2022

Your Best Estimate is Fine. Or is It?

Jerry Timbrook

Kristen Olson

Jolene D. Smyth

sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1097–1123, http://dx.doi.org/10.2478/JOS-2022-0047

Your Best Estimate is Fine. Or is It?

Jerry Timbrook[1], *Kristen Olson*[2], *and Jolene D. Smyth*[2]

Providing an exact answer to open-ended numeric questions can be a burdensome task for respondents. Researchers often assume that adding an invitation to estimate (e.g., "Your best estimate is fine") to these questions reduces cognitive burden, and in turn, reduces rates of undesirable response behaviors like item nonresponse, nonsubstantive answers, and answers that must be processed into a final response (e.g., qualified answers like "about 12" and ranges). Yet there is little research investigating this claim. Additionally, explicitly inviting estimation may lead respondents to round their answers, which may affect survey estimates. In this study, we investigate the effect of adding an invitation to estimate to 22 open-ended numeric questions in a mail survey and three questions in a separate telephone survey. Generally, we find that explicitly inviting estimation does not significantly change rates of item nonresponse, rounding, or qualified/range answers in either mode, though it does slightly reduce nonsubstantive answers for mail respondents. In the telephone survey, an invitation to estimate results in fewer conversational turns and shorter response times. Our results indicate that an invitation to estimate may simplify the interaction between interviewers and respondents in telephone surveys, and neither hurts nor helps data quality in mail surveys.

Key words: Estimation; enumeration; questionnaire design; data quality; respondent burden.

1. Introduction

Survey researchers often use open-ended questions to capture numeric responses for questions that require enumeration of events that occurred over a fixed time period (e.g., "How many cigarettes did you smoke in the last seven days?") or questions asking for financial information (e.g., income). Open-ended numeric questions are often used when a precise number is needed for the survey's analytic goals or when it is difficult to construct meaningful ranges for response options (Dillman et al. 2014). Yet providing an exact, numeric answer can be mentally taxing or impossible for respondents, especially when questions ask about hard-to-enumerate topics (Tourangeau et al. 2000; Conrad et al. 1998). As a result, some respondents may fail to answer a question altogether or give nonsubstantive answers (i.e., responses like "too many cigarettes to count"), which can

[1] RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709, U.S.A. Email: jtimbrook@rti.org
[2] University of Nebraska-Lincoln, 711 Oldfather Hall, Lincoln, NE 68588-0324, U.S.A. Emails: kolson5@unl.edu; jsmyth2@unl.edu

Acknowledgments: This work was supported in part by funds provided to the University of Nebraska-Lincoln under a Cooperative Agreement with the USDA-National Agricultural Statistics Service supported by the National Science Foundation National Center for Science and Engineering Statistics (58-AEU-5-0023 to J.D.S and K.O.). Additional funding was provided by the Office of Research and Economic Development and the Department of Sociology at the University of Nebraska-Lincoln, and the National Science Foundation (Grant Number SES-1132015 to K.O.). 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.

negatively affect data quality (Beatty and Herrmann 2002). Difficulty providing exact answers may also cause some respondents to report responses that must be recoded into a final answer by data processing staff (e.g., ranges like "five to ten cigarettes a week").

To ease the burden of answering numeric response questions and potentially limit undesirable answering behaviors, researchers often invite respondents to estimate their answer by adding phrases like "Your best estimate is fine" to the question stem (Dillman 2007). However, there is no research of which we are aware that establishes whether this phrase is helpful in mail and telephone surveys. Additionally, inviting estimation may also trigger changes in other response behaviors. For example, response heaping (i.e., giving rounded answers) is more prevalent when respondents estimate (Huttenlocher et al. 1990; Burton and Blair 1991; Holbrook et al. 2014), and may increase if respondents are told they can approximate their answer. Estimation can also lead respondents to overreport their behaviors, meaning that an invitation to estimate may inflate survey means (Burton and Blair 1991). Finally, it is possible that these effects vary across self-administered and interviewer-administered surveys because the cognitive and working memory demands on respondents differ across these modes (De Leeuw 2005).

In this article, we experimentally explore the effect of including an invitation to estimate worded as "Your best estimate is fine" on several indicators of data quality in a mail survey and in a separate telephone survey. We compare versions of questions with and without the invitation to estimate to answer the following research questions:

RQ1. Does the item nonresponse rate differ across versions?
RQ2. Does the rate of nonsubstantive answers differ across versions?
RQ3. Does the rate of range and qualified answers differ across versions?
RQ4. Does the heaping or rounding rate differ across versions?
RQ5. Do estimated means differ across versions?

2. Background

Questionnaire designers have several options when asking numeric questions such as how many miles respondents drive in an average week. A closed-ended question format with ranges for response options (e.g., "100 to less than 200" miles) can be simpler for respondents but yields imprecise responses. In contrast, an open-ended numeric item both cues and allows respondents to provide a precise answer (e.g., 150 miles), but respondents may have difficulty recalling such a precise answer (Tourangeau et al. 2000).

2.1. Response Strategies for Open-Ended Numeric Questions

Respondents generally use either enumeration or estimation to answer open-ended numeric questions about events or behaviors. *Enumerating* a precise answer requires respondents to retrieve each episode of the event or behavior from memory, decide if the episode fits the question's requirements, count up the total number of relevant episodes, and report that number (Blair and Burton 1987; Menon 1993; Tourangeau et al. 2000). For example, respondents often use *enumeration* to answer questions that ask about topics like *infrequent events* (e.g., trips to the emergency room) or events over *short and recent reference periods* (e.g., the past week) (Blair and Burton 1987; Burton and Blair 1991;

Conrad et al. 1998). Events that do not happen on a routine schedule (e.g., irregular events like the number of visits to the doctor for an injury) and events whose episodes are distinct from one another (e.g., number of serious illnesses) are often enumerated (Menon 1993; Conrad et al. 1998).

Alternatively, *estimation* involves providing an answer based on an approximate rate of occurrence for a target behavior or event (Blair and Burton 1987; Burton and Blair 1991; Tourangeau et al. 2000). Questions asking about *frequent events* (e.g., steps taken) or events that occur over *long reference periods* (e.g., your adult lifetime) lend themselves to estimation because their exact numbers often do not exist in memory (Blair and Burton 1987; Burton and Blair 1991; Conrad et al. 1998). Similarly, events that happen on a regular schedule (e.g., eating breakfast) and events whose episodes are similar to one another (e.g., minor illnesses) are often estimated (Menon 1993; Conrad et al. 1998). Exact answers to financial questions like one's annual income may also be difficult to remember and often estimated.

A respondent's use of enumeration versus estimation depends on how easily relevant information can be retrieved from memory (i.e., the *cognitive state*) (Beatty and Herrmann 2002); this answer may be modified if the respondent feels the need to edit the answer to a possibly *sensitive question* (i.e., the respondent's communicative intent). We start with discussing cognitive states and then move to communicative intent.

Information that requires minimal effort to retrieve (i.e., *available* information) can easily be enumerated and thus likely pose little burden to respondents to enumerate. Limits in memory, however, can prevent respondents from using enumeration (Huttenlocher et al. 1990). Estimation is commonly used when the exact information requested by a question cannot be retrieved, but a respondent can use other information in memory to approximate their answer (i.e., *generatable* information) (Beatty and Herrmann 2002). Respondents using estimation must decide whether their inexact answer meets the level of precision requested by the question (i.e., an adequacy judgement) (Beatty and Herrmann 2002). If respondents judge their estimated answer to be inadequate, they may skip the question entirely. Alternatively, respondents may indicate that their answer is estimated (i.e., potentially inadequate) by providing nonsubstantive answers (e.g., "too many to count"). Respondents may also indicate that their answer is "generated" by including additional information like ranges (e.g., "10–20") or qualifiers (e.g., "about 20") in their answer. Although these answers can ultimately be recoded into a final response using a set of rules (e.g., ranges like "10–20" coded as the lower bound of ten), they require costly post-survey processing that may introduce processing error.

Accessible information exists in a cognitive state between available and generatable and can be retrieved from memory only if a respondent exerts cognitive effort (Beatty and Herrmann 2002). The enumeration process for accessible information is burdensome; respondents must expend mental resources to retrieve each episode from memory. Respondents lacking sufficient motivation might avoid this burdensome retrieval task by estimating their answer instead, leading to less precise answers. Respondents again must make an adequacy judgment to determine if their estimated answer meets the precision requirements of the question. This judgment may lead respondents to not answer the question at all, resulting in item nonresponse, or to provide nonsubstantive, qualified, or range answers.

Finally, when exact information related to a question is not known and cannot be approximated, it is *inestimable* (e.g., apples eaten in your entire life) (Beatty and Herrmann 2002). In such cases, neither enumeration nor estimation will lead to an answer. The only honest recourse for respondents is to: (1) skip answering the question altogether; (2) select a "don't know" or "refusal" response option when available; or (3) provide a nonsubstantive answer (e.g., "Too many to count"). In general, survey designers are encouraged to avoid questions that ask for inestimable information.

Responses are modified by respondents' communicative intent, or the need to edit answers to sensitive questions (Beatty and Herrmann 2002). Respondents are more likely to skip sensitive questions (i.e., item nonresponse) if they worry their answers are socially undesirable (Beatty and Herrmann 2002; Tourangeau and Yan 2007). Alternatively, they may edit answers to fit within a social norm or expectations. This editing may manifest in behaviors that look like those of estimation – for instance, providing a range or reporting a more socially acceptable (and perhaps rounded) inexact answer.

2.2. *An Invitation to Estimate*

Appending the phrase "Your best estimate is fine" to a question is a common method for indicating that approximate answers are acceptable for numeric response questions (Dillman 2007). This phrase, inviting respondents to estimate their answers, is hypothesized to reduce item nonresponse rates by communicating that: (1) respondents do not have to engage in potentially burdensome enumeration (i.e., for accessible information), and (2) an imprecise answer is preferable to no answer (i.e., reporting generatable or estimated accessible information is acceptable), potentially making respondents' adequacy judgments less burdensome. Explicitly inviting estimation is also hypothesized to reduce nonsubstantive answers through the same mechanisms. By communicating that guesses are acceptable, an invitation to estimate is hypothesized to also reduce range or qualified answers; respondents may feel less inclined to communicate uncertainty over their answer (e.g., ranges, qualified answers) if they are told that estimation is permissible.

Despite the potential to decrease item nonresponse, an invitation to estimate may also have the unintended consequence of encouraging some respondents to estimate when they would have otherwise enumerated their answer. This can occur when the exact information requested by a question is retrievable with cognitive effort (i.e., accessible information), but the respondent approximates their answer when offered the invitation to estimate. An increase in estimation may in turn increase other potentially undesirable response behaviors. Primarily, respondents who estimate are more likely to provide rounded answers (i.e., heaping) (Burton and Blair 1991; Holbrook et al. 2014). An invitation to estimate may encourage this heaping behavior and foster inaccurate responses. Finally, as estimated answers tend to be larger than enumerated answers – that is, respondents round up rather than round down (Burton and Blair 1991) – survey means may be larger with an invitation to estimate. Therefore, when an invitation to estimate succeeds at encouraging estimation behavior, we would expect to see: lower rates of item nonresponse (H1a), nonsubstantive answers (H2a), and qualified or range answers (H3a); higher rates of heaping (H4a); and larger estimated means (H5a; summarized in Table 1).

Table 1. Summary of hypotheses.

Response outcome	Example	With an invitation to estimate (ItE). . .	And will vary by. . .	
			Cognitive state	Question sensitivity
Item nonresponse	Respondent does not answer the question	H1a: Item nonresponse rates will be lower.	H1b: Item nonresponse rates will be lower with an ItE for accessible information than other cognitive states.	H1c: Item nonresponse rates will be lower for sensitive questions with an ItE than for sensitive questions without an ItE.
Nonsubstantive answers	Respondent answers "Too many to count"	H2a: Nonsubstantive answer rates will be lower.	H2b: Nonsubstantive answer rates will be lower with an ItE for accessible information than other cognitive states.	H2c: Nonsubstantive answer rates will be lower for sensitive questions with an ItE than for sensitive questions without an ItE.
Qualified or range answers	Respondent answers "About 20" (qualified) or "10–20" (range)	H3a: Rates of qualified or range answers will be lower.	H3b: Rates of qualified or range answers will be lower with an ItE for accessible information than other cognitive states.	H3c: Rates of qualified or range answers will be lower for sensitive questions with an ItE than for sensitive questions without an ItE.
Heaping	Respondent answers with a rounded number like 20	H4a: Rates of heaped or rounded answers will be higher.	H4b: Rates of heaped or rounded answers will be higher with an ItE for accessible information than other cognitive states.	H4c: Rates of heaped or rounded answers will be higher for sensitive questions with an ItE than for sensitive questions without an ItE.
Survey estimates (means)	–	H5: Responses will be rounded up, yielding higher means.	Not testable in this study	Not testable in this study
Question administration length	–	H6: Fewer conversational turns, and response time will be lower.	Not testable in this study	Not testable in this study

ItE: Invitation to Estimate

We are aware of only one study that investigated using an invitation to estimate: Couper et al. (2011) found that adding this phrase had no effect on the rate of ill-formed answers (i.e., answers that do not conform to the response task like nonsubstantive and qualified/range answers) or response times on three open-ended numeric questions in a web survey. Evaluating more varied question types in different modes may yield different results. Further, nonsubstantive and qualified/range answers have not been evaluated separately despite having different effects on measurement (e.g., nonsubstantive answers are often set to item missing, while qualified/range answers are processed into a final response). Finally, the assumption that an invitation to estimate reduces item nonresponse rates remains uninvestigated.

2.2.1. Invitation to Estimate, Cognitive States, and Question Sensitivity

The effect of inviting estimation may also differ across cognitive states. For example, it may have no effect on requests for information that is available (because enumeration is not burdensome in this case), generatable (because this information could only be estimated in the first place), or inestimable (because this information cannot be retrieved). However, an invitation to estimate may lead respondents to approximate their answers to questions asking about accessible information by giving them permission to skip the cognitively taxing enumeration process and estimate their answer instead. This may lead to the outcomes that we hypothesize accompany a successful invitation to estimate (i.e., H1b-H4b).

The effect of an invitation to estimate may also vary with question sensitivity. An invitation to estimate may change a respondent's adequacy judgement for sensitive questions. In particular, permission to estimate may make respondents more comfortable answering the question by providing an inexact (e.g., rounded/heaped), less sensitive answer. For example, a respondent may be unwilling to report that they have received exactly nine speeding tickets in the past year. Including an invitation to estimate may encourage this respondent to instead provide an answer of five tickets: a rounded (i.e., plausibly estimated), less sensitive answer. Therefore, we hypothesize that adding an invitation to estimate on sensitive questions will lead to less item nonresponse by changing respondents' adequacy judgement to allow for estimated, less sensitive answers (H1c). Accordingly, we also hypothesize that this increase in estimated responses for sensitive questions will lead to the outcomes that we anticipate accompany a successful invitation to estimate (H2c-H4c).

2.2.2. Invitation to Estimate and Data Collection Mode

It is unclear if the effects of an invitation to estimate vary across self-administered and interviewer-administered modes. In self-administered surveys, respondents can see the text of a question and refer to it when considering their response. In a telephone survey, respondents must hold the question in their working memory while also considering their answer. This makes aural telephone surveys more cognitively taxing than visual modes (De Leeuw 2005) and suggests that an invitation to estimate might prove especially useful at reducing item nonresponse in the telephone mode. On the other hand, because telephone interviewers serve as motivating agents, item nonresponse rates are generally already low

in this mode (De Leeuw 2005); there may be little room for further reductions in item nonresponse rates due to an invitation to estimate.

Interviewers also interact with respondents to resolve nonsubstantive and qualified/range answers, for example, by probing these types of responses to obtain a single integer value. Because interviewers do this work during the question-asking process, there may be no detectable effect of an invitation to estimate on rates of nonsubstantive and qualified/range answers in the final recorded responses in telephone surveys (although we still expect increased heaping and larger estimated means). However, an invitation to estimate may decrease instances where respondents give nonsubstantive and qualified/range answers at any point during the question/answer conversation and reduce the need for interviewer intervention. Thus, we expect to see lower rates of respondents ever giving nonsubstantive or qualified/range answers when estimation is explicitly allowed.

Finally, unlike the mail mode, telephone surveys give researchers access to an additional indicator of respondent burden: question administration length. Question administrations lasting longer than a paradigmatic "question asked/question answered" sequence (i.e., around two conversational turns) can indicate that a question is burdensome (Schaeffer and Maynard 1996). Longer response times are also a common sign that survey questions are difficult to answer (Bassili and Scott 1996; Draisma and Dijkstra 2004). If an invitation to estimate reduces burden in the telephone mode, we would expect to see (H6): (1) fewer conversational turns between the interviewer and respondent and (2) shorter response times when estimation is allowed.

In this article, we evaluate the potential benefits and drawbacks of adding an invitation to estimate to open-ended numeric questions. For 22 questions in a mail survey we compare: rates of item nonresponse, nonsubstantive answers, qualified/range answers, and heaping; as well as estimated means across questions with and without an invitation to estimate. In the mail mode, we also explore the effect of inviting estimation across two question characteristics: cognitive state and sensitivity. For three questions in a telephone survey, we compare: rates of item nonresponse, ever giving nonsubstantive answers, ever giving qualified/range answers, and heaping; as well as estimated means, number of conversational turns, and response times across questions with and without an invitation to estimate.

3. Data and Methods

The data for the mail study comes from the National Health, Wellbeing and Perspectives Study (NHWPS) survey. NHWPS was conducted by the University of Nebraska-Lincoln's Bureau of Sociological Research (BOSR) in the spring of 2015. A total of 1,002 respondents completed and returned the survey (AAPOR RR1 = 16.7%). Respondents were randomly selected using the next birthday within-household selection method. The 12-page NHWPS questionnaire contained 77 questions asking about health, mental health, well-being, victimization, current events, and demographics. Sampled households were randomly assigned to one of two versions of the questionnaire (Version 1: $n = 522$, AAPOR RR1 = 17.4%; Version 2: $n = 480$, AAPOR RR1 = 16.0%). In Version 2, an invitation to estimate (i.e., "Your best estimate is fine.") was appended to 22 open-ended

numeric questions asking about the number of times that particular events occurred in their lifetime, number of hours spent on certain behaviors during a typical week, and an income question (see Online supplemental material A for question wording). This invitation was not included in Version 1.

The data for the telephone study come from the Work and Leisure Today 2 (WLT2) survey, a dual-frame random-digit dial telephone survey of U.S. adults. WLT2 was conducted by Abt SRBI in the summer of 2015. For landline numbers, survey respondents were randomly selected among adult household members using the Rizzo et al. (2004) within-household selection method (using the next birthday method for households with 3+ adults). For cell phone numbers, the adult who answered the phone was interviewed. The survey had 902 respondents (AAPOR RR3 = 7.8%), and contained 58 questions asking about leisure time, use of technology, and demographics. We again randomly assigned sampled cases to receive the invitation to estimate (Version 1: n = 451, AAPOR RR3 = 7.4%) for three behavioral questions, or not (Version 2: n = 451, AAPOR RR3 = 8.4%) (see Online supplemental material A for question wording). Three Version 1 cases were removed from analysis due to poor call quality (final n = 899).

3.1. Dependent Variables

3.1.1. Mail Survey

Our first dependent variable in the mail survey is an indicator for item nonresponse coded 1 if a respondent did not provide an answer to a question, and 0 otherwise. Next, we set an indicator for nonsubstantive answers to 1 if a respondent provided a non-numeric answer that could *not* be recoded as an integer (e.g., written notes like "too many to count" or "do the math") and 0 otherwise. Item nonrespondents are excluded from this indicator.

We operationalize qualified/range answers using an indicator variable coded as 1 for non-integer answers that *could* be processed into an integer based on a set of rules (e.g., range answers like "20 to 30" could be coded to 20, the lower limit of the range; qualified answers like "About 20" could be coded to 20; questions like "two hours a day per week" could be coded to 14; decimal answers like 25.2 could be rounded to 25; answers with units like "18 hours" could be coded as 18) and 0 otherwise. We make one exception to these rules for written, negative answers like "no" or "none". For the 21 event/behavior questions in the mail survey, we treat negative answers as a final answer of "0," and do not code them as qualified/range answers. However, these answers may have a different meaning in the context of an income question. Therefore, we code these answers as nonsubstantive for our income question, because negative responses like "no" may indicate a refusal to provide income information rather than a final answer of "0." This indicator of qualified/range answers excludes respondents previously coded as providing either item nonresponse or nonsubstantive answers.

Our main indicator of heaping was coded as a 1 if the response was a multiple of 5, and 0 if it was not. As respondents can also heap answers based on the calendar time in a question's reference period (Huttenlocher et al. 1990), for the five questions with a reference period of "in the last week," we also create a heaping indicator coded as a 1 if the response was a multiple of 7 (i.e., number of days in a week), and 0 if it was not. Both heaping measures exclude cases coded as a 1 for any previous indicator.

Our final dependent variable is the substantive responses to each question (including ranges/qualified answers). To account for outliers, we calculated the 99th percentile for each question, and replaced answers above the 99th percentile with that question's 99th percentile value.

3.1.2. Telephone Survey

In the telephone survey, our indicator of item nonresponse is coded as 1 if a respondent's final answer was "don't know" or a "refusal." All other answers are coded as 0.

Since instances of nonsubstantive and qualified/range answers are often resolved by telephone interviewers and not reflected in final answers, we use behavior coding to identify if these answers ever occurred during the question administration. Behavior coding is a systematic, objective method for identifying deviations from a paradigmatic "question asked/question answered" interviewer/respondent interaction (Fowler and Cannell 1996; Schaeffer and Maynard 1996). For each interview, we transcribe administrations of the three telephone questions at the conversational turn level (i.e., a period of uninterrupted speech by an interviewer or respondent that ends when an actor stops speaking or is interrupted by another actor). Trained undergraduates behavior-coded these turns using Sequence Viewer (Dijkstra 1999). To assess inter-coder reliability, two master coders also coded a 10% random subsample of the transcripts. Kappa values ranged from 0.54 to 0.81, all above the common 0.40 cutoff (Bilgen and Belli 2010).

We create an indicator for whether the respondent *ever* gave a nonsubstantive answer for each question using the behavior codes. For each question administration, we code the indicator as 1 if a respondent *ever* gave a non-integer answer to the question that could not be recoded (e.g., saying "I don't know," "a whole lot," or refusing to answer), regardless of whether they were an item nonrespondent or gave a final answer. All other answers were coded as 0. We similarly use behavior coding to create an indicator for qualified/range answers for each question coded as 1 if a respondent *ever* gave a non-integer answer to the question that could be recoded and 0 otherwise.

For the remaining indicators we exclude all cases coded as 1 for item nonresponse and focus only on those who gave a response. We create two indicators for heaping using final answers: one is coded 1 if a response is a multiple of 5 (to capture common rounding behavior) and 0 otherwise, and the other is coded 1 if a response is a multiple of 7 (because these questions use "week" as a reference period) and 0 otherwise. Means are again examined using the substantive integer responses to each question, accounting for outliers using the same 99th percentile method as the mail survey.

We operationalize the question administration length (i.e., an indicator of the question's burden) in two ways. First, the total number of conversational turns for each question is used as an indicator of administration burden overall. Second, we calculate the number of seconds it takes to reach a final answer after the interviewer has finished reading a question. We do this by summing the length of each conversational turn that occurs after the interviewer's first question-asking turn. To account for skew in this response time measure, we truncate response times below the first percentile and above the 99th percentile for each question (Yan and Olson 2013), and then use a natural log transformation.

3.2. Independent Variables

Our focal independent variable in both surveys is an indicator variable of whether a respondent was randomly assigned to receive an invitation to estimate their answer to a question (= 1) or not (= 0). The randomization in both modes was at the respondent level – for each respondent, all of the numeric questions examined here either included the invitation to estimate or did not include the invitation to estimate.

Our next set of independent variables capture key question characteristics: the likely cognitive state of information for the average person requested by the question and question sensitivity. These question characteristics (as well as those listed in the Controls section) were independently coded by two coders (two of the authors), and were coded relative to their perceptions of how the average person would view these questions (Online supplemental material (B) displays the questions used to rate these characteristics), as one might do when designing a questionnaire. Kappa values for all question characteristics were above 0.70 except for cognitive state and sensitivity. Disagreements between the two coders were resolved by a third coder (a third author) to create the final set of codes used in the analyses (Summarized in Table 2).

In the mail mode only, we use these codes to create three dichotomous indicators for the cognitive state of information requested by the question (available, accessible, and generatable) for the average person. No questions were coded as inestimable. We also create an indicator that describes the sensitivity (sensitive = 1; not sensitive = 0) of the question. In the telephone mode, because there are only three items included in this experiment, we use these coded characteristics for interpretation only (i.e., not as independent variables).

Table 2. Summary of question characteristics.

	NHWPS – mail		WLT2 – telephone	
	n	Percent/mean	n	Percent/mean
Cognitive state				
Available	8	38.10%	2	66.66%
Accessible	4	19.05%	0	0.00%
Generatable	9	42.86%	1	33.33%
Sensitivity				
Not sensitive	12	57.14%	1	33.33%
Sensitive	9	42.86%	2	66.66%
Reference period/similarity				
Short/similar	5	23.81%	3	100.00%
Long/dissimilar	16	76.19%	0	0.00%
Frequency				
Low	17	80.95%	1	33.33%
High	4	19.05%	2	66.66%
Regularity				
Regular	6	28.57%	1	33.33%
Irregular	15	71.43%	2	66.66%
Reading level		8.76		5.03
Number of words in stem		13.19		14.33

3.3. Controls

Question characteristics associated with a respondent's choice of an enumeration or estimation response strategy may affect many of the data quality indicators in this study as well as may influence the average respondent's cognitive state on a question (e.g., Blair and Burton 1987; Burton and Blair 1991; Conrad et al. 1998). Therefore, in the mail mode, we code and control for each question's reference period length (short = 0; long = 1); frequency (low = 0; high = 1), regularity (regular = 0; irregular = 1), and similarity (similar = 0; dissimilar = 1) of episodes of the event asked about; and number of words in the question stem and reading level using the Flesch-Kincaid grade level (both calculated excluding the phrase "Your best estimate is fine," then grand-mean-centered). Each question's reference period and similarity rating were perfectly collinear, so we analyze these two characteristics together (i.e., short/similar = 0; long/dissimilar = 1).

Likewise, decreased cognitive ability may negatively affect a respondent's ability to retrieve information from memory and thus also affect some of the data quality indicators examined here (i.e., skipping a question or estimating instead of enumerating; Krosnick 1991; Knäuper et al. 1997). In both the mail and telephone surveys, we control for respondents' age and level of education (two common proxies for cognitive ability) (Table 3). We multiply imputed missing values for age (12%) and education (6%) in the mail survey; due to the low item nonresponse rates in the telephone survey, we used grand mean imputation (for age, 4%, n = 34) and modal category imputation (for education, 1%, n = 5) in the telephone survey. The models include a grand-mean-centered continuous measure of age; level of education is represented by three dichotomous variables (high school or less, some college, or college graduate or higher).

Table 3. Summary of unimputed respondent characteristics.

	Overall	Without invitation to estimate	With invitation to estimate	t-test/ chi-square
NHWPS – mail				
Number of respondents	1,002	522	480	
Age (in years)	57.35	57.47	57.22	0.218
Education				
High school graduate or less	20.58%	20.25%	20.94%	2.797
Some college	31.98%	29.86%	34.30%	
College graduate or more	47.44%	49.90%	44.37%	
WLT2 – telephone				
Number of respondents	899	451	448	
Age (in years)	54.13	53.73	54.54	0.644
Education				
High school graduate or less	31.32%	31.03%	31.61%	6.022
Some college	26.29%	29.69%	22.87%	
College graduate or more	42.39%	39.29%	45.52%	

Note: There were no significant differences in respondent age or education across questionnaire version in either survey

3.4. Analysis

3.4.1. Mail Survey

We analyze our event/behavior questions (n = 21) together and our income question (n = 1) by itself. For the event/behavior questions, because the invitation to estimate was assigned at the respondent level (i.e., no variation within respondents on the experimental condition; for each respondent, all questions included the invitation to estimate or did not), we analyze each of our dichotomous data quality indicators using a population-averaged model with an exchangeable correlation structure (`xtgee` command in Stata 14.2 with respondents set as the clustering variable) (Agresti 2002; McNeish et al. 2017; Raudenbush and Bryk 2002; Rabe-Hesketh and Skrondal 2012; West et al. 2015). We use a logit link function because we have dichotomous outcomes. Clustering of questions within respondents is accounted for using cluster-robust standard errors, estimated using Huber-White sandwich estimators; the multivariate analysis of the mail survey also accounts for multiple imputation of the age and education variables. Age and education were multiply imputed $D = 5$ times (using adjustment cell random hotdeck imputation via `hotdeckvar` in Stata; Schonlau 2018) and combined for analysis using Rubin's Rules (via Stata's `micombine`).

We start our analyses by exploring the bivariate relationship between the invitation to estimate and each of our dichotomous data quality indicators (i.e., item nonresponse, nonsubstantive answers, qualified/range answers, and heaped answers). We estimate four population-averaged models, each predicting a different dichotomous quality indicator with our invitation to estimate indicator.

We then move to a multivariate framework, adding our remaining independent variables (each question's cognitive state and sensitivity) and controls (reference period/similarity, frequency, regularity, number of words, reading level, respondent age and education) to each of the bivariate models mentioned above. To test our moderation hypotheses, we estimate two new models per data quality indicator; each model includes an interaction term between the invitation to estimate indicator and a different question characteristic (i.e., an interaction with cognitive state in one model, an interaction with sensitivity in another). To interpret significant interactions, we calculate predicted probabilities using the `margins` command in Stata, holding all other variables at their means. Results from our main effects models and models with significant interactions are displayed in-text; results for models with non-significant interactions are displayed in the Online supplemental material

Our final analyses examine unweighted means for each behavior question across our two experimental conditions. For the bivariate tests, we estimate 21 ordinary least squares (OLS) regression models (one per question) predicting each question's average response with the invitation to estimate indicator entered as a predictor. We also tested this relationship using count models (e.g., Poisson and negative binomial regression), and our substantive conclusions do not change. We therefore describe the OLS results for ease of interpretation. For the multivariate models, we add respondent age and education as controls. For these analyses, one model is estimated per question so question characteristics (e.g., cognitive state, sensitivity) are not included as predictors.

For the income question, we conduct bivariate analyses using regression models predicting each of our data quality indicators (logistic regression for dichotomous outcomes, OLS regression for continuous outcomes) with the invitation to estimate as a predictor. We confirm our bivariate findings using multivariate models with our two respondent controls, age and education, as predictors.

3.4.2. Telephone

As the telephone portion of our study contains only three questions, we do not have enough observations within respondents to estimate the population-averaged models. Thus, we examine the data quality indicators separately for each question. Additionally, each respondent is nested within an interviewer, yielding a multilevel data structure (Hox, 1994; Olson and Bilgen, 2011; Olson and Peytchev, 2007). All telephone analyses account for this clustering using Stata's complex survey design procedures (`svy` procedures).

For our bivariate analyses, we estimate design-adjusted regression models predicting each of our data quality indicators (logistic regression for dichotomous outcomes, OLS regression for continuous outcomes) for each question with the invitation to estimate as a predictor. For the multivariate models, we add our two respondent controls, age and education, as predictors.

4. Results

4.1. Mail

Table 4 shows results of the bivariate tests of the relationship between an invitation to estimate and our dichotomous data quality indicators for our 21 event/behavior questions. Rates of item nonresponse, qualified/range answers, and heaped answers (multiples of 5 or 7) did not significantly differ across questions with and without an invitation to estimate ($p > 0.05$). However, question administrations with an invitation to estimate had significantly fewer nonsubstantive answers (0.78%) than those without the invitation (1.58%), although the difference is less than a percentage point (diff = 0.80%; $z = -2.10$, $p < 0.05$).

In the multivariate models, item nonresponse rates are again not significantly different across question administrations with and without an invitation to estimate ($p = 0.39$; Table 5). This indicates that, despite conventional wisdom and our H1a, inviting respondents to estimate does not reduce item nonresponse rates. Further, the interactions between the invitation to estimate indicator and our two independent variables (i.e., cognitive state and question sensitivity) were not significant ($p > 0.05$), indicating that the (non)effect of an invitation to estimate on item nonresponse did not differ across these question characteristics (i.e., hypotheses H1b and H1c were not supported). Table 6 summarizes the test statistics and p-values for each interaction tested in this study. Full results from models with non-significant interactions are displayed in Online supplemental material (Table C).

Consistent with H2a, an invitation to estimate did significantly reduce rates of nonsubstantive answers (e.g., "a lot" or "too many to count") (OR = 0.555, $p < 0.05$; Table 5), though these behaviors were rare overall. Predicted probabilities of nonsubstantive answers reduced by almost half from 0.014 without an invitation to 0.008 with an invitation to estimate. Explicitly accepting estimation, therefore, may communicate to respondents that a guess is preferred to admitting that the exact answer is not known.

Table 4. Percent of question administrations with each data quality indicator overall and by questionnaire version for mail.

Indicator	Overall		Without ItE		With ItE		Difference	
	%	SE	%	SE	%	SE	z/t-value	With ItE – Without ItE
Item nonresponse	5.68%	0.16%	6.19%	0.23%	5.14%	0.22%	-1.03	-1.05%
Nonsubstantive answers	1.20%	0.08%	1.58%	0.12%	0.78%	0.09%	-2.10	-0.80%*
Range or qualified answers	2.66%	0.11%	2.63%	0.16%	2.70%	0.17%	0.16	0.07%
Heaped/rounded answers								
Multiples of 5	60.74%	0.35%	60.36%	0.49%	61.15%	0.51%	0.81	0.80%
Multiples of 7								
Q51A: Working	44.78%	1.69%	44.52%	2.35%	45.06%	2.45%	0.16	0.54%
Q51B: Household work	15.05%	1.21%	15.89%	1.72%	14.15%	1.69%	-0.72	-1.74%
Q51C: Looking after family	56.11%	1.71%	57.24%	2.37%	54.90%	2.47%	-0.68	-2.34%
Q51D: Leisure activities	16.84%	1.27%	16.11%	1.74%	17.62%	1.86%	0.59	1.51%
Q51E: Sleeping	43.45%	1.69%	43.78%	2.34%	43.10%	2.44%	-0.20	-0.68%

SE: standard error; ItE: Invitation to Estimate

*$p < 0.05$

Table 5. Odds ratios and standard errors for models predicting data quality indicators for mail.

	Item nonresponse		Nonsubstantive answers		Nonsubstantive answers ItE x sensitivity		Qualified/range answers		Heaped/rounded answers (multiples of 5)	
	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)
Questionnaire version (ref = Without ItE)										
With ItE	0.813	(0.197)	0.555*	(0.153)	0.664	(0.185)	0.992	(0.186)	1.016	(0.076)
Question characteristics										
Cognitive State (ref = Available)										
Accessible	1.239***	(0.071)	1.874***	(0.207)	1.876***	(0.208)	2.030***	(0.269)	0.505***	(0.022)
Generatable	0.988	(0.063)	1.825***	(0.287)	1.832***	(0.290)	1.775***	(0.283)	0.377***	(0.019)
Sensitivity (ref = not sensitive)										
Sensitive	1.474***	(0.097)	0.851	(0.113)	1.068	(0.171)	0.962	(0.104)	2.363***	(0.143)
Controls										
Frequency (ref = Low)										
High	0.737***	(0.042)	2.289**	(0.631)	2.300**	(0.638)	1.131	(0.129)	1.306***	(0.079)
Reference Period/similarity (ref = short/similar)										
Long/dissimilar	0.316***	(0.062)	1.427	(0.365)	1.428	(0.367)	1.574**	(0.253)	1.202*	(0.096)
Regularity (ref = regular)										
Irregular	1.112	(0.164)	2.670**	(0.868)	2.702**	(0.880)	0.512***	(0.092)	0.422***	(0.043)

Table 5. Continued

	Item nonresponse		Nonsubstantive answers		Nonsubstantive answers ItE x sensitivity		Qualified/range answers		Heaped/rounded answers (multiples of 5)	
	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)	OR	(SE)
Question stem Reading level (GMC)	1.020*	(0.009)	1.043**	(0.016)	1.043*	(0.017)	1.054**	(0.016)	1.110***	(0.007)
Number of words in question (GMC)	0.986*	(0.006)	0.948***	(0.012)	0.948***	(0.012)	0.963***	(0.010)	0.997	(0.003)
Age (GMC)	1.033***	(0.011)	1.018	(0.012)	1.018	(0.012)	1.000	(0.006)	1.005**	(0.001)
Education (ref = HS or less)										
Some college	1.077	(0.536)	0.732	(0.297)	0.678	(0.269)	0.698	(0.190)	0.889	(0.066)
College +	0.860	(0.361)	0.324**	(0.141)	0.320**	(0.136)	0.688	(0.170)	1.037	(0.069)
Interactions										
ItE x question Sensitivity					0.468*	(0.140)				
N	21042		19846		19846		19608		19086	

OR: Odds Ratio; SE: standard error; ItE: Invitation to Estimate; GMC: grand-mean-centered
* $p < 0.05$; ** $p < 0.01$; ***$p < 0.001$

Table 6. Summary of test statistics and significance values for interactions by data quality indicator for mail.

Data quality indicator	χ^2	p
Item nonresponse		
ItE x Cognitive state	0.65	0.72
ItE x Sensitivity	1.96	0.16
Nonsubstantive answers		
ItE x Cognitive state	0.28	0.87
ItE x Sensitivity	6.44	0.01
Qualified/range answers		
ItE x Cognitive state	0.50	0.78
ItE x Sensitivity	0.49	0.48
Heaped/rounded answers (multiples of 5)		
ItE x Cognitive state	0.35	0.84
ItE x Sensitivity	0.05	0.83

ItE: Invitation to estimate

The effect of inviting estimation on nonsubstantive answers also differed across question sensitivity ($\chi^2 = 6.44$, $p < 0.05$; Table 6; supporting H2c), though the rates of occurrence remain low. For nonsensitive questions, there is no significant difference in the predicted probabilities of nonsubstantive answers between the version with the invitation to estimate (0.010) and the version without (0.014; Figure 1). However, for sensitive questions, the probability of nonsubstantive answers is significantly lower in the version with the invitation to estimate (0.005) than in the version without (0.015). This finding suggests that the invitation to estimate is encouraging some respondents to sensitive questions to provide an edited, usable answer rather than indicating that the question is sensitive by writing a nonsubstantive answer (e.g., "too revealing"). Contrary to H2b, the

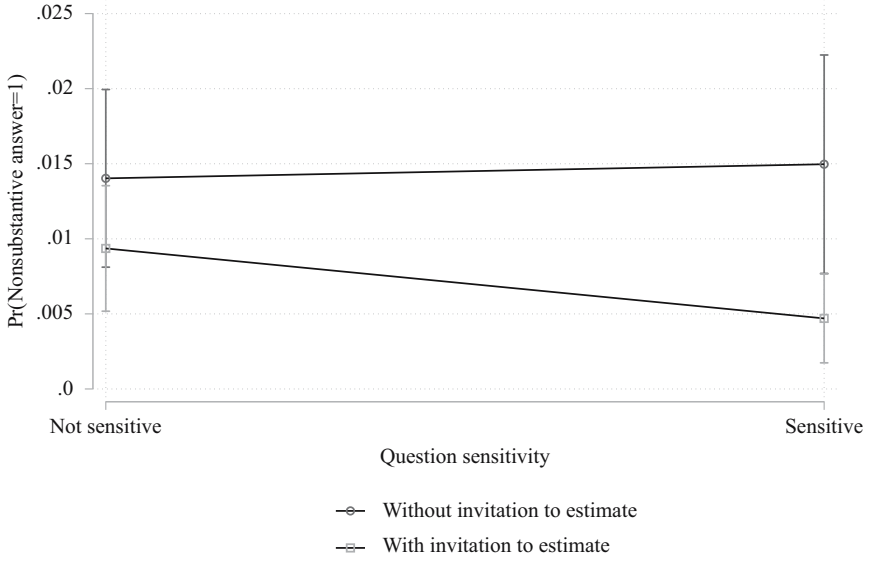

Fig. 1. Predicted probabilities for nonsubstantive answers by questionnaire version and question sensitivity.

interaction between the invitation to estimate indicator and cognitive state was not significant ($p = 0.87$; Table 6; full results in Online supplemental material, Table C).

Inviting estimation did not significantly affect the rate of qualified/range answers in the multivariate models ($p = 0.97$; Table 5). The invitation, therefore, did not help clarify the level of precision requested by the questions in the mail survey. Neither of the interactions between the invitation to estimate indicator and either cognitive state or question sensitivity were significant ($p > 0.05$; Table 6; full results in Online supplemental material, Table D). Thus, none of our H3a-H3c were supported.

Our next set of multivariate models demonstrates that respondents heaped their answers around multiples of 5 slightly more with an invitation to estimate, but the difference is not significant ($p = 0.36$; Table 5). The effect of an invitation to estimate on heaping does not differ across cognitive state or sensitivity ($p > 0.05$; Table 6; full results in Online supplemental material, Table D). Rates of heaping at multiples of 7 also do not significantly differ with or without an invitation to estimate in the multivariate models for the five questions with a reference period of a "week" ($p > 0.05$; Online supplemental material, Table E). Overall, we find no support for H4a-H4c.

Next, we examine substantive responses to the NHWPS questions. Table 7 displays means overall and by questionnaire version for each item. As expected, most event/behavior questions (76%) did have higher means with an invitation to estimate, but differences were generally small and none of the 21 bivariate tests were statistically significant ($p > 0.05$). These results remain unchanged when controlling for respondent age and education in the multivariate models (Online supplemental material, Tables F–H). Therefore, contrary to H5, any tendencies towards over-reporting when estimation was allowed were not large enough to cause significant differences in estimated means for these questions.

For our final set of analyses on the mail survey, we examine the effect of an invitation to estimate on an income question. We interpret and discuss our findings using the bivariate tests because the multivariate and bivariate results are the same for all indicators (multivariate results presented in Online supplemental material, Table I). We find that adding an invitation to estimate to an income question has no effect on rates of item nonresponse, nonsubstantive answers, qualified/range answers, or heaping ($p > 0.05$; Table 8). Further, we find that there are no significant differences in estimated means when an invitation to estimate is added versus when it is not ($p > 0.05$). This collection of results demonstrates that inviting respondents to estimate on an income question does little to affect data quality (i.e., no support for H1a-H4a or H5). It is possible that respondents assume that answers to yearly income questions should be estimated, as exact answers are typically not known; explicitly inviting estimation may not change response behaviors for this question.

4.2. Telephone

In the telephone survey, we explore the effect of an invitation to estimate across three behavioral questions asking about alcohol consumption and cigarette smoking in the past seven days, and number of miles driven in a typical week. We interpret and discuss our findings using only the design-adjusted bivariate tests because the multivariate and bivariate results are identical for all indicators across all questions, with one noted

Table 7. Item means and standard errors overall and by questionnaire version for mail and telephone.

Question	Overall		Without ItE		With ItE		Diff	
	Mean	SE	Mean	SE	Mean	SE	t-value	With ItE – Without ItE
NHWPS – Mail								
Q26A: New job	2.48	0.06	2.43	0.09	2.53	0.09	0.79	0.10
Q26B: Vacation	15.85	0.64	15.69	0.87	16.01	0.95	0.25	0.32
Q26C: Illness or accident	0.56	0.03	0.54	0.04	0.58	0.05	0.67	0.04
Q26D: Victim of crime	0.87	0.04	0.89	0.06	0.84	0.06	-0.61	-0.05
Q26E: Threatened	0.87	0.09	0.83	0.11	0.92	0.13	0.53	0.09
Q26F: Injured by someone else	0.25	0.03	0.23	0.04	0.27	0.05	0.73	0.04
Q26G: Unwanted sexual comments	1.95	0.20	1.92	0.29	1.97	0.28	0.12	0.05
Q26H: Bonus or promotion	4.92	0.22	5.10	0.31	4.73	0.32	-0.84	-0.38
Q27A: Witnessed other, illness or accident	7.46	0.71	6.97	0.91	7.98	1.11	0.71	1.02
Q27B: Witnessed other, victim of crime	1.37	0.08	1.31	0.11	1.44	0.12	0.80	0.13
Q27C: Witnessed other, threatened	1.24	0.09	1.28	0.12	1.19	0.14	-0.46	-0.09
Q27D: Witnessed other, injured by someone else	1.07	0.11	1.00	0.13	1.15	0.17	0.69	0.15
Q27E: Witnessed other, unwanted sexual comments	2.15	0.22	1.94	0.27	2.37	0.35	0.97	0.43
Q28: Told by other, victim of crime	4.57	0.25	4.34	0.34	4.83	0.36	1.00	0.49
Q29: Head injury w/o loss of consciousness	0.95	0.05	0.92	0.07	0.98	0.08	0.62	0.07
Q30: Head injury w/ loss of consciousness	0.41	0.03	0.40	0.04	0.41	0.04	0.20	0.01
Q51A: Working	23.44	0.74	23.87	1.05	22.97	1.06	-0.61	-0.90
Q51B: Household work	10.52	0.35	10.45	0.49	10.60	0.50	0.23	0.16
Q51C: Looking after family	12.23	0.96	12.24	1.33	12.21	1.39	-0.01	-0.03
Q51D: Leisure activities	19.00	0.57	18.98	0.79	19.02	0.81	0.04	0.04
Q51E: Sleeping	41.85	0.61	41.56	0.86	42.18	0.88	0.51	0.62
WLT2 – Telephone								
Q22: Alcohol	2.89	0.24	3.09	0.37	2.69	0.31	-0.96	-0.40
Q23: Cigarettes	12.72	1.18	13.23	1.69	12.21	1.66	-0.42	-1.02
Q31: Miles driven	172.23	7.76	173.18	10.93	171.29	11.04	-0.15	-1.89

SE: standard error; ItE: Invitation to Estimate

Note: There were no significant differences in means across questions with and without an invitation to estimate.

Table 8. Percent of question administrations with each data quality indicator and means overall and by questionnaire version for income question in mail.

Indicator	Overall		Without ItE		With ItE		Difference	
	%	SE	%	SE	%	SE	z/t-value	With ItE – Without ItE
Item nonresponse	17.76%	1.21%	19.16%	1.72%	16.25%	1.69%	-1.07	-2.91%
Nonsubstantive answers	4.98%	0.76%	5.69%	1.13%	4.23%	1.00%	-0.96	-1.46%
Range and qualified answers	5.24%	0.80%	5.53%	1.15%	4.94%	1.11%	-0.29	-0.59%
Heaped/rounded answers (Multiples of 5)	96.23%	0.70%	95.21%	1.10%	97.27%	0.85%	1.47	2.05%
	Mean	SE	Mean	SE	Mean	SE		
Q62: Income	80,045.04	2,486.35	80,519.19	3,644.50	79,554.88	3,377.39	-0.19	-964.31

SE: standard error; ItE: Invitation to Estimate

exception. For completeness, results from the multivariate models are presented in Online supplemental material (Tables J–Q).

Like the mail survey (i.e., contrary to H1a), we find that the relationship between item nonresponse and an invitation to estimate is not significant for any of the three questions in the telephone survey ($p > .05$; Table 9). As expected, interviewers likely motivated respondents to provide an answer across both questionnaire versions, nullifying the effect of an invitation to estimate on item nonresponse.

The percentage of respondents who ever gave a nonsubstantive answer (e.g., "I smoke a lot of cigarettes") was lower for questions with an invitation to estimate in the telephone survey, but this difference was only significant for the question asking about cigarettes ($t = -2.12$, $p < .05$; Table 9). However, this difference becomes not significant when controlling for respondent age and education ($p = .07$). Inviting estimation also did not significantly change the percentage of cases with at least one qualified/range answer (e.g., "between 5 and 10 drinks a week") for questions asking about alcohol or cigarettes ($p = 0.28$; Table 9).

When asking about number of miles driven in a typical week, cases with qualified/range answers were, however, significantly lower when the question included an invitation to estimate (47.33%) compared to when it did not (59.74%; diff = 12.41%; $t = -3.70$, $p < .01$; Table 9; partial support for H3a). This may be because the number of miles driven in a typical week is, by far, the highest-frequency behavior in either the mail or telephone survey (overall mean = 172.23 miles; Table 7) and cannot be reported exactly by most respondents (i.e., generatable information). Without an invitation to estimate, respondents may feel the need to notify the interviewer that they are uncertain of their answer. Including an invitation to estimate, however, may communicate that an imprecise number is acceptable, thus reducing respondent expressions of uncertainty.

Contrary to H4a, an invitation to estimate also does not significantly affect the percentage of respondents that heap their answers around multiples of 5 or 7 ($p > 0.05$; Table 9). For the alcohol and cigarette questions, this may be because these items are sensitive, making respondents more likely to partially conceal their answers via heaping with or without an invitation to estimate. Additionally, since the number of miles driven in a typical week generally cannot be enumerated for many respondents, answers to this question are likely heaped with or without permission to estimate.

As in the mail survey, we find that means to the three telephone questions do not differ with or without an invitation to estimate, contrary to H5. These results again indicate that an invitation to estimate does not trigger any meaningful changes in the magnitude of respondents' answers.

Finally, we examine two related indicators of question administration length that are unique to the telephone mode. We find that adding an invitation to estimate significantly reduces the number of conversational turns required to reach a final answer by about one turn for all three questions ($p < .05$; Table 9). Eliminating this extra turn significantly reduces response time by about one second for the question asking about alcohol ($t = -2.41$, $p < .05$), and by about two seconds for the questions asking about cigarettes ($t = -2.89$, $p < .01$) and miles driven ($t = -5.84$, $p < .001$). While an invitation to estimate may not affect the final answers, it does seem to simplify the interaction between interviewers and respondents required to achieve those final answers, supporting H6.

Table 9. Percent of question administrations with each data quality indicator overall and by questionnaire version for telephone.

	Overall		Without ItE		With ItE		Diff	
	%	SE	%	SE	%	SE	z/t-value	With ItE – Without ItE
Item Nonresponse								
Q22: Alcohol	0.89%	0.42%	0.89%	0.47%	0.89%	0.69%	0.010	0.01%
Q23: Cigarettes	0.56%	0.23%	0.67%	0.33%	0.45%	0.32%	-0.450	-0.22%
Q31: Miles driven	2.96%	0.87%	3.38%	1.48%	2.56%	0.89%	-0.500	-0.82%
Ever gave nonsubstantive answer								
Q22: Alcohol	6.90%	0.80%	7.78%	1.19%	6.01%	1.07%	-1.10	-1.76%
Q23: Cigarettes	4.49%	0.78%	6.08%	1.02%	2.91%	0.90%	-2.12*	-3.17%
Q31: Miles driven	13.24%	1.53%	15.32%	2.01%	11.20%	2.05%	-1.40	-4.13%
Ever gave qualified/range answer								
Q22: Alcohol	16.35%	1.20%	15.78%	1.35%	16.93%	1.97%	0.48	1.15%
Q23: Cigarettes	11.90%	1.18%	13.51%	1.48%	10.29%	1.77%	-1.34	-3.22%
Q31: Miles driven	53.47%	2.12%	59.74%	2.28%	47.33%	2.40%	-3.70**	-12.41%
Heaped final answers								
Multiples of 5								
Q22: Alcohol	66.33%	1.22%	67.11%	1.68%	65.54%	1.71%	-0.66	-1.57%
Q23: Cigarettes	97.20%	0.58%	96.65%	0.71%	97.76%	0.93%	0.86	1.11%
Q31: Miles driven	98.01%	0.47%	98.12%	0.58%	97.90%	0.73%	-0.24	-0.22%
Multiples of 7								
Q22: Alcohol	61.84%	1.38%	63.76%	1.62%	59.91%	2.04%	-1.48	-3.85%
Q23: Cigarettes	87.47%	1.20%	88.39%	1.56%	86.55%	1.69%	-0.80	-1.85%
Q31: Miles driven	9.03%	1.01%	9.14%	1.31%	8.92%	1.53%	-0.11	-0.22%
Number of turns								
Q22: Alcohol	3.69	0.16	4.04	0.22	3.33	0.14	-2.71*	-0.71
Q23: Cigarettes	3.69	0.17	4.17	0.26	3.22	0.08	-3.54**	-0.96
Q31: Miles driven	4.59	0.22	5.02	0.25	4.16	0.28	-2.27*	-0.86
Response time (secs)								
Q22: Alcohol	6.15	0.29	6.73	0.32	5.57	0.36	-2.41*	-1.16
Q23: Cigarettes	5.13	0.37	6.12	0.56	4.15	0.22	-2.89**	-1.97
Q31: Miles driven	13.01	0.39	14.04	0.41	12.00	0.54	-5.84***	-2.03

SE: standard error; ItE: Invitation to Estimate

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

5. Discussion

In this study, we evaluate the common assumption that inviting respondents to estimate when responding to numeric questions about events, behaviors, or income reduces burden. We explore the effect of an invitation to estimate on several indicators of data quality in both a mail and a telephone survey. This is the first study to our knowledge that investigates this questionnaire design choice in either mode. We have three main takeaways from our analyses.

First, we find no support for the idea that an invitation to estimate significantly reduces item nonresponse in either a mail or a telephone survey. Notably, this effect did not differ across the question characteristics in our mail study, meaning that even the most burdensome or sensitive questions did not benefit from inviting estimation. We do find that inviting estimation reduces nonsubstantive answers (i.e., answers that cannot be re-coded and would likely be set to item missing) for the event/behavior questions in our mail survey, especially when they ask about sensitive topics. However, the rate of these answers is quite small, and we do not see the same pattern for our mail income question or for any of the telephone questions. It is possible that the burden of numeric questions is not high enough to warrant item nonresponse even when respondents enumerate, making an invitation to estimate unnecessary. Alternatively, respondents may intuit that an estimated answer is preferable to item nonresponse when an exact answer is difficult or impossible to provide; they may not need encouragement to make a shift from enumeration. Future work could explore this further by, for example, investigating whether the effect of an invitation to estimate differs across respondent characteristics that may be associated with choice of retrieval strategy (e.g., education and age).

Second, if an invitation to estimate is used, researchers have little reason to worry that it will negatively affect data quality. Although the difference was not significant, questions with an invitation to estimate had slightly lower rates of item nonresponse than questions without the invitation. We also find that rates of qualified/range answers are unchanged with and without an invitation to estimate in the mail mode. Any measurement error introduced by re-coding answers like ranges into a single integer is not likely increased by inviting estimation. Further, explicitly accepting estimation does not change substantive responses: we find no effect of an invitation to estimate on rounding or means in either mode. This provides further evidence for the notion that respondents make decisions about when to use enumeration versus estimation on their own, and do not require an invitation to employ one strategy over the other.

Third, we find that an invitation to estimate may have some utility in simplifying the interaction between interviewers and respondents in CATI surveys. For example, inviting estimation reduced the percentage of respondents that gave qualified/range answers to a question asking about the number of miles driven in a typical week (i.e., a high frequency question that would be difficult to enumerate). Further, inviting estimation reduced the total number of turns required to achieve an acceptable final answer and reduced the length of response times for all three CATI questions. Although respondents ultimately provided an acceptable final answer in each of these cases, the invitation to estimate may have reduced the extra work interviewers had to perform (e.g., probing, providing clarifications) to obtain an acceptable answer. Therefore, these results are encouraging for using an invitation to estimate to reduce interactional burden in telephone surveys.

Future research should investigate whether the reduced response times found in the telephone survey can be replicated in a mail survey. Though such a study would likely be restricted to an observational, lab setting, these results would provide more insight into whether and how an invitation to estimate reduces respondent burden. It is also possible that some respondents do not actually read an invitation to estimate when it is included in visual surveys. Eye tracking studies could be used to determine if an invitation to estimate: (1) is seen by respondents, and (2) affects data quality when it is seen.

Although we experimentally examined an invitation to estimate in both a mail and a telephone survey, this was not an experimental study of mode differences. These two surveys had different questions and were fielded at different times. Instead, this study does provide a foundation for future work testing an invitation to estimate across these modes. Such a study would also provide insight into best practices for using an invitation to estimate in mixed-mode (e.g., interviewer- and self-administered) surveys. Further, the CATI survey only included three questions, which limits our ability to make strong conclusions about inviting estimation in the telephone mode. We also did not include a question asking for detailed financial information in the telephone survey. Future work should test the effects of the invitation to estimate on more questions asked in a telephone survey, including questions that vary on the theoretically motivated question characteristics identified here.

This study should also be replicated on the web; previous work in this mode (Couper et al. 2011) did not include all of the data quality indicators we explore here. We would expect that an invitation to estimate would have similar null effects across mail and web surveys, as both are self-administered and use visual channels of communication. Confirming these findings in a web survey that includes mobile respondents would allow questionnaire designers to have insights into whether one can drop the lengthy phrase "your best estimate is fine" and save space when displaying questions on the small screens of mobile devices. While this study investigated more than 20 questions, the characteristics of these questions were somewhat limited. For example, many of our questions asked about events that were unlikely to occur to a person in their lifetime (e.g., head injuries; crime victimization) or about the amount of time spent on activities that may occur with some regularity (e.g., work for pay). Additionally, the reference period and similarity of events of our questions were confounded. Also, none of our questions asked for information in the inestimable cognitive state (a good thing for our respondents, but not as useful for evaluating these questions). Future studies should explore the effect of explicitly accepting estimation on a wider variety of question types.

We also note that these question characteristics – and notably cognitive state – were not rated by the respondents themselves. For example, a respondent who has taken several out-of-state vacations in their lifetime may have more difficulty enumerating their answer than a respondent who has only taken two such vacations. It is possible, therefore, that an invitation to estimate may operate differentially based on a respondent's perception of a question and the utility of the invitation to estimate, rather than the perception of an outside rater. In general, raters disagree about question characteristics (Bais et al. 2019), although the average rating of question characteristics across expert raters has been shown to be related to measurement error (Olson 2010). Although our raters evaluated questions based on the "average respondent," asking a more diverse set of raters to evaluate question

characteristics and having a more diverse set of questions may provide more insight into the conditions under which an invitation to estimate is effective.

Overall, we find that inviting estimation in a mail survey has no significant effect for most of the data quality indicators in this study. Based on this collection of evidence, we see little reason to recommend using an invitation to estimate on questions asking about the frequency of events and behaviors or income in the mail mode, especially if questionnaire space is limited. Though we also note that using an invitation to estimate does not appear to negatively affect data quality for mail surveys either, so researchers do not be concerned about data from past studies where an invitation has been used. We are more optimistic that using an invitation to estimate in telephone surveys simplifies the interaction between interviewers and respondents. We encourage future research to replicate these findings in interviewer-administered modes.

6. References

Agresti A. 2002. *Categorical Data Analysis, Second Edition*, Hoboken, NJ: John Wiley & Sons.

Bais, F., B. Schouten, P. Lugtig, V. Toepoel, J. Arends-Tòth, S. Douhou, N. Kieruj, M. Morren, and C. Vis. 2019. "Can Survey Item Characteristics Relevant to Measurement Error Be Coded Reliably? A Case Study on 11 Dutch General Population Surveys." *Sociological Methods & Research* 48 (2): 263–95. DOI: https://doi.org/10.1177/0049124117729692.

Bassili, J.N., and B.S. Scott. 1996. "Response Latency as a Signal to Question Problems in Survey Research." *Public Opinion Quarterly* 60 (3): 390–99. https://doi.org/10.1086/297760.

Beatty, P., and D. Herrmann. 2002. "To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J. Little: 71–85. New York: Wiley.

Bilgen, I., and R.F. Belli. 2010. "Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires." *Journal of Official Statistics* 26 (3): 481–505. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/comparison-of-verbal-behaviors-between-calendar-and-standardized-conventional-questionnaires.pdf.

Blair, E., and S. Burton. 1987. "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions." *Journal of Consumer Research* 14 (2): 280–288. DOI: https://doi.org/10.1086/209112.

Burton, S., and E. Blair. 1991. "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys." *Public Opinion Quarterly* 55 (1): 50–79. DOI: https://doi.org/10.1086/269241.

Conrad, F.G., N.R. Brown, and E.R. Cashman. 1998. "Strategies for Estimating Behavioural Frequency in Survey Interviews." *Memory* 6 (4): 339–366. DOI: https://doi.org/10.1080/741942603.

Couper, M.P., C. Kennedy, F.G. Conrad, and R. Tourangeau. 2011. "Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys." *Journal of Official Statistics* 27 (1): 65–85. Available at: https://www.scb.se/contentassets/ca21efb41-

fee47d293bbee5bf7be7fb3/designing-input-fields-for-non-narrative-open-ended-responses-in-web-surveys.pdf.

De Leeuw, E.D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21 (2): 233–255. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/to-mix-or-not-to-mix-data-collection-modes-in-surveys.pdf.

Dijkstra, W. 1999. "A New Method for Studying Verbal Interactions in Survey Interviews." *Journal of Official Statistics* 15 (1): 67–85. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-new-method-for-studying-verbal-interactions-in-survey-interviews.pdf.

Dillman, D.A. 2007. *Mail and Internet Surveys: The Tailored Design Method*. Hoboken, N.J: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons.

Draisma, S., and W. Dijkstra. 2004. "Response Latency and (Para) Linguistic Expressions as Indicators of Response Error." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M.P. Couper, J. Lesser, E. Martin, J. Martin, and E. Singer. New York: Wiley.

Fowler Jr, F.J., and C.F. Cannell. 1996. "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman: 15–36. San Francisco: Jossey-Bass/Wiley.

Holbrook, A.L., S. Anand, T.P. Johnson, Y. Ik Cho, S. Shavitt, N. Chávez, and S. Weiner. 2014. "Response Heaping in Interviewer-Administered Surveys: Is It Really a Form of Satisficing?" *Public Opinion Quarterly* 78 (3): 591–633. DOI: https://doi.org/10.1093/poq/nfu017.

Hox, J.J. 1994. "Hierarchical Regression Models for Interviewer and Respondent Effects." *Sociological Methods & Research* 22 (3): 300–318. DOI: https://doi.org/10.1177/0049124194022003002.

Huttenlocher, J., L.V. Hedges, and N.M. Bradburn. 1990. "Reports of Elapsed Time: Bounding and Rounding Processes in Estimation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16 (2): 196. DOI:.https://doi.org/10.1037/0278-7393.16.2.196.

Knäuper, B., R.F. Belli, D.H. Hill, and A.R. Herzog. 1997. "Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality." *Journal of Official Statistics* 13 (2): 181–99. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/question-difficulty-and-respondents39-cognitive-ability-the-effect-on-data-quality.pdf.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (3): 213–36. DOI: https://doi.org/10.1002/acp. 2350050305.

McNeish, D., L.M. Stapleton, and R.D. Silverman. 2017. "On the Unnecessary Ubiquity of Hierarchical Linear Modeling." *Psychological Methods.* 22 (1): 114–140. DOI: http://dx.doi.org/10.1037/met0000078.

Menon, G. 1993. "The Effects of Accessibility of Information in Memory on Judgments of Behavioral Frequencies." *Journal of Consumer Research* 20 (3): 431–440. DOI: https://doi.org/10.1086/209359.

Olson, K., and I. Bilgen. 2011. "The Role of Interviewer Experience on Acquiescence." *Public Opinion Quarterly* 75 (1): 99–114. DOI: https://doi.org/10.1093/poq/nfq067.

Olson, K., and A. Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly* 71 (2): 273–286. DOI: https://doi.org/10.1093/poq/nfm007.

Olson, K. 2010. "An Examination of Questionnaire Evaluation by Expert Reviewers." *Field Methods* 22 (4): 295–318. DOI: https://doi.org/10.1177/1525822X10379795.

Rizzo, L., J.M. Brick, and I. Park. 2004. "A Minimally Intrusive Method for Sampling Persons in Random Digit Dial Surveys." *The Public Opinion Quarterly* 68 (2): 267–274. DOI: https://doi.org/10.1093/poq/nfh014.

Rabe-Hesketh S., and A. Skrondal A. 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*, Newbury Park, CA: Sage.

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*, edited by N. Schwarz and S. Sudman: 65–88. San Francisco: Jossey-Bass.

Schonlau, M. 2018. *HOTDECKVAR: Stata module for hotdeck imputation*, Statistical Software Components S458527, Boston College Department of Economics, revised 19 April 2022. Available at: https://EconPapers.repec.org/RePEc:boc:bocode:s458527.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. New York, NY: Cambridge University Press.

Tourangeau, R., and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83. DOI: https://doi.org/10.1037/0033-2909.133.5.859.

West B.T., K.B. Welch, and A.T. Galecki. 2015. *Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition*, Boca Raton, FL: CRC Press.

Yan, T., and K. Olson. 2013. "Analyzing Paradata to Investigate Measurement Error." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter: 73–96. New Jersey: John Wiley & Sons.

Received January 2021
Revised September 2021
Accepted May 2022