engaged in the behavior or have the concern (i.e., a “yes” answer with a more complicated mapping for those in the mismatched version) or to have not engaged in the behavior or had the concern (i.e., a “no” answer with a straightforward or one-to-one mapping). We expect most respondents will never or very rarely have engaged in the violent behaviors asked about in the Q31 NHWPS items. In the matched version, we expect those who have never done these behaviors to register a “never” response and those who have done them rarely to register a small number like “once” or “twice.” In the mismatched version, we expect those who have never done them to also

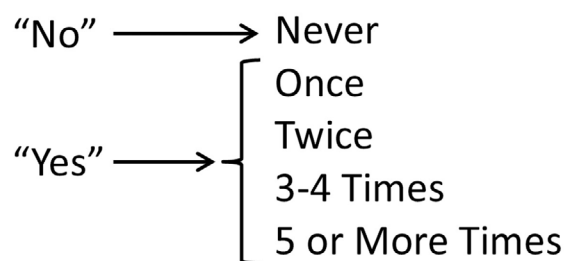


Figure 2. Illustration of How “No” Responses are Easier to Map to the Unipolar Ordinal Scale than “Yes” Responses.

map their answer to “never,” but some respondents who rarely engaged in violent behaviors to also formulate an initial “no” response, which maps most cleanly to “never.” Thus, for the Q31 violent behavior items, we hypothesize that the mismatched version will result in higher rates of selection of the “never” category (H3.Q31.A), lower means (H3.Q31.B), and lower variance (H3.Q31.C) than the matched version.

Items Q32 and Q46 in NHWPS and Q14 in WLT2 are about higher incidence and less sensitive behaviors. Because there is much less risk associated with these behaviors and opinions, we expect those who rarely do or have them to be more likely to lean toward an initial “yes” rather than an initial “no” response in the mismatched version, which then maps to one of the higher ordinal ROs. As a result, we hypothesize that there will be fewer selections of the “never” or “not at all” ROs (H3.Q32.A, H3.Q46.A, H3.Q14.A) and higher means (H3.Q32.B, H3.Q46.B, H3.Q14.B) in the mismatched than the matched versions. Having fewer initial “no” responses in the mismatched version also means there will be more people with the complicated mapping of an initial “yes” response into the corresponding ordinal categories, increasing the variability in responses at the mapping stage. Thus for these higher incidence and less sensitive items, we hypothesize higher variance in the mismatched than the matched version (H3.Q32.C, H3.Q46.C, H3.Q14.C).

The three WLT2 Q10 items are a different, but common, type of ordinal question. Notably, the “midpoint” on this type of question is sometimes placed between the other two ROs and other times placed at the end of the list of ROs, as done here. Placing the midpoint at the end of the list tends to occur in interviewer-administered surveys where the ROs may be read aloud. For example, the 2016 General Social Survey (General Social Service 2017) asks a series of questions about spending on space exploration, environment, health, and other topics. For each, respondents are asked, “are we spending too much, too little, or about the right amount on” For our items of this type, the mismatched question emphasizes “you,” while the matched version explicitly mentions both “you” and “others.” As a result, we hypothesize more respondents in the mismatched than the matched version will report that they define work objectives (H3.Q10A.A), control scheduling (H3.Q10B.A), and decide how to get their jobs done (H3.Q10C.A) because of the ease of acquiescing with the RO emphasized in the question stem.

The second type of mismatched question shown in table 1 has a yes/no question stem, but nominal ROs. Here again, the matched question stem fully represents the ROs, while the yes/no stem focuses strongly on home ownership and ignores the non-ownership ROs. Because of this, we hypothesize that compared to the matched version, respondents to the mismatched version will be more likely to choose the emphasized home ownership ROs (H3.Q63.A).

Next, we examine dichotomous nominal questions about how much one trusts and is suspicious of others. In these questions, the matched version is balanced, meaning it mentions both potential ROs. In contrast, the mismatched stems focus on only one concept from the ROs and falsely imply that the question is asking for an ordinal rating about this concept. As a result, in the mismatched version, some respondents are likely to initially put more focus on the emphasized RO and provide an ordinal response (e.g., “I can trust other people a lot”). Thus, we hypothesize that respondents in the mismatched version will be more likely than those in the matched version to select the RO that is emphasized in the mismatched question stem (i.e., “most people can be trusted” for the trust items—H3.Q44.A and H3.Q12.A—and “suspicious of other people” for the suspicion items—H3.Q45.A, H3.Q13.A).

The next type of mismatch, which only appeared in NHWPS, contained forced-choice ROs in which the question stem is either forced-choice (i.e., written to mention both ROs) or check-all (i.e., focused only on the affirmative RO). Previous research among university students has shown that respondents process check-all items less deeply than forced-choice items and, thus, are less likely to endorse individual items when they are presented in the check-all format (Smyth, Dillman, Christian, and Stern 2006); this pattern held for college students when forced-choice ROs are mismatched with a check-all question stem (Smyth 2008). Thus, we hypothesize that fewer items will be endorsed in the mismatched than the matched version (H3.FC.A), and by extension, that each individual item will be less likely to be endorsed in the mismatched than the matched version (H3.Q24.B, H3.Q35.B, H3.Q36.B). On the other hand, it is possible that the emphasis on only the positive RO in the mismatched version may make respondents more likely to endorse individual items in that version. In addition, because the mismatched check-all wording emphasizes only the positive RO, we hypothesize that respondents will treat the item as a check-all-that-apply question, and as a result, respondents will be more likely to mark answers only in the affirmative column and to leave the negative column blank in the mismatched version than in the matched version (H3.FC.C).

The final mismatch experiment differed from the others because it varied the ROs rather than the question stem. The question stem asked if one had a car, truck, or other vehicle and did not clearly communicate whether a simple “yes” or “no” response was sufficient or if the respondent was supposed to report the type of vehicle they have. For this item, we hypothesize higher item nonresponse in the version requiring more detailed answers (i.e., vehicle type) than the yes/no version (H1.Q30) because reporting the specific vehicle type rather than a simple “yes” answer requires more detailed retrieval and may cause confusion for those with multiple types of vehicles. In addition, we hypothesize that the version asking for detailed

vehicle types will take longer to answer than the yes/no version (H2.Q30) because it will take longer to retrieve the specific type of vehicle(s) and will take more words to report it. Moreover, in this version, respondents who mistakenly answer “yes” will require follow-up to get a codable answer. In the yes/ no version, if a respondent mistakenly answers with a specific type of vehicle, interviewers can confidently code their response as “yes” without additional probing. While we expect differences in data quality indicators, we hypothesize there will be no differences in the response distribution (H3.Q30) (i.e., percent reporting having a vehicle with car, truck, and other vehicle collapsed in the detailed version to be comparable to “yes” in the yes/no version) because, while it may be more difficult and time consuming, interviewers and respondents should be able to ultimately register a codable answer for those who answer the question.

Finally, we expect difficulties with the response process due to mismatches to be magnified among those with lower cognitive abilities (Krosnick 1991; Knäuper 1999); thus, we hypothesize that the effect of the mismatch will be greater for those with lower cognitive abilities compared with those with higher cognitive abilities (H4).

We examine these five types of mismatched questions using data from one postal mail and one telephone survey in the United States. Because postal mail is a visual mode, respondents have a better chance of spotting the mismatch and adjusting accordingly early in the response process. Likewise, evidence suggests that some respondents in visual modes look primarily at ROs, only looking briefly or partially at the question stem (Graesser, Cai, Louwerse, and Daniel 2006; Dillman et al. 2014) and thus may be less affected by mismatches. Therefore, we expect the effects of mismatches to be weaker in the mail mode. We expect somewhat larger effects in the telephone mode, where respondents cannot see all parts of the question at the outset and are more likely to initially formulate an answer that does not match the ROs, leading to increased problematic interactions, as have been previously seen in observational work in interviewer-administered surveys (Smit et al. 1997; Dijkstra and Ongena 2006). Our design does not allow us to explicitly or experimentally test the effects of mode, but our use of two surveys in different modes allows us to document whether effects happen in each mode and whether they seem to be of similar or different direction and magnitude.

3. Data and Methods

3.1 Data

We use two sources: the National Health, Wellbeing, and Perspectives Study (NHWPS) and the Work and Leisure Today II survey (WLT2). NHWPS was a mail survey conducted in English from April to August 2015. The survey was mailed to a simple random sample of six thousands addresses drawn from the US Postal Service's Delivery Sequence File by Survey Sampling International; 1,002 adults returned the NHWPS survey (AAPOR RR1 = 16.7 percent; AAPOR 2015). The next birthday within-household selection method was used to select one adult from each household. The questionnaire contained questions about current affairs, mental and physical health and health care, social engagement, financial well-being, crime victimization, substance use, household division of labor, and demographics.

Two experimental versions of the twelve-page questionnaire were developed. They contained the same seventy-seven individual questions, but features of each question were varied across the versions (e.g., matched versus mismatched question stems). Each version was randomly assigned to half (three thousand) of the six thousand sampled households. NHWPS also contained experiments on the use and timing of an incentive and on how the within-household selection instruction was provided.²

The WLT2 survey was a dual-frame random digit dial Computer-Assisted Telephone Interview (CATI) survey conducted by AbtSRBI in August and September 2015. For landline numbers, the Rizzo method was used to select an adult respondent from each household (Rizzo, Brick, and Park 2004). For cell phone numbers, the person answering the call was designated as the respondent. Overall, 902 people (451 in each version) completed WLT2 (AAPOR RR3 = 7.8 percent; AAPOR 2015). The questionnaire contained questions about employment, job satisfaction, volunteerism, exercise, substance use, technology use, leisure activities, and demographics. Sample members were randomly assigned to one of two experimental versions of the questionnaire in which characteristics of forty-three individual questions were systematically varied (Form A had fifty-six prompts; Form B had fifty-eight prompts).

Table 2 shows demographic distributions for respondents to both surveys by experimental treatment. All analyses are unweighted.

2. We examined interaction effects between the incentive and within-household instruction treatments and the matched versus mismatched treatments for each outcome. Out of eighty-eight tests, only four were statistically significant, consistent with Type I error, with no discernable pattern (analyses available upon request).

Table 2. Demographic Characteristics of Respondents to NHWPS and WLT2

	NHWPS				WLT2			
	Overall	Form A	Form B	χ^2	Overall	Form A	Form B	Design-adjusted F
Sex								
Male	38.60	41.09	35.91	2.59	47.88	47.77	47.99	0.00
Female	61.40	58.91	64.09		52.12	52.23	52.01	
n	917	477	440		896			
Age								
Under 65	65.77	67.05	64.38	0.79	67.29	65.63	69.96	1.53
65 and over	34.23	32.95	35.63		32.71	34.37	31.04	
n	1002	522	480		902			
Education								
HS or less	20.46	20.11	20.83	0.08	31.49	31.71	31.26	0.02
Postsecondary	79.54	79.89	79.17		68.51	68.29	68.74	
n	1002	522	480		902			
Hispanic								
Yes	5.67	5.66	5.69	0.00	7.49	8.04	6.95	0.53
No	94.33	94.34	94.31		92.51	91.96	93.05	
n	952	495	457		894			
Race								
White	83.95	84.02	83.89	0.00	79.53	80.49	78.59	0.69
Nonwhite	16.05	15.98	16.11		20.47	19.51	21.41	
n	941	488	453		816			
Income								
Less than 20K	13.61	13.35	13.87	1.20	20.00	20.33	19.69	0.25
20K-39,999	13.86	13.60	14.14		18.53	19.51	17.59	
40K-74,999	28.50	30.23	26.70		25.07	24.66	25.46	
75K+	44.03	42.82	45.29		36.40	35.50	37.27	
n	779	397	382		750			

3.2 Dependent Variables and Analysis Plan

Our first data quality indicator is the item nonresponse rate. Item nonresponse is operationalized as an indicator variable for each question, where “1” indicates that the question was unanswered in NHWPS or answered with “don’t know” or “refusal” in WLT2, and “0” indicates that a substantive response was provided.³ We compare the item nonresponse rate across the matched and mismatched versions using models that account for multiple items within persons. In the NHWPS, the experimental version is assigned at a respondent level; respondents either answer a questionnaire that contains the set of matched items or a questionnaire that contains the set of mismatched items. We thus estimate population-averaged models, also known as marginal models, in NHWPS: $\text{logit} [P(\text{miss}_{j_2})] = \beta_0 + \beta_1 I(\text{mismatched}_{j_2} = 1)$ predicting the logit of the probability of item nonresponse occurring on each question, where $\text{miss}_{j_2} = 1$ indicates that the respondent failed to answer question j_2 . We estimate these models as generalized estimating equations

with robust Huber-White sandwich estimators for the standard errors to account for the correlation of items within respondents (Agresti 2002, pp. 466–76; Raudenbush and Bryk 2002, pp. 303–4; Rabe-Hesketh and Skrondal 2012, pp. 517–9; West, Welch, and Galecki 2015, pp. 22–5). We use the “xtgee” command in Stata 15.0 with a logistic link function and an unstructured correlation matrix.

In WLT2, the matched versus mismatched version is assigned at the item level, varying within respondents. Thus, some items are matched and some are mismatched for each respondent, with responses cross-classified by both respondents and items, and both items and respondents are nested within interviewers. Thus, in WLT2, we use cross-classified multilevel random effects logistic regression models that account for the nesting of responses in interviewers, respondents and questions (Beretvas 2011, pp. 330–1). Following Beretvas’s (2011) notation, we predict the logit of the probability of item nonresponse occurring on each question, where $miss_{i(j_1, j_2)k} = 1$ indicates that the respondent failed to answer that question, as a function of an overall mean (γ_{0000}), the treatment effect on question j_2 , plus random effects due to the respondent ($u_{oj_1ok} \sim N(0, \tau_{j_1o})$), the question ($u_{ooj_2k} \sim N(0, \tau_{uj_2})$), and the interviewer ($v_{00ok} \sim N(0, \tau_{uk})$): $\text{logit}(\text{Pr}(miss_{i(j_1, j_2)k} = 1)) = \gamma_{0000} + I(\text{mismatched}_{ooj_2o} = 1) + v_{00ok} + u_{oj_1ok} + u_{ooj_2k}$. We estimate these models using “meqrlogit” in Stata 15.0, using the QR decomposition for the variance components. In both studies for both item nonresponse and response time, for parsimony, we report coefficients from models including only the experimental treatment effect; findings are unchanged if the type of mismatch is accounted for through a series of indicator variables.

Our second data quality indicator is response time, which is only available for items in WLT2. Response time was measured by the CATI system as the number of seconds the interviewer spent on each screen. To create a standardized measure, we divided response time by the number of words in the question. Because standardized item times are nested within interviewers, respondents, and questions, we use linear cross-classified random effects models to account for the clustering of item times within each of these units. We estimate these models using the mixed procedure in Stata 15.0.

We then compare the substantive responses across experimental versions. The dependent variables differ for each type of mismatch and are thus defined in Table 3. For questions in a battery in NHWPS, we estimate generalized estimation equations with robust standard errors to account for repeated measures of items within the battery and examine the variance-covariance matrix for the responses using the Box’s F-test for the equality of covariance matrices (Timm 2002), calculated using the “mvtest” covariances

3. WLT2 interviewers were trained to probe “don’t know” responses. We cannot tell from this data whether they actually did so.

Table 3. Summary of Dependent Variables and Statistical Tests for Substantive Responses by Survey

	<i>Dependent variable</i>	<i>Test</i>	<i>Hypothesis</i>
NHWPS – Mail survey			
Ordinal with yes/no mismatch			
Q31A-C	% Never	Repeated measures logistic	H3.Q31.A
	Mean	Repeated measures ANOVA	H3.Q31.B
	Covariance matrix	Box's F-test	H3.Q31.C
Q32A-E	% Never	Repeated measures logistic	H3.Q32.A
	Mean	Repeated measures ANOVA	H3.Q32.B
	Covariance matrix	Box's F-test	H3.Q32.C
Q46	% Not at all concerned	<i>t</i> test	H3.Q46.A
	Mean	<i>t</i> test	H3.Q46.B
	Variance	Levene's test	H3.Q46.C
Nominal with yes/no mismatch			
Q63	Response distribution	Chi-squared test	H3.Q63.A
Nominal with ordinal mismatch			
Q44 & Q45	Response distribution	Chi-squared test	H3.Q44.A H3.Q45.A
Forced-choice with check-all mismatch			
Q24, Q35 & Q36	# of items endorsed	Repeated measures ANOVA	H3.FC.A
	Endorsement of individual items	Repeated measures logistic	H3.Q24.B H3.Q35.B H3.Q36.B
	% using only affirmative option	Repeated measures logistic	H3.FC.C
WLT2 – Telephone Survey			
Ordinal with yes/no mismatch			
Q14	% Not at all concerned	Survey design adjusted <i>t</i> test	H3.Q14.A
	Mean	Survey design adjusted <i>t</i> test	H3.Q14.B
	Variance	Levene's test*	H3.Q14.C
Q10A-C	Response distribution	Survey design adjusted F-test	H3.Q10A.A H3.Q10B.A
Nominal with ordinal mismatch			
Q12 & Q13	Response distribution	Survey design adjusted F-test	H3.Q12.A H3.Q13.A
Unclear question stem			
Q30	Response distribution	Survey design adjusted F-test	H3.Q30

* There is no available survey design-adjusted test for equality of variances.

command in Stata 15.0. For NHWPS items that are not in batteries, we use the χ^2 test to compare the overall response distributions and the Levene's test for homogeneity of variances with "robvar" in Stata to compare variances. In WLT2, we account for clustering due to interviewers with survey design-adjusted analyses through the "svy" commands in Stata 15.0 (there are no batteries in WLT2). We compare response distributions using a survey design-adjusted *t* test or a design-adjusted *F* test (transformed from a

χ^2 test). In WLT2, we compare variances across the two versions on single item using the Levene's test.

For each outcome, we also test for interaction effects between the experimental version and age and education, which are commonly used as proxies for cognitive processing ability (Krosnick and Alwin 1987; Krosnick 1991; Knäuper 1999; Knäuper, Schwarz, Park, and Fritsch 2007). Age is categorized as "under sixty-five" versus "sixty-five and older," and education is categorized as "high school or less" versus "some postsecondary education or more." In NHWPS, age was missing for 120 respondents (11.98 percent) and education for 64 respondents (6.39 percent); in WLT2, age was missing for 34 respondents (3.77 percent) and education for 5 respondents (0.55 percent). In both surveys, we used a combination of logical imputation (e.g., years of military service included 1941 or earlier assigned to age 65+) and hot deck imputation to impute missing values.

Throughout the results, we present p values for two-sided tests. When differences are in the hypothesized direction, we interpret $p \leq 0.100$ as significant (i.e., converting to a one-sided test by dividing the p value in half for directional differences consistent with our hypothesized direction). We also present results for WLT2 Q30 separately because it differs in nature from the other mismatch items.

4. Results

4.1 Item Nonresponse

The top panel of Table 4 shows that, as hypothesized (H1), in NHWPS, the item nonresponse rate is higher in the mismatched than in the matched version. The odds of an item being left blank in the mismatched version are 1.6 times those in the matched version (two-sided $p = 0.059$; see supplementary materials for item-level nonresponse rates)⁴ In WLT2 (bottom panel of table 3), counter to H1, the item nonresponse rate was lower in the mismatched than in the matched version (odds

4.2 Response Timing

Table 5 shows that across all the items in WLT2, consistent with H2, response time was longer in the mismatched than in the matched version

4. We also examined whether the respondent skipped all items in the forced-choice grids. The mismatched version had a significantly higher rate of skipping the entire question than the matched version (mismatch coefficient = 0.89, two-sided $p = 0.053$). ratio = 0.52, $p = 0.045$).

Table 4. Odds Ratios and Robust Standard Errors Predicting Question-Level Item Nonresponse

<i>NHWPS: Population-averaged models</i>	<i>Odds ratio</i>	<i>95% CI</i>
Stem mismatch=1	1.60 [†]	(0.98, 2.60)
Constant	0.02 ^{***}	(0.013, 0.023)
Wald chi-square	3.56 [†]	
n	34,068	
# of respondents	1002	
<i>WLT2: Cross-classified multilevel random effects logistic regression models</i>		
Stem mismatch=1	0.52 [*]	(0.27, 0.99)
Constant	0.00 ^{***}	(0.00, 0.0001)
Variance Components		
Interviewer	0.00	
Question	0.69	
Respondent	73.52	
Likelihood ratio test for variance components	160.86 ^{***}	
Wald chi-square	4.01 [*]	
Log-likelihood	-337.50	
n	4,032	
# of respondents	902	

† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

(Coefficient = 0.334, $p_{0.000}$). Response times for individual questions are shown in the supplementary materials.^{5,6}

4.3 Substantive Results

4.3.1 Ordinal ROs with yes/no question stem mismatch. Table 6 compares substantive results for the first type of mismatch (see supplementary materials for full response distributions and full models for Q31 and Q32). For the Q31 items about violent behaviors, contrary to our hypotheses (H3.Q31.A, H3.Q31.B), there is no significant difference in the percentage selecting the

- The increase in seconds per word due to the mismatch significantly varied in magnitude across the different types of mismatches. Because there are other question characteristics that are confounded with the type of mismatch, such as question topic (e.g., one type of mismatch had only one question) and question format (e.g., items displayed in a grid versus individually), we cannot directly conclude whether the effect of mismatch is stronger for certain types of mismatches than for others.
- We also examined answer changes for WLT2 using the paradata from the CATI system. These are only answer changes that interviewers entered into the CATI system; the paradata misses changes that happened in conversation between the interviewer and respondent, but were not recorded in the system. Answer changes were rare, ranging from 0.22 percent to 1.33 percent across questions and versions, with no significant differences across the matched or mismatched versions (analyses available upon request).

Table 5. Coefficients and Standard Errors Predicting Seconds Per Word, Cross-Classified Random Effects Linear Regression Models, WLT2

	<i>Coef.</i>	<i>SE</i>
Stem mismatch=1	0.33***	0.019
Constant	0.98***	0.112
Random-effects		
Interviewer	0.04	0.012
Question	0.07	0.038
Respondents	0.08	0.008
Residual (response level)	0.34	0.009
Likelihood ratio test for variance components	886.12***	
Wald chi-square	308.71***	
Log-likelihood	-3905.81	
<i>n</i>	4030	
# of respondents	902	

† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 6. Substantive Results (Ordinal with Yes/No Mismatch)

	<i>Mean</i>		<i>% Never/Not at all</i>	
	<i>Matched</i>	<i>Mismatched</i>	<i>Matched</i>	<i>Mismatched</i>
NHWPS: Mail survey				
Q31A	1.17	1.15	91.71	92.80
Q31B	1.14	1.12	90.93	92.36
Q31C	1.12	1.08	93.26	94.92
<i>Coefficient</i>		-0.03		0.21
<i>z</i>		-1.04		1.04
<i>p</i>		0.297		0.297
Q32A	4.52	4.51	0.77	1.27
Q32B	3.24	3.45	16.12	11.37
Q32C	2.32	2.46	17.67	16.95
Q32D	2.76	2.88	10.06	10.81
Q32E	3.08	3.10	7.75	7.84
<i>Coefficient</i>		0.06		-0.13
<i>z</i>		1.85		-0.86
<i>p</i>		0.065		0.392
Q46	3.26	3.14	3.70	3.18
<i>t</i>		2.31		0.45
<i>p</i>		0.021		0.656
WLT2: Telephone survey				
Q14	3.13	3.00	6.05	10.51
<i>t</i>		2.06		-2.27
<i>p</i>		0.049		0.032

For the Q31 models, $n=2971$ with 992 respondents. For the Q32 models, $n=4937$ with 994 respondents. For Q46, $n=986$ and $df=984$. For Q14, $n=893$ and design adjusted $df=26$.

Two-sided *t* tests and corresponding *p* values are reported.

Bold indicates a *p* value that when divided in half to represent a one-sided test (consistent with the hypothesized direction) would be significant.

“never” category or in the mean ratings. For Q32.A-E, Q46, and Q14 (higher incidence items), we hypothesized fewer “never” or “not at all” responses and higher means in the mismatched than in the matched version (H3.Q32.A, H3.Q32.B, H3.Q46.A, H3.Q46.B, H3.Q14.A, H3.Q14.B). For both the Q32 items and Q46, there was no difference in the rate of selection of the “never” or “not at all” categories across versions. But in Q14, counter to our hypothesis, more respondents selected the “not at all” category in the mismatched than in the matched version. Consistent with our hypothesis, there was a significant positive effect of the mismatched version on the estimated means for Q32. However, for Q46 in NHWPS and Q14 in WLT2, the mean was lower in the mismatched than in the matched version, contrary to our hypotheses.

Table 7 shows the covariance matrices for these ordinal items with yes/no mismatched question stems. For the three Q31 items, the covariance matrices are not equal across the two treatments (Box $F = 17.85$, $p \leq 0.000$). As hypothesized (H3.Q31.C), the variances for each individual item (on the diagonal) are lower in the mismatched than the matched version. The same is not true for the items in Q32. Here, although the variance for three of the five individual items was higher in the mismatched version, we could not reject the hypothesis that the covariance matrices were different (Box $F = 0.94$, $p = 0.521$). The variance for Q46 was also statistically equivalent across versions, but for Q14, the variance was significantly higher in the mismatched than in the matched version as hypothesized ($F = 2.82$, $p = 0.094$).

Table 8 shows that the response distributions for WLT2 Q10 ordinal mismatched items differed across the experimental versions (Q10A: $F = 49.97$, $p = 0.000$; Q10B $F = 14.58$, $p = 0.001$; Q10C: $F = 5.86$, $p = 0.054$) in the hypothesized direction. Those in the mismatched version were more likely than those in the matched version to select the emphasized RO, reporting that they, not others, define job objectives (52.47 percent versus 19.51 percent, H3.Q10A.A), control scheduling (52.65 percent versus 39.07 percent, H3.Q10B.A), and decide how to get their job done (71.37 percent versus 61.86 percent, H3.Q10C.A).

In sum, each of the seven ordinal items had a significant difference in some form of substantive response, including the selection of a single response category, a measure of central tendency, or variability across the versions. Thus, mismatches affect substantive responses on questions with ordinal ROs.⁷

7. We also examined nondifferentiation among items that appeared in grids (a proxy for satisficing, Krosnick 1991; Couper, Tourangeau, Conrad, and Zhang 2013). We found high rates of straightlining, mostly on the “never” response option for the three low-incidence items in Q31, but no difference in rates across versions (matched 88.42% versus mismatched 87.92%) for this question. Straightlining rates on Q32 were considerably lower with the mismatched version, about twice as high (2.61%) as the matched version (1.17%), a statistically significant difference ($z = 21.66$, one-sided $p = 0.049$).

Table 7. Q31 - Covariance Matrices for Ordinal with Yes/No Mismatches

	<i>Matched</i>					<i>Mismatched</i>				
NHWPS – Mail Survey										
Q31	A	B	C			A	B	C		
A	0.420					0.351				
B	0.171	0.270				0.199	0.243			
C	0.152	0.201	0.260			0.097	0.101	0.164		
Box F (6,6890056.3)=17.58, p=0.000										
Q32	A	B	C	D	E	A	B	C	D	E
A	0.501					0.561				
B	-0.137	1.979				-0.087	1.776			
C	-0.157	0.476	0.840			-0.242	0.441	0.983		
D	-0.152	0.314	0.547	0.990		-0.191	0.355	0.645	1.102	
E	-0.080	0.207	0.434	0.414	1.180	-0.085	0.223	0.416	0.449	1.154
Box F (15,3681363.9)=0.94, p=0.521										
Q46 Variance	0.67					0.69				
Levene's Test F = 0.10, p = 0.756										
WLT2 – Telephone Survey										
Q14 Variance	0.81					1.00				
Levene's Test F=2.82, p=0.094										

For Q31, matched $n=518$, mismatched $n=471$. For Q32, matched $n=511$, mismatched $n=459$. For Q46, matched $n=514$, mismatched $n=472$. For Q14, matched $n=446$, mismatched $n=447$. For Q14 we use Levene's test, which does not account for interviewers because there is no available test for equality of variance in clustered data.

Table 8. Response Distributions for WLT2 Q10 Ordinal Items with Yes/No Mismatched Stems

	<i>Matched</i>	<i>Mismatched</i>	$\chi^2/Design$ <i>adjusted F</i>	<i>p</i>
Q10A. ($n = 428$)				
I mostly define the objectives	19.51	52.47	20.23	0.000
Others mostly define the objectives	32.68	18.83		
About equal	47.80	28.70		
Q10B. ($n = 441$)				
I mostly schedule my work	39.07	52.65	5.84	0.008
Others mostly schedule my work	38.60	37.17		
About equal	22.33	10.18		
Q10C. ($n = 442$)				
I mostly decide	61.86	71.37	2.40	0.1098
Others mostly decide	11.16	11.01		
About equal	26.98	17.62		

Q10 items were only asked of people who reported being employed.

Table 9. Response Distributions for NHWPS Question with Nominal Response Options and Matched or Yes/No Question Stems

	Matched	Mismatched	χ^2 /Design-adjusted F	<i>p</i>
Q63. (<i>n</i> = 938)				
Owned by you or someone in this household	44.40	49.66	5.81	0.121
Owned by you or someone in this household free and clear (with out a mortgage or loan)	34.83	28.19		
Rented	19.96	20.58		
Occupied without payment of rent	0.81	1.57		

4.3.2 Nominal ROs with yes/no question stem mismatch. For this item, we hypothesized higher selection of emphasized ROs in the mismatched version (H3.Q63.A). The response distribution does not significantly differ overall across the versions (Table 9). However, the differences between the versions for the individual ROs of owning one's home and owning one's home free and clear are sizeable (5.3 to 6.6 percentage points). Among those who own their home, 43.96 percent indicated owning it free and clear in the matched version compared with only 36.21 percent in the mismatched version, a significant difference of 7.75 percentage points ($\chi^2 = 4.59$, $p = 0.032$). That is, among homeowners, the mismatched question about home ownership (yes/no) increased the percentage choosing the first home ownership RO (owned) and decreased the percentage choosing the second (owned free and clear). It is unclear whether this effect of the mismatched stem is due to the relationship between the stem and the content of these two ROs or the order in which they were displayed.

4.3.3 Nominal ROs with ordinal question stem mismatch. Table 10 shows response distributions for the items with nominal ROs and ordinal question stem mismatches (H3.Q44.A, H3.Q45.A, H3.Q12.A, H3.Q13.A). The differences were in the expected direction (i.e., higher selection of the emphasized RO in the mismatched version) for three of the four items, but only one was statistically significant (Q13 in WLT2). Respondents who received the mismatched stem that focused on suspicion were more likely to report being suspicious (28.51 percent) than those who received the stem that mentioned both suspicion and openness (22.70 percent) ($\chi^2 = 3.80$, $p = 0.051$).

4.3.4 Forced-choice ROs with check-all-that-apply question stem mismatches. Contrary to our hypotheses, substantive responses across the three forced-choice items were largely unaffected by the question stem wording. We find no significant differences between the matched and mismatched versions

Table 10. Response Distributions for Questions with Nominal Response Options and Matched or Ordinal Question Stems

	<i>Matched</i>	<i>Mismatched</i>	χ^2 / <i>Design-adjusted F</i>	<i>p</i>
NHWPS: Mail survey				
Q44. (n = 976)				
Most people can be trusted	63.99	62.58	0.21	0.648
You cannot be too careful in dealing with people	36.01	37.42		
Q45. (n = 974)				
Suspicious of other people	25.20	27.68	0.77	0.379
Open to other people	74.80	72.32		
WLT2: Telephone survey				
Q12. (n = 883)				
Most people can be trusted	49.43	52.71	1.40	0.247
You cannot be too careful in dealing with people	50.57	47.29		
Q13. (n = 858)				
Suspicious of other people	22.70	28.51	3.51	0.072
Open to other people	77.30	71.49		

in the number of items endorsed (coefficient = 0.01, $p = 0.90$, analysis not shown, H3.FC.A) or in endorsement rates for individual items in these three questions (H3.Q24.B, H3.Q35.B, H3.Q36.B, see Table 11; for results by individual question and item, see the supplementary materials). Also unexpectedly, there was no significant difference across the versions in the percentage of respondents using only the affirmative ROs (“yes” or “okay”) and failing to use the negative ROs (coefficient = 20.01, $p = 0.945$, H3.FC.C).

4.4 Results for Item with Unclear Question Stem

Our final question asked whether respondents had a car, truck, or other vehicle. While the overall item nonresponse rate was very low (0.33 percent) (Table 12), as hypothesized (H1.Q30), it was significantly higher in the version requiring detailed answers (0.67 percent) than in the version requiring yes/no answers (0.0 percent) ($t = -1.94$, $p = 0.063$). Also as hypothesized, the detailed version took significantly more time to answer (11.07 seconds) than the yes/no version (7.78 seconds; $t = -6.82$, $p < 0.000$). As hypothesized, there was no difference across the two versions in the percent of respondents reporting owning a vehicle (H3.Q30). Thus, although the two versions produced similar response distributions and item missingness, the more detailed version may have required more work to produce those responses.

Table 11. Coefficients and Statistical Tests from Logistic Regression Estimated using Generalized Estimating Equations Predicting Endorsement of Items in Forced-Choice Questions

	<i>Version coefficient (mismatch=1)</i>	<i>z</i>	<i>p</i>	<i>n</i>	<i># Respondents</i>
Q24	-0.02	-0.42	0.671	5,901	995
Q35	0.12	0.94	0.347	9,768	984
Q36	-0.05	-0.65	0.516	5,865	995
All 3 questions together	0.03	0.57	0.569	21,534	997

Table 12. Question Wording and Outcomes Across Versions for a Nominal Question with Unclear Response Task in WLT2

	<i>Yes/No</i>	<i>Detailed</i>	<i>Sig. Test</i>	<i>p</i>
Item nonresponse	0.00	0.67	-1.94	0.063
Response time (seconds)	7.78	11.07		
Log response time	1.97	2.29	-6.82	0.000
Response distribution				
No	13.08	14.06	0.27	0.609
Yes	86.92	85.94		

Two-sided *t* tests are reported for item nonresponse and response time. **Bold** indicates where one-sided *t* tests consistent with hypotheses are statistically significant (i.e., two-sided *p* value divided in half). For item nonresponse, there were 902 observations, and for log response time, there were 899 observations. In both cases the design adjusted *df*=26. Chi-squared tests are reported for the response distribution. Here there were 899 observations and the design adjusted *df*= 26.

4.5 Education and Age Effects

We hypothesized that the effects of mismatches would be greater for those with lower cognitive abilities compared with those with higher cognitive abilities (H4) and tested this in each of our analyses using interaction terms between age (or education) and questionnaire version. This resulted in 27 tests for interactions with each of age and education. Only two age interactions (7 percent of all tests) were statistically significant, which is roughly what we would expect from chance alone. Thus, we conclude that the effect of mismatches does not differ by age.

For education, there were four significant interaction effects at the $p < 0.05$ level in NHWPS and none in WLT2 (16 percent of all of the tests), but two of the significant interactions occurred with different operationalizations of the same dependent variable (i.e., number of items endorsed and

endorsement of individual items for the forced choice questions). Thus, there was no evidence that the effect of mismatches differs by education in WLT2 and limited evidence that it does in NHWPS.⁸

5. Discussion

In this paper, we examined the effects of five different types of mismatched question stems on item nonresponse, response time, and response distributions using twenty-two experimental comparisons from two separate surveys. Our underlying question is whether providing question stems that do not allow respondents to correctly anticipate the ROs affects how they answer survey questions and thus affects survey estimates. While some questionnaire design texts assert that mismatches should be avoided, they appear in both preproduction and finalized questionnaires, and there is virtually no published empirical evidence of their effects. This lack of evidence leaves researchers with little empirical evidence supporting best-practices recommendations to avoid mismatches.

Our results, summarized in Table 13, indicate that mismatches are detrimental and should be avoided. In the mail survey, the mismatched version produced higher item nonresponse than the matched version. Unexpectedly, it produced lower item nonresponse in the telephone survey. In addition, the mismatched version took longer than the matched version in the telephone survey; we do not know how mismatches affect response time in the mail survey. The mismatches had mixed effects on substantive items. In the mail survey, they only significantly affected substantive results when the mismatch was a yes/no question stem with ordinal ROs, but they affected either response distributions or variance for both batteries with quantity-based response options (“never” to “5 or more times;” “never” to “always”). The difference in responses to the individual item with a “concern” response scale (Q46) was not in the hypothesized direction; future research and theoretical development should examine additional items outside batteries and without quasi-numeric response scales.

8. On Q31, lower education respondents had lower means in the mismatched than the matched version, but there was no difference for higher education respondents ($z = 1.97, p = 0.049$); this was consistent with our hypotheses. On Q44, the mismatch increased selection of the emphasized response option for low-education respondents, but there was no clear difference for high education respondents, also consistent with our hypotheses. On the forced-choice items, lower education respondents endorsed more items in the mismatched than the matched version (opposite our hypothesis), but endorsement rates did not differ by version for higher education respondents ($z = -2.15, p = 0.05$). This effect was likely driven by lower-education respondents being more likely to select “yes” on individual items in Q35 in the mismatched than in the matched version ($z = -3.46, p = 0.001$).

Table 13. Summary of Analysis Results, Overall and Interaction Effects

	Quest. #	Dependent variable	Hypothesis: In the mismatched version the DV will be . . .	Result	Effect of mismatch varies by . . .	
					Age	Edu.
NHWPS: Mail survey						
All types	All	Item nonresponse	H1: Higher	Supported	No	No
Ordinal with yes/no mismatch	Q31A-C	% Never	H3.Q31.A: Higher	Not Supported	No	No
		Mean	H3.Q31.B: Lower	Not Supported	No	Yes
		Covariance matrix	H3.Q31.C: Lower var.	Supported	n/a	n/a
	Q32A-E %	Never	H3.Q32.A: Lower	Not Supported	No	No
		Mean	H3.Q32.B: Higher	Supported	No	No
	Q46 %	Covariance matrix	H3.Q32.C: Higher var.	Not Supported	n/a	n/a
		Not at all concerned	H3.Q46.A: Lower	Not Supported	No	No
Mean		H3.Q46.B: Higher	Not Supported	No	No	
		Variance	H3.Q46.C: Higher	Not Supported	n/a	n/a
Nominal with yes/no mismatch	Q63	Response Distribution	H3.Q63.A: Higher home ownership	Not Supported	No	No
Nominal with ordinal mismatch	Q44	Response Distribution	H3.Q44.A: Higher most people trusted	Not Supported	No	Yes
	Q45	Response Distribution	H3.Q45.A: Higher suspicious	Not Supported	No	No
Forced-choice with checkall mismatch	Q24, 35, & 36	# of items endorsed	H3.FC.A: Lower	Not Supported	No	Yes
		% using only affirmative column	H3.FC.C: Higher	Not Supported	No	No
	Q24	Endorse individual items	H3.Q24.B: Lower	Not Supported	No	No
	Q35	Endorse individual items	H3.Q35.B: Lower	Not Supported	No	Yes
	Q36	Endorse individual items	H3.Q36.B: Lower	Not Supported	No	No
WLT2: Telephone survey						
All types	All but Q30	Item nonresponse # seconds on question screen/# words in question	H1: Higher H2: Higher	Not Supported Supported	No No	No No
Ordinal with yes/no mismatch	Q14	% Not at all concerned	H3.Q14.A: Lower	Not Supported	No	No
		Mean	H3.Q14.B: Higher	Not Supported	No	No
		Variance	H3.Q14.C: Higher	Supported	n/a	n/a
	Q10A	Response Distribution	H3.Q10A.A: Higher R defines work objectives	Supported	No	No
	Q10B	Response Distribution	H3.Q10B.A: Higher R controls scheduling	Supported	Yes	No
	Q10C	Response Distribution	H3.Q10C.A: Higher R decides how to do job	Supported	No	No
Nominal with ordinal mismatch	Q12	Response Distribution	H3.Q12.A: Higher most trusted	Not Supported	No	No
	Q13	Response Distribution	H3.Q13.A: Higher suspicious	Supported	No	No
Unclear question stem	Q30	Item Nonresponse	H1.Q30: Higher in detailed version	Supported	No	No
		Response Time	H2.Q30: Higher in detailed version	Supported	Yes	No
		Response Distribution	H3.Q30: No Difference	Supported	No	No

In the telephone survey, the mismatches significantly affected response distributions in five of the six questions tested (excluding Q30, which will be discussed below) and for all mismatch types examined. This is probably because the mismatches triggered additional interviewer prompts and corrections (Smit et al. 1997; Ongena and Dijkstra 2010) that may have affected responses. The evidence of substantive differences across versions was particularly strong on ordinal items with yes/no question stems in the telephone survey, when the mismatched question stems emphasized some ROs and deemphasized others. Overall, the mismatches impacted response distributions and/or data quality in both modes, although more strongly in the telephone mode, confirming our initial worries about this question design feature.

In most instances, across all of the items, the effect of the mismatch did not vary by our proxies of cognitive ability. However, where there were significant interaction effects, they tended to be with education and to indicate that lower education respondents were more strongly impacted by the mismatch than higher-education respondents.

We hypothesized and found higher item nonresponse rates in the mismatched version in the mail survey because respondents who can see the mismatch ahead of time (i.e., mail respondents) are likely confused by it, and it is easier to skip confusing items when there is no interviewer present. However, we did not anticipate that the mismatched version would produce lower item nonresponse than the matched version in the telephone survey. In retrospect, we think two forces may be at play to produce this effect. First, for telephone respondents who cannot see the ROs, the mismatched versions of the questions in this study seem to suggest an easier type of response than the matched versions. For example, for the items in Q10 and Q14, the mismatched version suggests that only a “yes” or “no” response is needed, which seems relatively easy. In the matched version, the response task, providing an ordinal response, may seem harder from the outset. Thus, more people in this version may be starting with a “don’t know” or “refuse” response. Second, telephone respondents likely do not discover the mismatch in design until they are already committed to answering the question. It is harder to revert to a “don’t know” or “refuse” answer after an initial substantive response (“yes” or “no”) is provided. As a result, telephone respondents in the mismatched version likely see the process through, ultimately providing an acceptable response and resulting in lower item nonresponse rates.

Future research should analyze interviewer/respondent interactions to explore how interviewers and respondents negotiate the answering process on matched versus mismatched questions in telephone surveys and how this affects outcomes. Such research would provide insights into why the mismatched version had lower item nonresponse, took longer (likely due to additional conversation needed to resolve mismatch-related problems), and

had different substantive responses than the matched version. In addition, we only tested one nominal item with yes/no mismatch (Q63) and only in the mail mode where respondents could see the ROs upfront; this type of mismatch was not tested in the telephone mode where it may have different effects. Further testing needs to be done on additional questions with this type of mismatch in both the mail and telephone mode before firm conclusions can be drawn. In this study, our mail and telephone surveys were two separate surveys conducted by different survey organizations, meaning our comparisons across modes are observational. Future experimental tests across modes are therefore also needed.

In the telephone survey, we examined one item with a question stem that suggested two different types of acceptable answers. While this type of item should be further tested, our results suggest researchers should carefully consider this type of mismatch and only require detailed responses when absolutely necessary, as doing so increases item nonresponse and response time. Even though the response distribution did not differ across versions, respondents to the detailed version had more difficulty answering. Future analysis of interviewer/respondent interactions should be used to examine this difficulty directly. In the meantime, if a simple yes/no response is needed, we suggest avoiding more detailed ROs. If the detail is needed, a series of yes/no questions for each detailed item captures the information adequately and has the added benefit of reducing potential confusion for respondents with multiple applicable responses.

Taken together, our findings provide heretofore missing empirical justification for advice to be wary of mismatches between the question stem and ROs in question design. They suggest that mismatches can undermine data quality in both mail and telephone modes, although they may be less detrimental when respondents can see both the question stem and response options in self-administered modes. Given these findings, analysts should carefully review both the wording and ROs in existing surveys to evaluate whether a mismatch could have affected their substantive results. More generally, survey researchers should design questions holistically, with the question stem, response options, and relationships between them simultaneously taken into consideration, and experts should continue to advocate such design.

Acknowledgments — NHWPS data collection was supported by the Office of Research and Economic Development and the Department of Sociology at the University of Nebraska-Lincoln and was fielded with assistance from the Bureau of Sociological Research at UNL. NHWPS data processing and analysis were supported by Cooperative Agreements with the USDA-National Agricultural Statistics Service supported by the National Science Foundation National Center for Science and Engineering Statistics (58-3AEU-0-0020, 58-AEU-5-0023 to JS and KO). Collection,

processing, and analysis of WLT2 were supported by a grant from the National Science Foundation (SES- 1132015 to KO). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, NCSES, USDA, or other funders.

Supplementary Materials follow the References.

References

- Agresti, A. (2002), *Categorical Data Analysis, Second Edition*, Hoboken, NJ: John Wiley & Sons, Inc.
- American Association for Public Opinion Research (2015), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. <http://www.aapor.org> (accessed March 13, 2018).
- Beretvas, S. N. (2011), "Cross-Classified and Multiple-Membership Models," in *Handbook of Advanced Multilevel Analysis*, eds. J. J. Hox, and J. K. Roberts, pp. 313–334, New York: Routledge.
- Bureau of Justice Statistics (2015), "National Crime Victimization Survey Crime Incident Report." https://www.bjs.gov/content/pub/pdf/ncvs15_cir.pdf (accessed March 13, 2018).
- Couper, M. P., R. Tourangeau, F. G. Conrad, and C. Zhang (2013), "The Design of Grids in Web Surveys," *Social Science Computer Review*, 31, 322–345.
- Dijkstra, W., and Y. Ongena (2006), "Question-Answer Sequences in Survey-Interviews," *Quality and Quantity*, 40, 983–1011.
- Dillman, D. A., J. D. Smyth, and L. M. Christian (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, Hoboken, NJ: John Wiley & Sons.
- Federal Reserve (2015), "Report on the Economic Well-Being of US Households in 2014" <https://www.federalreserve.gov/econresdata/2014-report-economic-well-being-us-households-201505.pdf> (accessed March 13, 2018).
- Fowler, F. J. (1995), *Improving Survey Questions*, Thousand Oaks, CA: Sage.
- General Social Survey (2017), "English—GSS2016 Ballot1." <http://gss.norc.umd.edu/Get-Documentation/questionnaires> (accessed March 13, 2018).
- Graesser, A. C., Z. Cai, M. M. Louwerse, and F. Daniel (2006), "Question Understanding Aid (QUAID): A Web Facility That Tests Question Comprehensibility," *Public Opinion Quarterly*, 70, 3.
- Holbrook, A., Y. I. Cho, and T. Johnson (2006), "Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties," *Public Opinion Quarterly*, 70, 565–595.
- Jenkins, C., and D. A. Dillman (1997), "Towards a Theory of Self-Administered Questionnaire Design," in *Survey Measurement and Process Quality*, eds. L. E. Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, pp. 165–196, New York: John Wiley & Sons.

- Knäuper, B. (1999), "The Impact of Age and Education on Response Order Effects in Attitude Measurement," *Public Opinion Quarterly*, 63, 347-370.
- Knäuper, B., N. Schwarz, D. Park, and A. Fritsch (2007), "The Perils of Interpreting Age Differences in Attitude Reports: Question Order Effects Decrease with Age," *Journal of Official Statistics*, 23, 515-528.
- Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5, 213-236.
- Krosnick, J. A., and D. F. Alwin (1987), "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement," *Public Opinion Quarterly*, 51, 201-219.
- National Survey of College Graduates (n.d.), "2015 National Survey of College Graduates New Respondents Questionnaire." <https://www.nsf.gov/statistics/srvygrads/#qs> (accessed March 13, 2018).
- Olson, K., and J. D. Smyth (2015), "The Effect of CATI Questions, Respondents, and Interviewers on Response Time," *Journal of Survey Statistics and Methodology*, 3, 361-396.
- Ongena, Y. P., and W. Dijkstra (2010), "Preventing Mismatch Answers in Standardized Survey Interviews," *Quality & Quantity*, 44, 641-659.
- Rabe-Hesketh, S., and A. Skrondal (2012), *Multilevel and Longitudinal Modeling Using Stata, Third Edition, Volume II: Categorical Responses, Counts, and Survival*, College Station, TX: Stata Press.
- Raudenbush, S. W., and A. S. Bryk (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.), Newbury Park, CA: Sage.
- Rizzo, L., J. M. Brick, and I. Park (2004), "A Minimally Intrusive Method for Sampling Persons in Random Digit Dial Surveys," *Public Opinion Quarterly*, 68, 267-274.
- Schaeffer, N. C., and D. W. Maynard (1996), "From Paradigm to Prototype and Back Again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews" in *Answering Questions Methodology for Determining Cognitive and Communicative Processes in Survey Research*, eds. N. Schwarz, and S. Sudman, pp. 65-88, San Francisco: Jossey-Bass.
- Smit, J. H., W. Dijkstra, and J. van der Zouwen (1997), "Suggestive Interviewer Behavior in Surveys: An Experimental Study," *Journal of Official Statistics*, 13, 19-28.
- Smyth, J. D. (2008). "Unresolved Issues in Multiple-Answer Questions," paper presented at the American Association for Public Opinion Research annual conference, May 15-18, 2008, New Orleans, LA.
- Smyth, J. D., D. A. Dillman, L. M. Christian, and M. J. Stern (2006), "Comparing Check-All and Forced-Choice Formats in Web Surveys," *Public Opinion Quarterly*, 70, 66-77.
- Substance Abuse and Mental Health Services Administration (n.d.), "National Substance Abuse, HIV, and Hepatitis Prevention Initiative, Cohort 6 Adult Baseline Questionnaire," <https://www.samhsa.gov> (accessed March 13, 2018).

- Timm, N. H. (2002), *Applied Multivariate Analysis*, New York: Springer-Verlag.
- Tourangeau, R., L. J. Rips, and K. Rasinski (2000), *The Psychology of Survey Response*, New York: Cambridge University Press.
- van der Zouwen, J. (2000), "An Assessment of the Difficulty of Questions Used in the ISSP Questionnaires, the Clarity of Their Wording, and the Comparability of the Responses," *ZA-Information*, 46, 96–114. <https://www.ssoar.info/ssoar/handle/document/19936> (accessed March 13, 2018).
- van der Zouwen, J., and W. Dijkstra (2002), "Testing Questionnaires Using Interaction Coding." In *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, eds. D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, and J. van der Zouwen, pp. 427–448, New York: Wiley.
- West, B. T., K. B. Welch, and A. T. Galecki (2015), *Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition* (2nd ed.), Boca Raton, FL: CRC Press.
- Willis, G., and J. T. Lessler (1999), *Question Appraisal System QAS-99*, Rockville, MD: Research Triangle Institute.

Supplemental Materials

Additional Examples of Mismatches from Existing Surveys

<p>B3 PolActiv CARD 7 Do you think that you could take an active role⁶ in a group involved with political issues? Please use this card.</p> <table><tr><td>Definitely not</td><td>1</td></tr><tr><td>Probably not</td><td>2</td></tr><tr><td>Not sure either way</td><td>3</td></tr><tr><td>Probably</td><td>4</td></tr><tr><td>Definitely</td><td>5</td></tr><tr><td>(Don't know)</td><td>8</td></tr></table>	Definitely not	1	Probably not	2	Not sure either way	3	Probably	4	Definitely	5	(Don't know)	8
Definitely not	1											
Probably not	2											
Not sure either way	3											
Probably	4											
Definitely	5											
(Don't know)	8											
<p>Source: European Social Survey. (2002), ESS1 Source Main Questionnaire. Retrieved April 19, 2016 from http://www.europeansocialsurvey.org/data/round-index.html.</p>												
<p>J3 Did you feel that the respondent tried to answer the questions to the best of his or her ability?</p> <table><tr><td>Never</td><td>1</td></tr><tr><td>Almost never</td><td>2</td></tr><tr><td>Now and then</td><td>3</td></tr><tr><td>Often</td><td>4</td></tr><tr><td>Very often</td><td>5</td></tr><tr><td>Don't know</td><td>8</td></tr></table>	Never	1	Almost never	2	Now and then	3	Often	4	Very often	5	Don't know	8
Never	1											
Almost never	2											
Now and then	3											
Often	4											
Very often	5											
Don't know	8											
<p>Source: European Social Survey. (2015), ESS7 Source Main Questionnaire. Retrieved April 19, 2016 from http://www.europeansocialsurvey.org/data/round-index.html.</p>												
<p>Q30. [From [INSERT 12-MONTH REFERENCE PERIOD]], have any eligible employees been denied Family and Medical Leave because FMLA did not cover the reason for their leave? [HYPERLINK “eligible employees” AND “FMLA”]</p> <table><tr><td>1</td><td>All</td></tr><tr><td>2</td><td>Most</td></tr><tr><td>3</td><td>Some</td></tr><tr><td>4</td><td>None</td></tr><tr><td>9</td><td>REF</td></tr></table>	1	All	2	Most	3	Some	4	None	9	REF		
1	All											
2	Most											
3	Some											
4	None											
9	REF											
<p>Source: Daley, Kelly, Courtney Kennedy, Marci Schalk, Julie Pacer, Allison Ackermann, Alyssa Pozniak, and Jacob Klerman. (2013), Family and Medical Leave in 2012 Methodology Report—Appendices. Retrieved February 23, 2017 from https://www.dol.gov/asp/evaluation/completed-studies/Family_Medical_Leave_Act_Survey/METHODOLOGY_APPE NDIX_family_medical_leave_act_survey.pdf.</p>												

8) Have you had sex with a woman in the last year (12 months)? (mark all that apply ☒)

(1) Vaginal sex (penis in vagina)

(1) Anal sex (penis in anus (butt))

(1) Oral sex (mouth on penis, vagina, or anus)

(1) I have not had sex with a woman in the last year.

Source: California Health and Human Services. (n.d.), Client Assesment Questionnaire. Retrieved April 19, 2016 from <http://www.cdph.ca.gov/programs/aids/Documents/LEOcdph8458CHIVClientAsmntQuest110107.pdf>.

QA15_B21 Were you told that you had Type 1 or Type 2 diabetes?

AB51

[IF NEEDED, SAY: "Type 1 diabetes results from the body's failure to produce insulin and is usually diagnosed in children and young adults. Type 2 diabetes results from insulin resistance and is the most common form of diabetes."]

TYPE 12

TYPE 22

ANOTHER TYPE (SPECIFY: _____)91

DOUBLE DIABETES (TYPE 1 AND TYPE 2)4

REFUSED-7

DON'T KNOW-8

Source: California Health and Human Services. (2016), CHIS2015 Adult Questionnaire Version 2.73. Retrieved February 23, 2017 from <http://healthpolicy.ucla.edu/chis/design/Pages/questionnairesEnglish.aspx>

147a. Did YOU lose any (other) time from work because of this incident for such things as cooperating with a police investigation, testifying in court, or repairing or replacing damaged or stolen property?

Probe: Any other reason?

Enter all that apply.

874

1 Police related activities

2 Court related activities

3 Repairing damaged property

4 Replacing stolen items

5 Other - Specify - ASK 147b

6 None (did not lose time from work for any of these reasons) - SKIP to 151

Source: Bureau of Justice Statistics. (2015), National Crime Victimization Survey Crime Incident Report. Retrieved February 23, 2017 from https://www.bjs.gov/content/pub/pdf/ncvs15_cir.pdf.

H24. Did that payment include any of the following?
READ OUT

	Yes	No	Don't know
a) A mortgage protection policy?	1	2	8
b) Building structure insurance?	1	2	8
c) Contents or possessions insurance?	1	2	8
d) Any other extra payments?	1	2	8

Source: British Household Panel Survey. (n.d.), BHPS Questionnaires and Survey Documents – Wave 16 Questionnaires and Showcards. Retrieved February 23, 2017 from https://www.iser.essex.ac.uk/bhps/documentation/pdf_versions/survey_documents/wave16/index.html.

K10. Which of the following are sources of funds for you and your spouse in

retirement?	Yes	No
a. Social Security		
b. I have a job		
c. My spouse/partner has a job		
d. Defined benefit pension from work (i.e., pension based on a formula, your earnings, and years of service)		
e. 401(k), 403(b), thrift, or other defined contribution pension plan from work		
f. Individual Retirement Account (IRA)		
g. Savings outside a retirement account (e.g., a brokerage account, savings account)		
h. Income from real estate or the sale of real estate		
i. Income from a business or the sale of a business		
j. Relying on children, grandchildren, or other family		
k. Other retirement savings		
Source: Federal Reserve. (2015), Report on the Economic Well-Being of U.S. Households in 2014. Retrieved February 23, 2017 from https://www.federalreserve.gov/econresdata/2014-report-economic-well-being-us-households-201505.pdf . This question was retyped to save space by eliminating programming notes.		

<p>If M3 = 1 ASK</p> <p>RM1a) 'Has a doctor ever told you that you suffer from osteoarthritis, rheumatoid arthritis, gout or osteoporosis?'</p>	<p>PHLPRXA</p> <p>Yes..... 1</p> <p>No..... 2</p>
Source: British Household Panel Survey. (n.d.), BHPS Questionnaires and Survey Documents – Wave 16 Questionnaires and Showcards. Retrieved February 23, 2017 from https://www.iser.essex.ac.uk/bhps/documentation/pdf_versions/survey_docs/wave16/index.html .	

<p>b. If this company had a computer system business continuity or disaster recovery program, was it tested, used in an emergency situation and/or updated in 2001? <i>Mark (X) all that apply.</i></p> <p>211</p> <p>01 <input type="checkbox"/> Tested</p> <p>02 <input type="checkbox"/> Used in emergency situation</p> <p>03 <input type="checkbox"/> Updated</p> <p>04 <input type="checkbox"/> None of the above</p> <p>05 <input type="checkbox"/> Don't know</p> <p>06 <input type="checkbox"/> Not applicable</p>
Source: Department of Justice. (n.d.), 2001 Computer Security Survey. Retrieved February 23, 2017 from https://www.census.gov/eos/www/css/cssprimary.pdf .

Q68. Do you have specific computer software or a person in human resources that tracks use of family and medical leave?

1 Computer software
 2 Designated person in human resources
 3 Both computer software and designated HR person
 4 Other method of tracking FMLA leave, please specify: _____
 [HYPERLINK "FMLA"]
 5 Do not track family and medical leave
 9 REF

Source: Daley, Kelly, Courtney Kennedy, Marci Schalk, Julie Pacer, Allison Ackermann, Alyssa Pozniak, and Jacob Klerman. (2013), Family and Medical Leave in 2012 Methodology Report—Appendices. Retrieved February 23, 2017 from https://www.dol.gov/asp/evaluation/completed-studies/Family_Medical_Leave_Act_Survey/METHODOLOGY_APPE_NDIX_family_medical_leave_act_survey.pdf.

<p>36a.</p> <p>Being a victim of crime affects people in different ways. Next I would like to ask you some questions about how being a crime victim may have affected you.</p> <p>Did being a victim of this crime lead you to have significant problems with your job or schoolwork, or trouble with your boss, coworkers, or peers?</p>	<p style="text-align: center;">969</p> <p>1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No</p>
<p>Source: Bureau of Justice Statistics. (2015), National Crime Victimization Survey Crime Incident Report. Retrieved February 23, 2017 from https://www.bjs.gov/content/pub/pdf/ncvs15_cir.pdf.</p>	

Item Nonresponse Rates by Item and Version

	Matched	Mismatched	t	p
NHWPS – Mail Survey				
Ordinal response scales and matched or yes/no question stems.				
Q31				
A	0.57	1.67	-1.66	0.098
B	0.77	1.87	-1.55	0.122
C	0.57	1.67	-1.66	0.098
Q32				
A	0.38	1.67	-2.04	0.041
B	1.34	2.92	-1.74	0.082
C	1.34	1.67	-0.42	0.672
D	0.96	1.67	-0.99	0.322
E	1.15	1.67	-0.70	0.486
Nominal response options and matched or yes/no question stems.				
Q63	5.94	6.88	-0.61	0.545
Nominal response options and matched or ordinal question stems.				
Q44	2.11	3.13	-1.01	0.312
Q45	2.68	2.92	-0.23	0.822
Forced-choice response options with matched or check-all question stems				
Q24				
A	1.34	2.71	-1.55	0.122
B	1.15	3.13	-2.18	0.029
C	1.15	3.33	-2.36	0.018
D	1.15	1.46	-0.43	0.666
E	0.77	1.67	-1.31	0.191
F	1.34	3.33	-2.11	0.035
Q35				
A	1.92	2.92	-1.03	0.301
B	2.11	2.92	-0.82	0.412
C	1.92	2.92	-1.03	0.301
D	2.30	3.13	-0.81	0.420
E	2.49	3.13	-0.61	0.543
F	2.30	3.13	-0.81	0.420
G	1.72	2.71	-1.06	0.289
H	2.30	2.71	-0.42	0.678

I	2.30	2.71	-0.42	0.678
J	2.30	2.71	-0.42	0.678
Q36				
A	1.72	2.08	-0.42	0.677
B	0.96	2.50	-1.89	0.059
C	2.87	2.92	-0.04	0.968
D	1.34	3.54	-2.28	0.023
E	2.68	3.54	-0.79	0.433
F	1.92	3.54	-1.59	0.113

WLT2 – Telephone Survey

Ordinal response scales and matched or yes/no question stems.

Q14	1.11	0.89	0.37	0.715
Q10A	4.65	1.76	1.78	0.087
Q10B	0.00	0.44	-1.01	0.322
Q10C	0.00	0.00	--	--

Nominal response options and matched or ordinal question stems.

Q12	2.22	2.00	0.26	0.794
Q13	6.21	3.55	1.42	0.168

Note: For NHWPS items, n=1,002 and df=1000. For WLT2 Q12, Q13, and Q14, n=902. For WLT2 Q10A-C n=442 because these questions were only asked of those who previously reported being employed. For all WLT2 items, the design adjusted degrees of freedom is 26. Two-sided t-tests are reported, but bold is used in the p-values to designate items that are significant with a one-sided t-test consistent with the hypothesized direction of effect.

Response Times by Question for WLT2 Telephone Survey Items

		<u>Matched</u>	<u>Mismatched</u>	<u>t</u>
Ordinal Response Scales with Matched or Yes/No Question Stems				
Q14	Response time in seconds (n=893)	14.96	14.39	
	Log response time (n=893)	2.59	2.63	-0.88
Q10A	Response time in seconds (n=428)	18.41	15.01	
	Log response time (n=428)	2.83	2.58	3.44**
Q10B	Response time in seconds (n=440)	13.97	10.28	
	Log response time (n=440)	2.56	2.19	5.11***
Q10C	Response time in seconds (n=441)	12.00	9.47	
	Log response time (n=441)	2.41	2.11	3.91***
Nominal Response Options with Matched or Ordinal Question Stems				
Q12	Response time in seconds (n=883)	18.42	20.34	
	Log response time (n=883)	2.81	2.93	-2.35*
Q13	Response time in seconds (n=858)	11.04	13.55	
	Log response time (n=858)	2.31	2.49	-3.91***
Nominal Question with Unclear Response Task				
Q30	Response time in seconds (n=899)	<u>7.78</u>	<u>11.07</u>	
	Log response time (n=899)	1.97	2.29	-6.82***

Notes: T-tests are two-sided and adjusted to account for interviewers. Cases with negative response times resulting from back-ups in the questionnaire have been eliminated from calculations of response time. Response times reported here do not account for question length.

Response distributions for questions with ordinal response scales with matched or yes/no question stems.

	Matched	Mis- matched	χ^2
NHWPS – Mail Survey			
Q31A. Threaten to hit or hurt another person (n=991)			
Never	91.71	92.80	0.94
Once	3.28	2.54	
Twice	2.89	2.97	
3-4 Times	0.58	0.64	
5 or More Times	1.54	1.06	
Q31B. Push or shove another person (n=989)			
Never	90.93	92.36	1.33
Once	5.41	4.88	
Twice	2.70	1.70	
3-4 Times	0.39	0.42	
5 or More Times	0.58	0.64	
Q31C. Slap, hit, or kick another person (n=991)			
Never	93.26	94.92	2.44
Once	3.47	3.18	
Twice	1.93	1.27	
3-4 Times	0.77	0.21	
5 or More Times	0.58	0.42	
Q32A. I feel safe where I live (n=992)			
Never	0.77	1.27	1.81
Rarely	0.77	1.06	
Sometimes	5.58	4.24	
Often	31.35	32.20	
Always	61.54	61.23	
Q32B. I avoid places in my town where I do not feel safe (n=981)			
Never	16.12	11.37	7.24
Rarely	15.15	13.30	
Sometimes	22.91	22.53	
Often	20.00	24.25	
Always	25.83	28.54	
Q32C. I worry about becoming a victim of a crime (n=987)			
Never	17.67	16.95	9.63*
Rarely	43.69	35.81	

Sometimes	30.68	35.17	
Often	5.24	8.05	
Always	2.72	4.03	

Q32D. I worry about someone I care for becoming a victim of a crime (n=989)

Never	10.06	10.81	8.28+
Rarely	29.59	22.25	
Sometimes	39.65	42.58	
Often	15.67	16.95	
Always	5.03	7.42	

Q32E. I worry about identity theft (n=988)

Never	7.75	7.84	1.95
Rarely	19.77	18.64	
Sometimes	42.05	40.89	
Often	17.83	21.19	
Always	12.60	11.44	

Q46. (n=986)

Not At All Concerned	3.70	3.18	9.17*
A Little Concerned	12.65	18.86	
Somewhat Concerned	37.94	39.19	
Very Concerned	45.72	38.77	

WLT2 – Telephone Survey

Q14. (n=893)

Not At All Concerned	6.05	10.51	1.97
A Little Concerned	16.37	18.34	
Somewhat Concerned	35.65	31.54	
Very Concerned	41.93	39.60	

Note: Design adjusted F(2.82, 73.40) reported for Q14 to account for interviewers. + $p \leq 0.100$; * $p \leq 0.05$; ** $p \leq 0.010$; *** $p \leq 0.001$

Full substantive results models for the percent selecting “never” and means for Q31 and Q32 (ordinal response options with yes/no question stem mismatches)

Question 31				
	<u>Percent Selecting “Never”</u>		<u>Mean</u>	
	<u>Coefficient</u>	<u>Robust SE</u>	<u>Coefficient</u>	<u>Robust SE</u>
Stem mismatch=1	0.21	0.20	-0.03	0.03
Constant	2.45***	0.13	1.14***	0.02
Wald chi-square	1.09		1.09	
n	2,971		2,971	
# of respondents	992		992	

Question 32				
	<u>Percent Selecting “Never”</u>		<u>Mean</u>	
	<u>Coefficient</u>	<u>Robust SE</u>	<u>Coefficient</u>	<u>Robust SE</u>
Stem mismatch=1	-.13	0.15	0.06+	0.03
Constant	-2.49***	0.10	3.41***	0.02
Wald chi-square	0.73		3.41+	
n	4,937		4,937	
# of respondents	994		994	

Notes: +p<.10, *p<.05, **p<.01, ***p<.001,

Response distributions for questions with forced-choice response options with matched or check-all question stems

	Matched	Mis-matched	Sig. Test^a
Q24.			
You took a class or finished school. (n=982)	23.11	25.70	0.89
You took professional development training for work. (n=981)	33.72	32.90	0.07
You learned a new language or improved your language skills. (n=980)	13.76	10.13	3.04+
You tried to eat healthier. (n=989)	87.79	87.53	0.02
You tried to exercise more. (n=990)	80.31	80.08	0.01
You set up a household budget. (n=979)	38.45	42.03	1.30
Mean Number of Items Endorsed	2.75	2.76	-0.13
Percent using only the positive category (n=995)	3.07	3.59	-0.45
Q35.			
Stores (n=978)	37.89	35.84	0.44
Schools (n=977)	13.70	16.31	1.31
Restaurants (n=978)	34.57	33.69	0.08
Banks (n=975)	27.65	32.26	2.47
Public Parks (n=974)	38.11	35.70	0.61
College Campuses (n=975)	21.37	23.44	0.60
Daycare Centers (n=980)	13.45	11.99	0.47
Bars (n=977)	14.71	19.06	3.31+
Concerts (n=977)	17.84	20.34	0.99
Airplanes (n=977)	8.82	12.85	4.12*
Mean Number of Items Endorsed	2.27	2.39	-0.63
Percent using only the positive category (n=948)	4.09	5.96	-1.35+
Q36.			
Legalize marijuana for medical use (n=983)	79.14	78.94	0.01
Legalize marijuana for personal use (n=985)	40.62	40.81	0.00
Legalize same-sex marriage (n=973)	52.07	55.58	1.20
Allow same-sex couples to adopt children (n=978)	61.17	60.91	0.01
Legalize carrying concealed firearms (n=971)	44.29	40.39	1.51
Legalize the sale of alcohol on Sundays (n=975)	64.45	62.20	0.53
Mean Number of Items Endorsed	3.36	3.33	0.26
Percent using only the positive category (n=995)	11.90	9.49	1.22

+ $p \leq 0.100$; * $p \leq 0.05$; ** $p \leq 0.010$; *** $p \leq 0.001$

^a One-sided t-test used to test differences in mean number of items endorsed and the percent using only the positive category to reflect directional hypotheses. Chi-squared used to test response distributions.