Survey Research Methods (2022)
Vol. 16, No. 2, pp. 207-226
doi:10.18148/srm/2022.v16i2.7771
European Survey Research Association

The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration

Steven Hope
University College London, UK

Pamela Campanelli
The Survey Coach, UK

Gerry Nicolaas
Natcen, UK

Peter Lynn
Institute for Social and Economic Research
University of Essex, UK

Annette Jäckle
Institute for Social and Economic Research
University of Essex, UK

The presence of an interviewer (face-to-face or via telephone) is hypothesized to motivate respondents to generate accurate answers and reduce task difficulty, but also to reduce the privacy of the reporting situation. To study this, we used respondents from an existing face-to-face probability sample of the adult general population who were randomly assigned to face-to-face, telephone and web modes of data collection. The prevalence of indicators of satisficing (e.g., non-differentiation, acquiescence, middle category choices and primacy/recency effects) and socially desirable responding were studied across modes. Results show differences between interviewer-administered modes and web in levels of satisficing (non-differentiation, and to some extent acquiescence and middle category choices) and in socially desirable responding. There was also an unexpected finding of how satisficing can differ by mode.

Keywords: mode effects, interviewer presence, interviewer effects, satisficing, non-differentiation, acquiescence, middle category effects, primacy and recency, social desirability

1 Introduction

Due to the costs of data collection for interviewer-administered surveys of household respondents, there has been a push towards internet first, mixed modes studies. For example, in several National Statistical Institutes (NSIs) around the world (see Betts & Lound, 2010a; Couper, 2012) the aim is now for as many respondents as possible to complete web surveys rather than using traditional telephone and face-to-face modes of data collection. The European Statistical System (ESS) started investigation of using web and mixed mode surveys for data collection of social surveys in 2012 (see Blanke, Luiten, Betts, & Lound, 2014). Another example is Understanding Society: The UK Household Longitudinal Survey (UKHLS), which experimentally tested web-first data collection in an separate "innovation panel" from 2009 to 2015 before introducing it in the main survey from 2016 (Burton, Lynn, & Benzeval, 2020; Jäckle, Lynn, & Burton, 2015). U.S. Census Bureau (2010) started with a mixed modes design in 2005. Here mail was *the least expensive method of data collection, and the success of the program depend[ed] on high levels of mail response.*This was followed by telephone and face-to-face methods. In 2018, the data collection method changed to web, followed by mail and then face-to-face (U.S. Census Bureau, 2020).

Using more than one mode of data collection can increase response rates and improve coverage of certain subpopulations.[1] However, having different respondents fill in the same questionnaire in different modes can lead to differential measurement error.[2] With the push to mixed modes

Contact information: Steven Hope, Population, Policy and Practice Research and Teaching Department, UCL GOS Institute of Child Health, 30 Guilford Street London WC1N 1EH, UK. E-mail: s.hope@ucl.ac.uk

[1] For a detailed explanation of mixed mode possibilities see de Leeuw (2005).

[2] Groves (1989) suggests that measurement error can emerge from the mode of data collection, the instrument [questionnaire], the respondents and the interviewer.

designs, there is increasing need to understand these measurement error differences. In this paper we examine measurement error differences between interviewer-administered surveys (face-to-face and telephone) in comparison to web surveys; that is, we focus on measurement differences due to the presence or absence of an interviewer.

Differences in measurement error between survey modes can be difficult to quantify, although a number of designs have been employed in an attempt to assess it. For example, these include (1) repeated measures designs where the same respondents in different modes are used to detect whether the change in mode is associated with changes in answers (Ofstedal, McClain, & Couper, 2021); (2) a random group of respondents change mode during the interview (Heerwegh, 2009); (3) comparisons with "gold-standard" data such as administrative records linked to survey data (Beste, Sakshaug, & Trappmann, 2021); (4) randomized field experiments (Laaksonen & Heiskanen, 2013) and (5) randomised hall tests (Jäckle, Roberts, & Lynn, 2008). All of these methods have limitations[3], but in other disciplines, such as evidence-based medicine, randomised designs are considered to be at the top of the hierarchy of evidence in terms of the quality of design (Guyatt et al., 2000).

This paper makes several contributions to the mixed mode literature. First, our focus on mode differences is with respect to measurement error rather than nonresponse or noncoverage. Second, our exploration of the presence or absence of the interviewer goes beyond a focus on social desirability bias. Third, we used a fully randomized experiment, assigning respondents to face-to-face, telephone and web data collection. Fourth, participants were drawn from the adult population of the UK; designs that compare with web often use student populations (e.g. Smyth, Christian, & Dillman, 2008) or existing internet panels (e.g. Grandjean, Nelson, & Taylor, 2009). Fifth, we used a probability sample of those who had access and used the internet rather than a convenience (opt-in) sample as is often used for internet panels (Baker et al., 2010; Scherpenzeel, 2011). Probability samples offer better quality data than convenience samples (Erens et al., 2014; Yeager et al., 2011). Sixth, our analysis controlled for the effects that interviewers have on their workloads as measured by correlated interviewer variance. This aspect is often ignored (for example, none of the mode comparison studies cited in Section 2.1 accounted for correlated interviewer variance). Seventh, our examination of question format difficulty as a facet of mode effects is a new approach to the evaluation of measurement error.

2 Background

Each mode, whether face-to-face, telephone or web, comprises a number of measurement factors. Figure 1 provides our overarching conceptual model, the features of which are explained in more detail in both this section and Section 2.1. For example, the model includes whether the stimulus is visual versus aural and the reply is spoken or typed. In the "reporting situation" under "task difficulty" there are the skills required by the respondent to complete the survey (e.g., literacy, numeracy and IT skills) and the impact of the respondent's comfort with a particular mode (called media-related factors by de Leeuw, 1992). In terms of questionnaire factors ("cognitive demands of Q"), question wording and format can vary in level of difficulty and differ by mode. In terms of respondent factors, the impact of the respondent's knowledge, opinion, ability and motivation to answer a survey question are at play. Tourangeau, Rips, and Rasinski (2000) suggest that the quality of a respondent's answer depends on thoroughness of the following four cognitive steps: comprehension of the survey question retrieving relevant information from memory, formulating a judgment, and selecting a response. Krosnick (1991) labels the respondent's behavior of going through the four steps comprehensively as "optimizing"; while "strong satisficing" represents the opposite, with no retrieval or information integration, and "weak satisficing" is in between, with "incomplete or biased information retrieval and/or information integration" (p. 213). Examples of weak satisficing include respondents selecting the first response option that constitutes a reasonable answer (primacy effects)[4] or agreeing with questions that make a one-sided assertion (acquiescence bias)[5]. Examples of strong satisficing include respondents selecting "don't know" when an answer is known, choosing the status quo response on an attitude question when an opinion is held, selecting the same response for every question (non-differentiation; sometimes called "straightlining"), or answer randomly. Kros-

[3] In addition to the measurement error differences between modes, each design has its own limitations. (1) Repeated measures approach—a time lag between data collection periods can confound inconsistent responses with true change; (2) The random group changing mode during the interview—this design could suffer from differential nonresponse bias if those completing the later mode are different than those who do not; (3) Comparison to gold-standard data method—access to such data are often limited and the administrative records may only contain a few variables of interest or contain errors; (4) Randomised field experiments – issue ex ante identical samples to different modes, but can have differential non-response bias between modes; (5) Randomised "hall tests"—randomly allocate participating respondents to mode, removing differential nonresponse bias in mode comparisons, but do not reflect standard data collection methods (e.g., respondents are not in their homes and, for web and telephone, the ability to do something else at the same time Campanelli, Blake, Mackie, & Hope, 2015).

[4] "Primacy refers to the [respondent's] tendency to more frequently choose from among the first categories" and recency the reverse, with the last categories more likely to be chosen. In both cases, this is done regardless of the category content (Dillman, Smyth, & Christian, 2014, p. 104).

[5] Acquiescence bias is agreeing to an agree/disagree statement regardless of its content (Holbrook, 2008).

nick's model of satisficing suggests that the likelihood of satisficing increases with the overall difficulty of providing an answer to a survey question ('task difficulty') and decreases with the respondent's "ability" and "motivation". Krosnick represents this relationship as task difficulty divided by the product of ability and motivation.

The presence or absence of an interviewer can influence whether, and to what extent, respondents engage in satisficing behaviors. But a negative aspect of the presence of an interviewer is social desirability bias. This occurs when respondents overstate socially desirable behavior and understate socially undesirable behavior. As suggested by Figure 1, whether the survey is "interviewer administered" influences the "reporting situation", which in turn influences the respondent's "willingness to disclose". As can be seen, the "reporting situation" includes more than just the "privacy of reporting". It includes the "quality of interviewer/respondent rapport" and the respondent's "perceived legitimacy of the survey".

In summary, we propose that features of the survey mode, such as the presence or absence of an interviewer, can influence how respondents process a survey question. If respondents completing a survey in one mode systematically differ in how they process survey questions compared to respondents completing the same questions in another mode, this will lead to differences in measurement error between modes. In addition to showing Tourangeau et al. (2000) four cognitive steps, Krosnick (1991) model for satisficing and the path to social desirability, Figure 1 also shows how "context information" can affect how the survey question is processed, which can lead to "other response effects" than satisficing and social desirability bias (e.g. "characteristics of visual layout", "ability to see the whole questionnaire" and question context effects through the "influence of prior questions"). In turn, the "context information" can be affected by whether the survey is interviewer-administered or not, how much control the respondent has over the questionnaire, and whether the question stimulus is presented visually or aurally. More details on the presence or absence of the interviewer and the outcomes of satisficing or social desirability bias are provided in Section 2.1.

2.1 The Role of the Interviewer in Producing Mode Effects

There are several potentially beneficial consequences of the use of interviewers. Interviewers can actively encourage respondents to make sufficient effort in processing the survey question. As shown in Figure 1 (under "respondent motivation"), this can be done both verbally and non-verbally. Non-verbal influence can be expressed in a number of ways. For example, "if interviewers are professional and diligent and model their engagement in the process effectively for respondents, this may be contagious and may inspire respondents to be more effortful than they would be without such modeling. Likewise, the presence of an interviewer may create a sense of accountability in respondents, who may feel that they could be asked at any time to justify their answers to questions... Such accountability is believed to inspire more diligent cognitive effort and more accurate answering of questions" (Baker et al., 2010, p. 737). Holbrook, Green, and Krosnick (2003) suggest that non-verbal communication (e.g., smiling, nodding) between the face-to-face interviewer and respondent is both positive and motivating. de Leeuw (1992) points out that face-to-face interviewers can also monitor the non-verbal expressions of the respondent. Telephone interviewers can use non-verbal minimal sounds, such as "mmmm" and "uhuh" to motivate respondents, while also listening to these from the respondent.

In addition, in face-to-face surveys the interviewer can reduce "task difficulty" by improving the "reporting situation". He or she can reduce "respondent distractions" (e.g., asking for the television to be turned off, or for the interview to take place is a quieter room). According to de Leeuw (1992), in face-to-face interviewing the respondent and the interviewer share the locus of control (see "R control over Qaire"). As a consequence, the interview can be carried out at a pace that is comfortable for both interviewer and respondent (less "time pressure"). In contrast, as telephone interviewers are not physically present, they are less able to reduce respondent distraction, such as multi-tasking while participating in the survey (Holbrook et al., 2003). In addition, as suggested by de Leeuw (1992), the locus of control is not balanced but favors the telephone interviewer. "There is a tendency to avoid silences in a telephone conversation" (p. 15) which can increase the feel of "time pressure".

Interviewers can also reduce "task difficulty" by lessening the "cognitive demands of the question". They can take on the role of facilitator, actively helping to reduce "task difficulty" through clarifying how the task should be completed and offering standard definitions for a question, when asked. Indeed, it is a duty of both face-to-face and telephone interviewers to "answer the respondent's queries, and to probe to clarify answers" (de Leeuw, 1992, p. 18). In summary, these aspects of the interviewer's role should lead to less satisficing in interviewer-administered modes.

In contrast, the absence of an interviewer can lead to satisficing in web surveys. In a review of the literature on online panels, the Baker et al. (2010) taskforce commented that "... replacing an interviewer with a computer for self-administration has the potential to increase the likelihood of satisficing due to the ease of responding (simply clicking responses without supervision)" (p. 736). That is, respondents might feel that they are less likely to be held accountable ('perceived accountability') for their answers and are therefore less motivated to process the survey questions diligently. The absence of an interviewer also increases op-

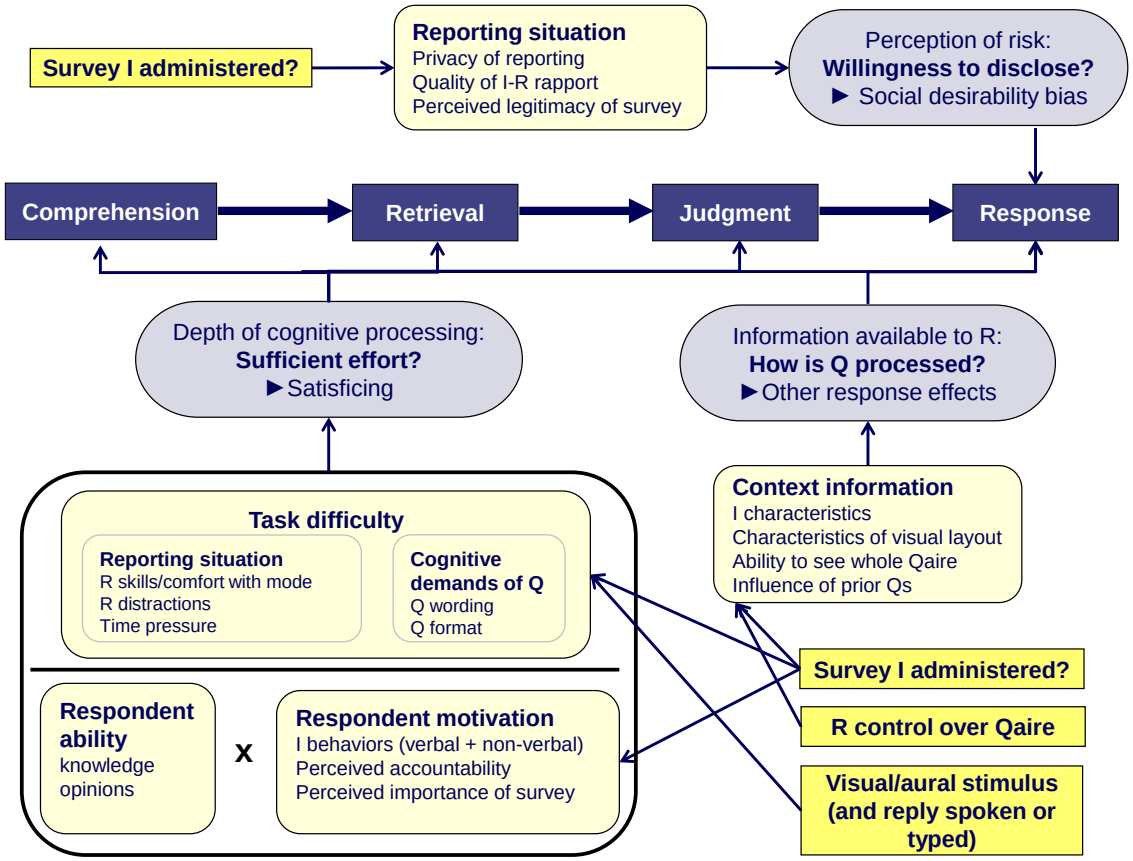

Figure 1. Conceptual model for measurement error differences between interviewer-administered and web surveys (I = interviewer, R = respondent and Q = survey question)

portunities to lie in web surveys (e.g., about age, income or place of residence) (Revilla, Saris, Loewe, & Ochoa, 2015) or "to rush through a self-administered questionnaire without reading the questions carefully or thinking thoroughly when generating answers" (Baker et al., 2010, p 373). An example of this latter point is given by Zhang and Conrad (2013), who focused on the consequences of the speeding behaviour of respondents in an online panel. They found that early speeders continued to speed, that speeding was connected to non-differentiation in batteries of questions and that this was true for respondents with varied characteristics, although more common in those with low levels of education. The following studies have explored other indicators of satisficing, comparing web surveys to other modes (e.g., face-to-face and/or telephone interviewing and one with aural delivery over an intercom to simulate telephone interviewing). There were mixed results for non-differentiation. Heerwegh and Loosveldt (2008) and Chang and Krosnick (2010) found more non-differentiation in web surveys while Cernat and Revilla (2020) found no differences. Duffy, Smith, Terhanian, and Bremer (2006) found more middle category selection in web surveys. Heerwegh and Loosveldt (2008) and the U.K. Cabinet Office Consultation (2013) found more selection of "don't know" answers in web surveys. Chang and Krosnick (2010) found more recency effects for aural delivery and Cernat and Revilla (2020) found more primacy in visual delivery. Cernat and Revilla (2020) found more item nonresponse in web surveys. de Leeuw (1992) and Groves and Kahn (1979) found that respondents gave less complete answers to open-ended questions in self-completion surveys. Comparing open-ended and closed-ended questions, Reja, Lozar Manfreda, Hlebec, and Vehovar (2003) found more inadequate and missing answers for the open-ended questions in web surveys.

Despite the strengths inherent in interviewer administration of surveys, it also has drawbacks. The impact of interviewer errors (e.g., not writing down verbatim responses on open questions or not marking the reply that corresponds to the respondent's answer) can be mitigated through training and supervision. However, not all interviewer effects can be reduced so easily. There are specific circumstances where "interviewer characteristics" can affect the respondent's an-

swer. According to (Fowler & Mangione, 1990, p. 105), this occurs "when the topic of a survey is very directly related to some interviewer characteristic so that potentially a respondent might think that some of the response alternatives would be directly insulting or offensive or embarrassing to an interviewer". An example of this was shown by Schuman, Steeth, and Bobo (1985). White Southern respondents in the US were asked a sensitive question of their views of Black people; more truthful answers were obtained with local southern face-to-face interviewers than with the telephone interviewers with a northern accent. Interviewer effects can also have an impact on statistical outputs. Interviewers' random errors will be captured in the statistical simple variance associated with a survey question. However, interviewers' consistent errors across a number of interviews (referred to as correlated variance) inflate the standard errors associated with a survey question (see full details in Section 3.3). Finally, it is known that the presence of an interviewer can lead to social desirability bias (see Figure 1). For both face-to-face and telephone interviewing, the interviewer's presence can lead to a reduction in the privacy of the reporting situation (de Leeuw, 2005; Kreuter, Presser, & Tourangeau, 2008; Tourangeau, Conrad, & Couper, 2013). If the questions are sensitive, this can have a detrimental effect on the respondent's willingness to answer truthfully. Here we would also expect differences between face-to-face (where the interviewer and respondent are in the same location) and telephone (where the interviewer is physically separated from the respondent), with higher levels of social desirability in face-to-face interview settings. However, this is likely to be mitigated by a higher quality of "interviewer-respondent rapport". It is also easier for the respondent to validate the "legitimacy of the survey" with face-to-face interviewers and their ID badges. This is consistent with the results of Jäckle, Roberts, and Lynn (2006), who found that "telephone respondents [as opposed to face-to-face respondents] were more likely to give socially desirable responses across a range of indicators" (p. 4). Similar conclusions had been reached by Holbrook et al. (2003).

In this paper we concentrate on the satisficing and social desirability bias aspects of the overarching model in Figure 1, and on question format rather than question wording. We also account for the effects of non-random interviewer error by correcting for correlated interviewer variance in the analyses.

2.2 The Impact of Question Format on Mode Effects

Very little has been written about the impact of question format on mode effects. There is evidence that certain question formats are intrinsically more difficult for respondents to complete. As reviewed in Section 2.1, we expect the presence or absence of an interviewer to have a noticeable effect on answers for difficult question formats. Some examples of difficult question formats are:

- Ranking task. This requires the respondent to understand the nature of ranking a list of options. Ranking has been shown to demand "considerable cognitive sophistication and concentration" on the part of the respondent, particularly when there is a long list of options (Alwin & Krosnick, 1985, p. 536; Fowler, 1995). Another potential indicator of difficulty is that a ranking task takes longer to complete than the equivalent series of separate rating questions (Reynolds & Jolly, 1980; Taylor & Kinnear, 1971).
- Agree/disagree statements. For example, rating your health with five categories from "excellent" to "poor" is easier than answering the agree/disagree statement, "my health is excellent" with five categories from "strongly agree" to "strongly disagree" (Fowler, 1995). If your health is "very good', 'good', 'fair" or "poor", you could validly choose "disagree". Agree/disagree statements also confound two dimensions: the respondent's attitudinal position (i.e., agree or disagree) and the intensity of his/her feeling (e.g., strongly agree versus agree) (Fowler, 1995). Negatively worded statements also add to respondent difficulty; for example: "disagreeing that one is seldom depressed is a complicated way of saying one is often depressed" (Fowler, 1995, p. 56).[6]
- Number of answer categories. The cognitive complexity of answering a survey question increases with the number of answer categories (Krosnick & Presser, 2010). Similarly, DeCastellarnau (2018, p. 1531), notes "too many categories may reduce the clarity of the meaning of the options and limit the capacity of respondents to make clear distinctions between them". Fowler (1995, p. 53) suggests "5 to 7 categories is probably as many categories as most respondents can use meaningfully for most rating tasks".
- Aurally administered questions with five or more categories with no showcard. Answering a question is easier when it is presented visually (chapter 6 of Dillman et al., 2014; chapter 5 of Tourangeau et al., 2013).

2.3 Hypotheses

We hypothesize the interviewer-administration (in both face-to-face and telephone modes) will have an impact both on the extent of (A) satisficing and (B) socially desirable responding in comparison to self-administration (web mode). This will potentially lead to differences in measurement error between modes.

For point A: satisficing, we expect interviewers to motivate respondents (and assist, where appropriate) and for web respondents to be prone to satisficing behavior (as described in Section 2.1). More specifically, we hypothesize that the following satisficing behaviors by difficult question format

[6] In addition, the agree / disagree format has been found to be less reliable and valid than comparable rating items (Saris, Revilla, Krosnick, & Shaeffer, 2010) and is prone to acquiescence bias particularly among the less-educated (Fowler, 1995).

will be more likely to occur without the presence of an interviewer.

Hypothesis 1—Duplicates and non-differentiation in ranking tasks: Without the motivation and help of an interviewer, we expect poorer quality data for the complex task of ranking responses, with more duplicates and non-differentiation in web than in the interviewer-administered modes.

Hypothesis 2—Acquiescence response bias on agree/disagree questions: Although acquiescence falls under Krosnick (1991) list of satisficing behaviors and Betts and Lound (2010b, p. 39) confirm that "it requires less effort to agree than to generate reasons to disagree", there are alternative explanations[7]. The literature is inconclusive, and therefore we are using a two-tailed hypothesis for Hypothesis 2, i.e., that acquiescence will differ between interviewer-administered and web surveys.

Hypothesis 3—Middle category satisficing: Without the motivation of an interviewer, respondents may satisfice by choosing the middle option when they have a positive or negative opinion. We therefore expect that web respondents will be more likely than those in interviewer-administered modes to satisfice by selecting the middle category.[8]

Hypothesis 4—Primacy and recency effects on questions with five or more categories: Without the motivation of an interviewer, we firstly expect respondents using the web to be more likely to show primacy effects than respondents in interviewer-administered modes that use visual aids, such as showcards in face-to-face interviewing. Secondly, we expect more evidence of primacy effects in the visual web mode than recency effects in the aural interviewer-administered modes (telephone and face-to-face without showcards).[9,10]

For point B: social desirability, we hypothesize that, by reducing the privacy of reporting, the presence of an interviewer will reduce the respondent's willingness to disclose potentially sensitive information.

Hypothesis 5—Social desirability: We expect socially desirable responding to be more prevalent in the interviewer-administered modes compared to web mode. We also expect there will be a difference between face-to-face and telephone interviewers. Although the face-to-face interviewers are physically present, we hypothesise that the better quality of rapport in face-to-face will lead to less social desirability bias in CAPI than CATI.

3 Methods

3.1 Data and Experimental Design for the Mixed-Modes Study

The NatCen Social Research Omnibus survey used a probability sample of adults aged 16 and over in Great Britain, with clients buying questionnaire space.[11] The survey was administered quarterly to a fresh sample of respondents. 1,600 face-to-face interviews were typically completed using CAPI at each wave, with response rates averaging 54% (American Association for Public Opinion Research, 2016, response rate RR5). To achieve a large enough sample size for the mixed modes study[12], two waves of the Omnibus survey were combined. The sample comprised respondents from the Omnibus survey who had agreed to participate in future research (83% agreed to be re-contacted) and who had indicted in the Omnibus Survey that they had access to the internet. A subset of these respondents were randomly allocated to one of three modes: CAPI, CATI and web based on the expected response rate (See Table 1). Separate surveys for each of the three modes were collected between January and June of 2009.

The sample sizes for some parts of the questionnaire were further reduced because mode of data collection was crossed with seven format experiments.[13]

Nonresponse from the Omnibus survey and from those who refused to be re-contacted does not affect the mode comparisons. However, there may be differential nonresponse bias across the modes of the experiment. To account for this, we decided not to use standard weighting (such as post-stratification weighting) because our focus was on the subset of the population who have internet access and usage. Propensity score weighting would have been difficult to use

[7] Alternative explanations for acquiescence bias include ambiguity of the agree/disagree statement itself (McBride & Moran, 1967; Peabody, 1966), characteristics of the respondent e.g., less educated (Landsberger & Saavedra, 1967; Schuman & Presser, 1981), deference to the interviewer (Javeline, 1999; Lenski & Leggett, 1960) and category fallacy (i.e., choosing a "safe" category because of a concern about looking foolish or ignorant) (Jackman, 1973; Warnecke et al., 1997). In a systematic review of satisficing, Roberts, Gilbert, Allum, and Eisner (2019) found eight studies that supported acquiescence as an indicator of satisficing and six which did not.

[8] This is consistent with the findings of Duffy et al. (2006, p. 629) that "online survey respondents seem much more inclined to select the neutral point ('neither agree nor disagree') than face-to-face respondents."

[9] Assuming all answer categories are equally plausible, the expectation is that primacy effects are most likely to occur in the visual modes and recency effects in aural modes (Dillman et al., 2014).

[10] It was not possible to investigate whether nonresponse differed by mode (a possible marker of satisficing) because there was virtually no item nonresponse in any mode of this experiment.

[11] At the time this paper was written, NatCen Social Research was no longer running its Omnibus survey.

[12] The data from this mixed mode experiment are available through the UK Data Service (NatCen Social Research, 2014).

[13] (1) few answer categories versus many, (2) rating versus ranking, (3) agree/disagree statements versus forced choice, (4) "yes/no" versus "mark all that apply", (5) branching versus non-branching, (6) fully-labeled versus end-labeled items and (7) showcard versus no showcard on long lists in CAPI. Experiments (1) and (7) were crossed for CAPI.

Table 1
Allocation of cases, response rates and achieved sample sizes

	CAPI	CATI	Web	total
Randomly allocated cases (based on expected response rate)	521	596	829	1946
Cases after exclusions due to non-use of the internet in web mode (ineligible - unable to do survey)	NA	NA	744	
Cases in dataset	380	409	349	1138
Response rate (cases in dataset/eligible allocated cases) - response rate RR5	73%	69%	47%	
Non-use of internet cases in interviewer modes (excluded for comparability with web)[a]	98	95	NA	
Cases used in analysis	282	314	349	945

[a] Weighting and modelling are useful techniques, but it is not possible to know if all the bias is removed. The exclusion of interviewer-administered cases who had access to the internet, but were not internet users, guaranteed that any possible bias due to these cases would be removed before comparison with web.

as there were more than two modes. Therefore we decided to include key sociodemographic variables as covariates in our regression models in order to account for differences in the characteristics of respondents between the three modes. In this internet access and use sample, sex, age, ethnicity, marital status and labor force status were related to mode differences, but education was not.[14]

Using the "uni-mode" approach (Dillman, Smyth, & Christian, 2007)[15], the questionnaire was designed to be as similar as possible across modes, with identical question order and wording. Data were collected and processed in the same way for each mode by NatCen Social Research.[16] Interviewers were trained to read the survey questions as worded. Interviewer-administered surveys often include "definitions" of ambiguous terms in look-up screens and hyperlinks can be provided in web surveys for "definitions", but this study purposely did not use any of these. All instructions read out by interviewers were duplicated in the web version. Similarly, there were no error checks or error messages in web surveys or interviewer administered modes. However, interviewers were allowed to answer respondent questions about the task itself, to probe on any inadequate answers from the respondent and face-to-face interviewers could pick up on and react to visual cues of miscomprehension. These interviewer behaviors reflect the potential positive interviewer influence which is of interest in the experiment.

3.2 Analysis Methods

Using SPSS, we estimated logistic regression models for indicators of satisficing or social desirability (dependent variables), with mode of data collection as the independent variable, accounting for the nonresponse bias control variables (sex, age, ethnicity, marital status and labor force status) listed in Section 3.1. Because of small sample sizes, findings at the $\alpha < .10$ level are reported. One-sided tests were used for all hypotheses except Hypothesis 2 on acquiescence. As there were multiple tests for each hypothesis, we used the Šidák correction (Abdi, 2007). The Šidák adjusted α is equal to $1 - (1 - \alpha)$ raised to the power of $1/m$, where m is the number of tests. We have also listed the percentage of cases for a given dependent variable across the modes to show patterns of results. All of the variables included in the analyses are listed in Table 2. In some cases, new dependent variables were derived:

1. For the ranking questions, we explored duplicates and non-differentiation. The duplicate ranks indicator was given a value 1 if the respondent had assigned any of the options the same ranking, and 0 otherwise. The non-differentiation indicator was given a value 1 if the respondent had picked the same ranking throughout or the same ranking for all but one of the questions, and 0 otherwise.

2. For the agree/disagree questions, the first acquiescence indicator took the value 1 if the respondent answered "strongly agree" or "agree" on a question, and 0 otherwise. Agree/disagree questions are often analysed in this way because research has shown that "response style may have more to do with people's willingness to choose the extreme response than with differences in the opinions being compared" (Fowler, 1995, p. 66; see also "response contraction bias" in Tourangeau et al., 2000). The second acquiescence indicator was at the scale level. In psychometric multi-item scales,

[14] The control variables were used in the following formats: sex (male, female), age (16–24, 25–34, 35–44, 45–54, 55–64, 65 and older), ethnicity (white, other), marital status (single, married/civil partner, separated/divorced/widowed) and labor force status (working, unemployed/sick, retired, other). They were entered as a series of dummy variables in the logistic regression with the first category being the omitted one.

[15] The 2007 version of this book provides a more complete description of the "uni-mode" approach than later editions.

[16] At the time of the data collection, smart phone usage was not prevalent. The Office of Communications (2018)—the regulatory body of UK's broadcasting, telecommunications and postal communications—suggests that only 17% of UK residents in 2008 owned a smartphone. In this study, no special instructions were given to respondents randomly allocated to web regarding type of device to use.

acquiescence behavior is typically identified by respondents agreeing to opposite statements (DeVellis, 2016). For each scale, the indicator took a value of 1 if the respondent agreed to two or more opposite statements, and 0 otherwise.[17]

3. For middle category satisficing the indicator took a value of 1 if the respondent had chosen the middle category, and 0 otherwise.

4. The primacy indicator took the value 1 if the respondent selected the first response option and 0 otherwise; the recency indicator took the value 1 if the respondent selected the last option, and 0 otherwise.

5. For the sensitive questions, the indicator took a value of 1 if the respondent had chosen "strongly agree or agree" when this was the socially desirable response, and 0 otherwise.

3.3 Accounting for Correlated Interviewer Variance

Correlated variance occurs when the results within a cluster (in this case, the interviewers' workloads) are more homogeneous than the sample as a whole. Homogenous clusters inflate sampling variance. A general indicator of the effect of a complex survey design on sampling variance is a design effect (DEFF). The DEFF is the complex design variance divided by the equivalent simple random sample variance.[18]

The complex survey design created by correlated interviewer variance is a one-stage cluster design. The clusters are designated by interviewer identification number and are classified as primary sampling units (PSUs).[19] Web survey respondents have no interviewer identification number and are therefore excluded from the analysis because they are not assigned to a PSU. Assigning an overall interviewer identification number to web respondents leads to unusual standard errors in the results because the homogeneity of web respondents has been factored in.[20] As a consequence, it was not possible to use the standard software approaches for handling complex designs for comparisons between interviewer and web modes. Thomas and Heck (2001, p. 530) suggested three ways to compare interviewer and web modes, accounting for correlated interviewer variance and the absence of an interviewer in the web mode: (1) "adjust the estimated standard errors ... [by] a known DEFT" (the square root of DEFF), (2) use "the effective sample size [NEFF] by adjusting the relative weight downward" (NEFF = $\frac{n}{\text{DEFF}}$) and (3) use "a more conservative critical alpha value". We adopted a variation of methods 1 and 2, using the effective sample size as the criterion for reducing the sample size and therefore adjusting the estimated standard errors. We created the effective sample size by taking a random sub-selection of CAPI cases and a separate random sub-selection of CATI cases, bringing n down to NEFF based on the DEFFs for that particular variable and data collection mode. We then re-ran the logistic regressions.

4 Results

4.1 Results of Interviewer Effects on Satisficing

Hypothesis 1: Duplicates and non-differentiation in ranking tasks. This hypothesis investigated the extent to which duplicates and non-differentiation in ranking tasks would be greater in web than CAPI. CATI was excluded as long ranking tasks are not possible without visual aids. There were two ranking tasks (Table 2: questions 19 and 20). The first ranking task asked which geographical unit of the respondent's address (from street, city, rising to UK and European level) was most important, a task that was phrased in terms of an address game. The second task asked which changes to the respondent's neighborhood would be most important to improve it.

First, we examined an indicator of duplicate ranks. Hypothesis 1 was supported (Table 3); duplicate ranks were significantly more prevalent among web respondents than CAPI respondents, especially for the address game question.

[17] Note that the quality of neighborhood scale only had one pair or opposite statements: "compared to other neighborhoods, this neighborhood has more properties that are in a poor state of repair" as opposed to "compared to other neighborhoods, this neighborhood has more properties that are well-kept". The sensitive scale had two pairs of opposite statements: "I would be concerned for my family's safety if housing were provided near my home for people who were leaving prison" as opposed to "people who have been in prison have as much right to live in my neighbourhood as any other people" (and an identical pair of statements for people with mental health problems living in the neighborhood). The scale for the thoroughness of preparation before making a large financial decision had two statements showing more preparation for the decision ("I would do a lot of research before making an important financial decision" and "I definitely would talk to family and friends before making an important financial decision") and two showing less preparation ("I would rarely talk to a financial advisor before making an important financial decision" and "I would rarely read all the small print before making important financial decisions").

[18] According to Kish (1965), the formula for the DEFF for cluster sampling is $1+\rho(b-1)$ where ρ is the intracluster correlation coefficient (measuring cluster homogeneity) and b is the cluster size. The DEFF for interviewers is equivalent: $1 + \rho_{\text{interviewers}}(m - 1)$ where $\rho_{\text{interviewers}}$ is the intraclass correlation coefficient and m is the interviewer workload size. (m-bar may be substituted for m if interviewers have different workload sizes.) In most face-to-face survey fieldwork, an interviewer is assigned to a geographic cluster; confounding these two sources of correlated variance. In our study, the face-to-face interviewer workload sizes were very small (1 to 5 respondents), and so any confounding is minimal.

[19] For the exploration of the complex survey design, we used Stata.

[20] For example, the standard error for a comparison between CAPI with a DEFF greater than 1.0 and web (or equivalently for CATI with a DEFF greater than 1.0 and web) would sometimes be decreased rather than increased.

Second, we looked at an extreme form of duplicate ranking: non-differentiation. The prevalence of non-differentiation was lower than that for duplicate ranks. There was a significantly greater level of non-differentiation in web than CAPI for the address game question. Although in the same direction, the mode differences for non-differentiation in the neighbourhood question were small (1% for CAPI and 4% for web) and did not reach significance.

Hypothesis 2: Acquiescence response bias on agree/disagree questions. This hypothesis investigated whether acquiescence would differ between web and interviewer-administered modes. As described in Section 2.3, we did not predict a direction for the hypothesis because the literature is mixed. The hypothesis was tested using twelve agree/disagree questions from three multi-item scales (Table 2: questions 7-18[21]). The first scale comprised questions on the quality of the neighborhood (questions 7-10); the second scale contained questions on the thoroughness of preparation for a financial decision (questions 11-14), and the third scale had questions about people with mental health problems and former prisoners living in the respondent's neighborhood (questions 15-18) and was designed to be sensitive. Each of the three scales included both positively and negatively worded statements.

To investigate acquiescence, we first looked first at question level comparisons, focusing on the eight non-sensitive statements (questions 7–14) and the percentage of respondents who had agreed or strongly agreed to a question.[22] For the four neighbourhood questions, the percentages were similar across CAPI, CATI and web and there were no significant differences. For the four financial decision questions, CAPI and CATI respondents had higher acquiescence than web respondents, but this was only significant on two of the four questions (questions 11 and 12, for CATI>Web). Second, we examined acquiescence at the scale level. Table 5 shows the percentage of respondents who agreed to two or more opposite statements. Although the percentages of acquiescence were higher for the CAPI and CATI respondents compared to web, none of the results reached significance. The two significant results suggest that acquiescence is not due to satisficing, as suggested by Krosnick (1991), because it was more prevalent in the interviewer-administered modes than web. Overall, there was only some support for differences in acquiescence by mode.

Hypothesis 3: Middle category satisficing. We hypothesized that web respondents would be more likely to choose a middle category option compared to respondents in interviewer-administrated modes. This involved five category questions (question numbers 7–18).[23]

The percentage comparisons show that for eleven out of the twelve questions, web respondents were more likely than CAPI respondents to select middle categories, and this difference was statistically significant for five questions (Table 6). Similarly, for 9 of 12 questions, web respondents were more likely than CATI respondents to select middle categories, a difference that was statistically significant for three of the questions. Overall, these results provide some support for the hypothesis.

Hypothesis 4: Primacy and recency effects on questions with long lists of categories. We hypothesized that there would be more response order effects in web than in interviewer-administered modes. Six questions with seven or eight responses were used from the experiment that compared these to three category responses (Table 2: questions 1–6), crossed with a showcard/no showcard in CAPI. Primacy effects could be compared between web and CAPI with showcard, which allowed for the isolation of the effects of interviewer presence, holding visual presentation of the response options constant across modes. In the case of CAPI without a showcard and CATI, the comparison was between primacy in web and recency effects in the interviewer-administered modes. We expected the odds ratio for any significant primacy effects in web to be larger than the odds ratio for any significant recency effects for CAPI without a showcard or CATI.

Primacy effects (CAPI with showcard versus web). None of the results were significant and there was no clear pattern (Table 7).

Primacy effects in web versus recency effects in CAPI without showcard and CATI). There was only one instance of a CATI recency effect for any of the six questions in comparison to CAPI and CATI (Table 8), question 4 (amount spent on leisure activities). CATI respondents were more likely than web respondents to choose the last category (OR = 5.91, $p = 0.001$). Question 4 also showed the expected pattern of web primacy effects, with web respondents being more likely than CATI respondents to choose the first category, but this did not reach significance.

In summary, there is little evidence to support Hypothesis 4.

[21] Note that the three multi-item scales were not consecutive in the questionnaire. The actual question numbers were N35-N38, FM52-FM55, and N64-N67.

[22] The sensitive items were excluded, as both socially desirable responding to the positive statements (items 16 and18) and potentially more truthful answers to the negative statements (items 15 and 17) could be confused with acquiescence. For example, a respondent admitting the truth that he/she would worry if people with mental health problems were provided housing near his/her home (item 15) would be indicated by agreement with the statement.

[23] Although questions with 7 categories are known to be more difficult for respondents to answer than those with fewer categories, Weems (2004) suggests that increasing the number of categories decreases the selection of the middle category. We therefore did not include these questions.

Table 2
Wording and source of questions analysed in the mixed mode experiment

Question Formats	Showcard in CAPI?	Question ID (Actual Question #	Question Topic and Wording	Response Options for the Long Versions	Source
Long questions (from a long versus short question experiment) crossed with showcard/ no showcard in CAPI	A random half of CAPI respondents received a showcard and others did not	1 (N44)	SATISFACTION WITH STREET CLEANING: And how satisfied or dissatisfied are you with street cleaning?	7 categories (very satisfied, moderately satisfied, slightly satisfied, neither satisfied nor dissatisfied, slightly dissatisfied, moderately dissatisfied, very dissatisfied)	Citizenship Survey, 2007
		2 (N43)	SATISFACTION WITH WASTE AND RECYCLING COLLECTION: I would like you to tell me how satisfied or dissatisfied you are with local household waste collection, recycling collection and other recycling collection points. Would you say you are...	7 categories (very satisfied, moderately satisfied, slightly satisfied, neither satisfied nor dissatisfied, slightly dissatisfied, moderately dissatisfied, very dissatisfied)	Citizenship Survey, 2007 (modified to make question more difficult)
		3 (FM82)	LENGTH LIVED IN AREA: How long have you lived in this area?	7 categories (less than 12 months, 12 months or more but less than 2 years, 2 years or more but less than 3 years, 3 years or more but less than 5 years, 5 years or more but less than 10 years, 10 years or more but less than 20 years, 20 years or longer)	British Crime Survey, 2006
		4 (FM81)	AMOUNT SPENT ON LEISURE ACTIVITIES: How much do you personally spend in an average month on leisure activities, and entertainment and hobbies, other than eating out?	8 categories (less than £20, £20 - £39, £40 - £59, £60 - £79, £80 - £99, £100 - £119, £120 - £139, £140 or more)	British Household Panel Study, Wave 17
		5 (FM75)	TYPE OF DWELLING: Which of these best describes your home?	8 categories (detached house, semi-detached house, terraced house, bungalow, flat in a block of flats, flat in a house, maisonette, other)	Survey of Public Attitudes and Behaviors Towards the Environment, 2007
		6 (N39)	LOCATIONS NEAREST TO HOUSE: Which of the following is closest to where you live?	7 categories (a primary school, a secondary school, a 6th form college, a river, a lake, a cinema, a theatre)	New

Continues on next page

Continued from last page

Question Formats	Showcard in CAPI?	Question ID (Actual Question #	Question Topic and Wording	Response Options for the Long Versions	Source
Set of four agree/disagree statements	Without showcards	7–10 (N35–N38)	QUALITY OF NEIGHBORHOOD: The next few questions are about the extent to which you agree or disagree with statements about your neighborhood. Here is the first statement. • This neighborhood is not a bad place to live. • Compared to other neighborhoods, this neighborhood has more properties that are in a poor state of repair. • Compared to other neighborhoods, this neighborhood does not suffer from things like litter, dog mess and graffiti. • Compared to other neighborhoods, this neighborhood has more properties that are well kept.	5 categories (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)	Modified and extended from a Southern Housing Association questionnaire
Set of four agree/disagree statements	Without showcards	11–14 (FM64–FM67)	THOROUGHNESS OF PREPARATION BEFORE MAKING A LARGE FINANCIAL DECISION: To what extent do you agree or disagree with the following statements about making important financial decisions such as taking out a mortgage, loan or pension. • I would rarely read all the small print before making important financial decisions. • I would do a lot of research before making an important financial decision. • I would rarely talk to a financial advisor before making an important financial decision. • I definitely would talk to family and friends before making an important financial decision.	5 categories (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)	Attitudes to Pensions Survey, 2006 (question 3 newly added)

Continues on next page

Continued from last page

Question Formats	Showcard in CAPI?	Question ID (Actual Question #	Question Topic and Wording	Response Options for the Long Versions	Source
Set of Four SENSITIVE Agree/disagree statements	Without showcards	15–18 (N52–N55)	PEOPLE WITH MENTAL HEALTH PROBLEMS AND FORMER PRISONERS IN R'S NEIGHBORHOOD: How strongly do you agree or disagree with the following 4 statements. • I would worry if housing were provided near my home for people with mental health problems leaving hospital. • People who have serious mental health problems have just as much right to live in my neighborhood as any other people. • I would be concerned for my family's safety if housing were provided near my home for people who were leaving prison. • People who have been in prison have just as much right to live in my neighborhood as any other people.	5 categories (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)	Extended from the British Social Attitudes, 2006
Ranking task (from a rating versus ranking experiment)	With showcards	19 (N29)	CHILDREN'S ADDRESS GAME: Sometimes for their amusement, children give their address as Home Street, This town, Localshire, My country, United Kingdom, Europe, The World. Thinking in this way about where you live now and what is important to you generally in your everyday life, please rank the following 6 questions from 1 (meaning most important) to 6 (meaning least important).	6 categories (the street in which you live, the city or town in which you live, the county or region, for instance, Yorkshire, Lothian or East Anglia, the country in which you live for instance, England, Northern Ireland, Scotland, Wales, the United Kingdom, Europe).	British Social Attitudes, 2006
		20 (N45)	IMPROVEMENTS TO THE NEIGHBORHOOD: What would you consider most important in improving the quality of your neighborhood? Please rank the following 7 questions from 1 (meaning most important) to 7 (meaning least important).	7 categories (less traffic, less crime, more/better shops, better schools, more/better facilities for leisure activities, better transport links, more parking spaces)	National Survey of Culture, Leisure and Sport, 2005/6 (modified to ranking)

Table 3
Mode differences in two ranking tasks

		CAPI		Web		Logistic Regression results[a]		
ID	Dependent variables	%	n	%	n	comparison	OR	*p*-value
Duplicate ranks								
19	Geographic unit of respondent's address: Children's address game	18	133	49	179	Web>CAPI	5.1*	0.001
20	Importance of different changes to the neighborhood	16	141	29	157	Web>CAPI	2.2*	0.008
Non-differentiation								
19	Geographic unit of respondent's address: Children's address game	5	135	14	183	Web>CAPI	2.9*	0.023
20	Importance of different changes to the neighborhood	1	147	4	166	Web>CAPI	8.8	0.057

[a] Control variables: sex, age, ethnicity, marital status and labor force status
* $p \leq 0.026$ (Šidák adjusted α for 4 tests)

Table 4
Mode differences in acquiescence as measured by the percentage choosing 'strongly agree' or 'agree'

		CAPI		CATI		Web		Logistic Regression results[a]		
ID	Dependent variables	%	n	%	n	%	n	comparison	OR	*p*-value
Quality of neighborhood questions										
7	Neighborhood not a bad place (5 categories, agree/disagree)	90	282	89	314	85	349	CAPI>Web	1.6	0.060
								CATI>Web	1.7	0.048
8	More properties in bad state of repair (5 categories, agree/disagree)	10	134	13	159	10	183	Web>CAPI	1.0	0.964
								CATI>Web	1.1	0.863
9	Neighborhood does not suffer from litter, dog mess, graffiti (5 categories, agree/disagree)	60	282	57	314	56	349	CAPI>Web	1.2	0.259
								CATI>Web	1.1	0.562
10	More properties are well kept (5 categories, agree/disagree)	76	135	73	161	76	183	Web>CAPI	1.0	0.905
								Web>CATI	1.0	0.932
Thoroughness of preparation before financial decision questions										
11	Rarely read small print before financial decision (5 categories, agree/disagree)	35	281	43	314	29	349	CAPI>Web	1.4	0.090
								CATI>Web	1.9*	0.002
12	Would do a lot of research before financial decision (5 categories, agree/disagree)	90	282	91	314	85	349	CAPI>Web	1.6	0.056
								CATI>Web	1.9*	0.012
13	Rarely talk to financial advisor before financial decision (5 categories, agree/disagree)	41	281	48	314	39	349	CAPI>Web	1.4	0.523
								CATI>Web	1.4	0.067
14	Would talk to friends and family before financial decision (5 categories, agree/disagree)	71	134	71	161	64	183	CAPI>Web	1.2	0.450
								CATI>Web	1.3	0.480

[a] Control variables: sex, age, ethnicity, marital status and labor force status
* $p \leq 0.013$ (Šidák adjusted α for 8 tests for CAPI versus Web and 8 tests for CATI versus Web)

Table 5
Mode differences in acquiescence as measured by percentage agreement to opposite statements in multi-item scales

		CAPI		CATI		Web		Logistic Regression results[a]		
IDs	Dependent variables	%	n	%	n	%	n	comparison	OR	*p*-value
7-10	Neighborhood scale	4	134	4	159	2	183	CAPI>Web	2.0	0.332
								CATI>Web	2.2	0.252
11-14	Financial decisions scale	42	134	53	161	39	183	CAPI>Web	1.1	0.686
								CATI>Web	1.7	0.211
15-18	Sensitive scale	36	146	35	153	28	166	CAPI>Web	1.4	0.096
								CATI>Web	1.9	0.092

[a] Control variables: sex, age, ethnicity, marital status and labor force status
[*] $p \leq 0.035$ (Šidák adjusted α for 3 tests for CAPI versus Web and 3 tests for CATI versus Web)

4.2 Results of Interviewer Effects on Social Desirability

Hypothesis 5: Interviewer effects on social desirability. We examined the impact of interviewer presence on socially desirable responding using sensitive agree/disagree statements about people with mental health problems and former prisoners living in the respondent's neighborhood (Table 2: questions 15–18).

The biggest differences were for CAPI and/or CATI vs web, with significantly more socially desirable responding shown in three of four CAPI vs web comparisons and three of four CATI vs web comparisons (Table 9). Differences between the interviewer-administered and web modes occurred regardless of the direction of the statement, indicating that this was a separate phenomenon to acquiescence, as for two of the four sensitive statements the socially desirable response required disagreement with the statement. Overall, the results support the first part of Hypothesis 5.

We were also interested in whether there would be differences between CAPI and CATI. Although there are no significant differences, on questions 16–18, the pattern of results was in the direction of more social desirability bias for CATI than CAPI,[24] offering some support for this secondary hypothesis.

5 Discussion

5.1 Review of Our Findings

These results illustrate how the presence or absence of interviewers may influence survey responses, contributing to measurement differences between self-administered modes (such as web) and interviewer-administered modes. We had hypothesized that on difficult non-sensitive questions, the presence of an interviewer would reduce satisficing behaviour. As seen in Figure 1 and discussed in Sections 2.1, there are many reasons why this may occur. For example, the nonverbal behavior of the interviewer can show professionalism and encourage respondent accountability. Similarly the smiling and nodding of face-to-face interviewers and the "mmmm's" and "uhuh's" of telephone interviewers can motivate respondents. Face to face interviewers, in particular, can reduce task difficulty by improving the reporting situation by reducing respondent distractions and reducing time pressure. Both face-to-face and telephone interviewers can also reduce task difficulty by reducing the cognitive demands of answering the survey question by clarifying aspects of the task, offering standard definitions and answering queries. As shown in Figure 1, there is an elaboration of Krosnick's 1991 model on satisficing (task difficulty divided by the product of respondent ability and motivation), and satisficing in turn influences how the respondent completes the four cognitive steps (outlined in the model by Tourangeau et al., 2000). In contrast, we hypothesized that on sensitive questions, the presence of an interviewer would reduce the privacy of the reporting situation, which could have an impact on respondents' willingness to disclose truthful answers but this would be tempered by the greater rapport in face-to-face interviewing.

There was evidence that the presence of an interviewer did help respondents carry out complicated tasks: CAPI respondents were more likely than web respondents to complete ranking tasks correctly (Hypothesis 1). There was also some evidence that interviewers motivated respondents to fully consider a question and the response options, reducing the extent of satisficing: respondents in the interviewer-administered modes were less likely to select middle response categories than web respondents (Hypothesis 3).

With respect to acquiescence, there were few significant differences by mode. The two questions that were significant showed greater acquiescence in interviewer-administered modes. This provides some support that acquiescence is related to interviewer presence rather than satisficing (Hypoth-

[24] Item 17 sig at $p < 0.10$ level before Šidák adjustment.

Table 6
Mode differences in middle category selection

		CAPI		CATI		Web		Logistic Regression results[a]		
ID	Dependent variables	%	n	%	n	%	n	comparison	OR	*p*-value
7	Neighborhood not a bad place (5 categories, agree/disagree)	4	282	7	314	10	349	Web>CAPI Web>CATI	2.7* 2.0	0.008 0.053
8	More properties in bad state of repair (5 categories, agree/disagree)	10	134	10	159	14	183	Web>CAPI Web>CATI	1.6 1.9	0.183 0.086
9	Neighborhood does not suffer from litter, dog mess, graffiti (5 categories, agree/disagree)	14	282	15	314	21	349	Web>CAPI Web>CATI	1.7 1.5	0.013 0.210
10	More properties are well kept (5 categories, agree/disagree)	17	135	19	161	16	183	CAPI>Web CATI>Web	1.0 1.1	0.997 0.829
11	Rarely read the small print before a financial decision (5 categories, agree/disagree)	6	281	7	314	14	349	Web>CAPI Web>CATI	2.4* 2.2*	0.002 0.004
12	Would do a lot of research before financial decision (5 categories, agree/disagree)	4	282	5	314	13	349	Web>CAPI Web>CATI	4.2 3.1*	0.000 0.000
13	Rarely talk to financial advisor before financial decision (5 categories, agree/disagree)	9	281	12	314	17	349	Web>CAPI Web>CATI	2.0* 1.2	0.006 0.505
14	Would talk to friends and family before financial decision (5 categories, agree/disagree)	10	134	12	161	15	183	Web>CAPI Web>CATI	1.4 1.2	0.399 0.657
15	Would worry if people with mental health problems lived in neighbourhood (5 categories, agree/disagree)	28	282	39	313	32	349	Web>CAPI CATI>Web	1.2 1.4	0.306 0.152
16	People with mental health problems have just as much right to live in neighbourhood (5 categories, agree/disagree)	18	281	16	313	31	349	Web>CAPI Web>CATI	2.0* 2.4*	0.002 0.000
17	Would worry if former prisoners lived in neighbourhood (5 categories, agree/disagree)	17	147	24	153	20	166	Web>CAPI CATI>Web	1.3 1.4	0.380 0.392
18	Former prisoners have just as much right to live in neighbourhood (5 categories, agree/disagree)	31	146	27	153	35	166	Web>CAPI Web>CATI	1.3 1.5	0.332 0.162

[a] Control variables: sex, age, ethnicity, marital status and labor force status
* $p \leq 0.009$ (Šidák adjusted α for 12 tests for CAPI versus Web and 12 tests for CATI versus Web)

esis 2).

We found little evidence of traditional primacy and recency effects (Hypothesis 4). Although primacy and recency are widely known and researched, according to Dillman et al. (1995, p. 674) "the prevalence of primacy and recency effect has been over-estimated".

For Hypothesis 5, our study examined the impact of interviewer presence on respondents' answers to sensitive statements. In contrast to the mixed pattern of results for satisficing behavior between modes, the evidence clearly showed that more socially desirable answers were provided in the interviewer-administered modes compared to web. There was weak evidence that more social desirability bias is present in CATI rather than CAPI interviews.

Table 7
Mode differences in primacy effects between CAPI with showcard and web

			CAPI		Web		Logistic Regression results[a]		
ID	Dependent variables	Categories	%	n	%	n	comparison	OR	*p*-value
1	Satisfaction with street cleaning	7	8	79	17	166	Web>CAPI	2.7	0.045
2	Satisfaction with waste and recycling collection	7	26	80	31	166	Web>CAPI	1.3	0.428
3	Length lived in area	7	1	70	4	183	Web>CAPI	6.7	0.137
4	Amount spent on leisure activities	8	24	70	30	183	Web>CAPI	1.6	0.251
5	Type of dwelling	8	30	67	27	166	CAPI>Web	1.3	0.415
6	Locations nearest to home	7	67	70	62	183	CAPI>Web	1.2	0.485

[a] Control variables: sex, age, ethnicity, marital status and labor force status
[*] $p \leq 0.017$ (Šidák adjusted α for 6 tests for CAPI versus Web)

Table 8
Mode differences between CAPI (no showcard)/CATI recency effects and Web primacy effects

				CAPI		CATI		Web		Logistic Regression results[b]		
ID	Dependent variables	Categ.	Type[a]	%	n	%	n	%	n	comparison	OR	*p*-value
1	Satisfaction with street cleaning	7	Primacy	14	66	24	153	17	166	Web>CAPI	1.3	0.490
										CATI>Web	1.5	0.171
			Recency	8	66	6	153	3	166	CAPI>Web	3.4	0.074
										CATI>Web	1.9	0.257
2	Satisfaction with waste and recycling collection	7	Primacy	25	67	35	153	31	166	Web>CAPI	1.7	0.133
										CATI>Web	1.1	0.590
			Recency	8	67	5	153	2	166	CAPI>Web	3.6	0.089
										CATI>Web	2.5	0.191
3	Length lived in area	7	Primacy	0	65	4	161	4	183	Web>CAPI	0.0	0.997
										Web>CATI	2.2	0.216
			Recency	37	65	35	161	39	183	Web>CAPI	1.1	0.807
										Web>CATI	1.0	0.938
4	Amount spent on leisure activities	8	Primacy	40	65	21	160	30	183	CAPI>Web	1.5	0.168
										Web>CATI	1.7	0.087
			Recency	5	65	11	160	3	183	CAPI>Web	1.9	0.392
										CATI>Web	5.9[*]	0.001
5	Type of dwelling	8	Primacy	31	80	21	153	26	166	CAPI>Web	1.4	0.297
										Web>CATI	1.1	0.632
			Recency	0	80	1	153	2	166	Web>CAPI	0.0	0.996
										Web>CATI	2.2	0.550
6	Locations nearest to home	7	Primacy	68	65	67	161	62	183	CAPI>Web	1.4	0.225
										CATI>Web	1.2	0.372
			Recency	2	65	3	161	3	183	Web>CAPI	0.0	0.999
										Web>CATI	0.0	0.999

[a] Whether primacy or recency effects were explored [b] Control variables: sex, age, ethnicity, marital status and labor force status
[*] $p \leq 0.009$ (Šidák adjusted α for 12 tests for CAPI versus Web)

Table 9
Mode differences in selection of the social desirability answer – 'strongly agree' and 'agree' to the sensitive questions

		CAPI		CATI		Web		Logistic Regression results[a]		
ID	Dependent variables and direction of socially desirable answer	%	n	%	n	%	n	comparison	OR	*p*-value
15	Would worry if people with mental health problems lived in neighbourhood (Lower percentage)	38	282	36	313	53	349	Web>CAPI Web>CATI	1.8* 2.0*	0.001 0.000
16	People with mental health problems have just as much right to live in neighbourhood (Higher percentage)	65	281	69	313	44	349	CAPI>Web CATI>Web	2.3* 2.9*	0.000 0.000
17	Would worry if former prisoners lived in neighbourhood (Lower percentage)	70	147	60	153	73	166	Web>CAPI Web>CATI	1.1 1.8	0.784 0.049
18	Former prisoners have just as much right to live in neighbourhood (Higher percentage)	46	146	51	153	33	166	CAPI>Web CATI>Web	1.8* 2.1*	0.012 0.002

[a] Control variables: sex, age, ethnicity, marital status and labor force status
* $p \leq 0.026$ (Šidák adjusted α for 12 tests for CAPI versus Web and 4 tests for CATI versus Web)

We also identified an unexpected finding, with patterns of satisficing differing by question format as well as by mode. For example, interview respondents were more likely to acquiesce but web respondents were more likely to choose a middle category on questions in the agree/disagree format.

5.2 Strengths, Limitations and Conclusions

Our study has several strengths compared to other published research on mixed modes, outlined in Section 1 on contributions to the extant literature. This includes a substantive focus on measurement error as opposed to nonresponse or noncoverage biases, an exploration of the presence or absence of the interviewer beyond social desirability effects, a fully randomized experimental design of three modes (CAPI, CATI and web), a sample of the adult population of Great Britain as opposed to students or other more homogeneous populations, a probability sample of those who have access to the internet rather than an online opt-in panel or purposive sample, analyses that accounted for the correlated variance created by interviewers, and a focus on the difficulty of question format rather than wording. Our findings suggest that the difficulty of question format should be taken into consideration in mixed modes research that includes both interview and self-administered modes. This is particularly pertinent because we found that respondents may satisfice differently according to mode.

There are also some limitations of our study. First, although our sample was drawn from the population of Great Britain, there were three levels of nonresponse (to the original Omnibus survey, to being re-contacted, to the mixed modes experiment). Thus, overall response rates are low (for example, for the CAPI mode this resulted in 54% x 83% x 73% = 33% of the original sample taking part in the experiment). As a consequence, participants in the experiment may differ from the population as a whole, in particular, by having been more cooperative. This would suggest they would be less likely to satisfice. Nevertheless, the results do provide evidence for satisficing. Since the focus of the experiment was a comparison of mode differences between the CAPI/CATI and web, the first two levels of nonresponse are not relevant. To correct for the differential nonresponse in the experiment itself, we controlled for socio-demographic variables that accounted for differences between respondents in each mode. Despite the three levels of nonresponse and the random sub-selection of respondents to then be randomly allocated to mode, this was still a probability design, because a random sample of a random sample is still a random sample (Blair & Blair, 2015). And "probability samples, even ones without especially high response rates, yielded quite accurate results" (Yeager et al., 2011, p. 737).

A second limitation was the use of pre-existing survey questions from established surveys without a further pretest. Nevertheless, the questions did reflect those that are included in real social surveys.

In conclusion, both expected and unexpected findings make a valuable contribution to the research literature on measurement error in mixed mode survey designs. We have provided evidence that interviewers can motivate and help respondents while at the same time that their presence may lead respondents to give socially desirable answers. We have

also shown that findings vary by question format, and that, even on the same question, satisficing may manifest differently between modes and question formats.

While survey cost pressures encourage increasing use of mixed mode designs, these results highlight that some questions are less effective across modes. Before deciding to use mixed mode designs therefore, practitioners should consider the potential impact of question format and sensitivity on answers provided, and include pre-testing in different modes where this is possible.

6 Acknowledgements

We wish to acknowledge the funding from the Economic and Social Research Council [grant number RES-175-25-0007] which made this research possible. We are grateful to Alita Nandi (ISER), David Hussey (NatCen Social Research) and Rebecca Taylor, Margaret Blake, Michelle Gray, Chloë Robinson (who were at NatCen Social Research at the time the experiment was implemented) for their contributions to the design, management and analysis. We are also grateful to NatCen Social Research interviewers and operations staff for collecting and processing the data. Finally, we would like to thank all those members of the public who gave their time and co-operation in responding to the surveys.

References

Abdi, H. (2007). The Bonferonni and šidák corrections for multiple comparisons. The University of Texas at Dallas. Retrieved from https://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf

Alwin, D., & Krosnick, J. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, *49*(4), 535–552.

American Association for Public Opinion Research. (2016). Standard definitions: Final dispositions of case codes and outcome rates for surveys. Retrieved from https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Baker, R., Blumberg, S., Brick, M., Couper, M., Courtright, M., Dennis, J., & Zahs, D. (2010). AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711–781.

Beste, J., Sakshaug, J., & Trappmann, M. (2021). Effects of a CAPI-CATI mixed-mode panel design on non-response, measurement error and total survey error. *Survey Research Methods*, *14*(2), 235–239. Paper presented at the Digital expert discussion: The future of the German mikrozensus. Retrieved from https://dstatg.de/wpcontent/uploads/2021/07/Expertengespraech-Tagungsdokumentation.pdf

Betts, P., & Lound, C. (2010a). *The application of alternative modes of data collection in UK government social surveys: A report for the Government Statistical Service*. London: Office for National Statistics.

Betts, P., & Lound, C. (2010b). *The application of alternative modes of data collection in UK government social surveys: Review of literature and consultation with national statistical institutes*. London: Office for National Statistics.

Blair, E., & Blair, J. (2015). *Applied survey sampling*. Thousand Oaks, CA: Sage.

Blanke, K., Luiten, A., Betts, P., & Lound, C. (2014). *Query on data collection for social surveys: ESSnet project data collection for social surveys using multiple modes.* Luxembourg: Eurostat. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/Query_report_DCSS.pdf_en

Burton, J., Lynn, P., & Benzeval, M. (2020). How understanding society: The uk household longitudinal study adapted to the COVID-19 pandemic. *Survey Research Methods*, *14*(2), 235–239.

Campanelli, P., Blake, M., Mackie, M., & Hope, S. (2015). *Mixed modes and measurement error: Using cognitive interviewing to explore the results of a mixed modes experiment*. Institute for Social and Economic Research Working Paper Series (2015-18), University of Essex.

Cernat, A., & Revilla, M. (2020). Moving from face-to-face to a web panel: Impacts on measurement quality. *Journal of Survey Statistics and Methodology*, *9*(4), 1–19.

Chang, L., & Krosnick, J. (2010). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *74*(4), 641–678.

Couper, M. (2012). *Advantages and disadvantages of internet survey methods for official statistics*. 4th International Workshop on Internet Survey Methods, Daejeon, South Korea.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity*, *52*(4), 1523–1559.

de Leeuw, E. (1992). *Data quality in mail, telephone, and face-to-face surveys*. Amsterdam: TT Publications.

de Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2), 233–255.

DeVellis, R. (2016). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: Sage.

Dillman, D., Brown, T., Carlson, J., Carpenter, E., Lorenz, F., Mason, R., & Songster, R. (1995). Effects of category order on answers to mail and telephone surveys. *Rural Sociology*, *60*(4), 674–687.

Dillman, D., Smyth, J., & Christian, L. (2007). *Internet, mail and mixed-mode surveys: The tailored design method* (2nd ed.). Hoboken, NJ: Wiley.

Dillman, D., Smyth, J., & Christian, L. (2014). Aural versus visual design of questions and questionnaires. In D. Dillman, J. Smyth, & L. Christian (Eds.), *Internet, phone, mail and mixed-mode surveys: The tailored design method* (4th ed., pp. 169–225). Hoboken, NJ: Wiley.

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2006). Comparing data from online and face-to-face surveys. *International Journal of Market Research*, *47*(6), 615–639.

Erens, B., Burkill, S., Couper, M., Conrad, F., Clifton, S., Tanton, C., & Copas, A. (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: A comparison with a probability sample interview survey. *Journal of Medical Internet Research*, *16*(12), 276.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage.

Fowler, F. J., & Mangione, T. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Thousand Oaks, CA: Sage.

Grandjean, B., Nelson, N., & Taylor, P. (2009). *Comparing an internet panel survey to mail and phone surveys on willingness to pay for environmental quality: A national mode test*. Proceedings 2009 annual conference of the American Association for Public Opinion Research, Hollywood, Florida.

Groves, R. (1989). *Survey errors and survey costs*. Hoboken, New Jersey: Wiley.

Groves, R., & Kahn, R. (1979). *Surveys by telephone: A national comparison with personal interview*. Cambridge, Massachusetts: Academic Press.

Guyatt, G., Haynes, R., Jaeschke, R., Cook, D., Green, L., Naylor, C., & Richardson, W. (2000). Users' guides to the medical literature: XXV. evidence-based medicine: Principles for applying the users' guides to patient care, evidence-based medicine working group. *Journal of the American Medical Associate*, *284*(10), 1290–1296.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*(1), 111–121.

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846.

Holbrook, A. (2008). Acquiescence response bias. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage.

Holbrook, A., Green, M., & Krosnick, J. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*(1), 79–125.

Jäckle, A., Lynn, P., & Burton, J. (2015). Going online with a face-to-face household panel: Effects of a mixed mode design on item and unit non-response. *Survey Research Methods*, *9*(1), 57–70.

Jäckle, A., Roberts, C., & Lynn, P. (2006). *Telephone versus face-to-face interviewing: Mode effects on data quality and likely causes*. Institute for Social and Economic Research Working Paper Series (2006-41). University of Essex.

Jäckle, A., Roberts, C., & Lynn, P. (2008). *Assessing the effect of data collection mode on measurement*. Institute for Social and Economic Research Working Paper Series (2008-08). University of Essex.

Jackman, M. (1973). Education and prejudice or education and response set? *American Sociological Review*, *38*(3), 327–339.

Javeline, D. (1999). Respond effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly*, *63*(1), 1–28.

Kish, L. (1965). *Survey sampling*. Hoboken, New Jersey: Wiley.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in cati, ivr, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J., & Presser, S. (2010). Questionnaire design. In J. Wright & P. Marsden (Eds.), *Handbook of survey research* (2nd ed.). San Diego, CA: Elsevier.

Laaksonen, S., & Heiskanen, M. (2013). *Comparison of three survey modes*. Department of Social Research Working Paper 2. University of Helsinki.

Landsberger, H., & Saavedra, A. (1967). Response set in developing countries. *Public Opinion Quarterly*, *31*(2), 214–229.

Lenski, G., & Leggett, J. (1960). Caste, class, and deference in the research interview. *American Journal of Sociology*, *65*(5), 463–467.

McBride, L., & Moran, G. (1967). Double agreement as a function of ambiguity and susceptibility to demand implications of the psychological situation. *Journal of Personality and Social Psychology*, *6*(1), 115–118.

NatCen Social Research. (2014). Mixed modes and measurement error, 2009. doi:10.5255/UKDA-SN-7515-1

Office of Communications. (2018). A decade of digital dependency. Retrieved from https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/decade-of-digital-dependency

Ofstedal, M., McClain, C., & Couper, M. (2021). Measuring cognition in a mulit-mode context. In P. Lynn (Ed.), *Advances in longitudinal survey methodology* (pp. 250–271). Hoboken, NJ: Wiley.

Peabody, D. (1966). Authoritarianism scales and response bias. *Psychological Bulletin*, *65*(1), 11–23.

Reja, U., Lozar Manfreda, K., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. In A. Ferligoj & A. Mrvar (Eds.), *Developments in applied statistics* (pp. 159–177). Ljubljana: Fakulteta za družbene vede.

Revilla, M., Saris, W., Loewe, G., & Ochoa, C. (2015). Can a non-probabilistic online panel achieve question quality similar to that of the European Social Survey? *International Journal of Market Research*, *57*(3), 395–412.

Reynolds, T., & Jolly, J. (1980). Measuring personal values: An evaluation of alternative methods. *Journal of Marketing Research*, *17*(4), 531–536.

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, *83*(3), 598–626.

Saris, W., Revilla, M., Krosnick, J., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 61–79.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel. *Bulletin de Méthodologie Sociologique*, *109*, 56–61.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Cambridge, Massachusetts: Academic Press.

Schuman, H., Steeth, C., & Bobo, L. (1985). *Racial attitudes in America: Trends and interpretations*. Cambridge, Massachusetts: Harvard University Press.

Smyth, J., Christian, L., & Dillman, D. (2008). Does 'yes or no' on the telephone mean the same as 'check-all-that-apply' on the web? *Public Opinion Quarterly*, *72*(1), 103–113.

Taylor, J., & Kinnear, T. (1971). Empirical comparison of alternative methods for collecting proximity judgments. *Bureau of Business Research Working Paper*, (46).

Thomas, S., & Heck, R. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education*, *42*(5), 517–540.

Tourangeau, R., Conrad, F., & Couper, M. (2013). *The science of web surveys*. Oxford: Oxford University Press.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

U.K. Cabinet Office Consultation. (2013). Community life survey: Development of content and methodology for future survey years. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/

U.S. Census Bureau. (2010). Design and methodology: American Community Survey. Retrieved from https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/

U.S. Census Bureau. (2020). People and households represented in each American Community Survey data collection mode. Retrieved from https://www.census.gov/library/visualizations/interactive/acs-collection.html

Warnecke, R., Johnson, T., Chavez, N., Sudman, S., O'Rourke, D., Lacey, L., & Horm, J. (1997). Improving question wording in a survey of culturally diverse populations. *Annals of Epidemiology*, *7*(5), 334–342.

Weems, G. (2004). Impact of the number of response categories on frequency scales. *Research in the Schools*, *11*(1), 41–49.

Yeager, D., Krosnick, J., Chang, L., Javatz, H., Levendusky, M., Simpser, A., & Wang, R. (2011). Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709–747.

Zhang, C., & Conrad, F. (2013). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135.