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

Pamela Campanelli
The Survey Coach

Gerry Nicolaas
NatCen Social Research

Peter Lynn
Institute for Social and Economic Research
University of Essex

Annette Jäckle
Institute for Social and Economic Research
University of Essex

No. 20141-20
April 2014

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Non-technical summary

This paper focuses on the role of the interviewer in causing mode effects, contrasting modes in which an interviewer is used (face-to-face and telephone) with a mode without an interviewer (web). Other differences between modes, such as aural versus visual transmission of information, are held constant where possible.

The presence of an interviewer is hypothesised to motivate respondents to generate an accurate answer and to reduce the difficulty of the task by offering support and providing explanations of what is needed. However, an interviewer will reduce the privacy of the reporting situation, which may have an impact on respondents' willingness to answer truthfully.

Based on an UK ESRC-funded mixed modes experiment, this paper compares (1) the prevalence of indicators of satisficing (e.g., non-differentiation, acquiescence, middle categories, primacy and recency, and item nonresponse) and (2) the prevalence of socially desirable responding between interviewer and self-completion modes. Results provide evidence that interviewers do motivate respondents, with fewer middle category endorsements in the interviewer modes than web, and that interviewers do help respondents, with fewer errors of duplication and non-differentiation in ranking tasks in the interviewer modes than web. We found clear differences by mode on agree/disagree questions, with acquiescence more prevalent in the interviewer modes than web. This suggests that acquiescence is due to another cause than satisficing. Acquiescence due to socially desirable responding was also ruled out with a scale of sensitive questions, as respondents clearly gave different views for positive and negative statements. Socially desirable responding occurred, as expected, between interview modes and web (although social desirability did not differ between the interviewer modes). There was little evidence for standard primacy and recency effects and little or no missing data across all the modes, but there was some evidence of an unexpected primacy/positivity bias in the telephone interview mode.

One noteworthy result was the different manifestation of satisficing behaviour between mode and item format. For example, on agree/disagree items respondents in the interview modes were more likely to show acquiescence bias, but respondents completing the web survey were more likely to choose the middle category. Comparative results from a cognitive interviewing follow-up study supports some of the empirical findings, but also leads us to question some of the indicators frequently used in the literature to test for acquiescence effects.

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, UCL Institute of Child Health
Pamela Campanelli, The Survey Coach
Gerry Nicolaas, NatCen Social Research
Peter Lynn, Institute for Social and Economic Research
Annette Jäckle, Institute for Social and Economic Research

Abstract

The presence of an interviewer is hypothesised to motivate respondents to generate an accurate answer and reduce task difficulty, but also to reduce the privacy of the reporting situation.

The prevalence of indicators of satisficing (e.g., non-differentiation, acquiescence, middle categories, primacy and recency, and item nonresponse) and socially desirable responding were studied experimentally across modes and also through cognitive interviewing. Results show differences between interviewer and self-completion modes: in levels of satisficing for non-differentiation, acquiescence, and middle categories and socially desirable responding. There were also unexpected findings of a CATI primacy/positivity bias and of different ways of satisficing.

Keywords: Mode Effects, Interviewer presence, Satisficing, Primacy, Recency, Middle category effects, Non-differentiation, Social desirability.

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 Rebecca Taylor, Margaret Blake, Michelle Gray and Chloë Robinson, David Hussey from NatCen Social Research and Alita Nandi from ISER 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.

Contact: Steven Hope, Centre for Policy Research, Population, Policy and Practice Programme, UCL Institute of Child Health, 30 Guilford Street London WC1N 1EH. Email: s.hope@ucl.ac.uk.

1 Introduction

In 2003, Biemer and Lyberg described mixed-modes as the 'norm' for survey design, and combining modes of data collection in surveys remains a topical issue. The interest in a mixed modes approach is driven in part by a desire to improve survey response rates at a time when rates for unimode surveys are declining, and in part by the need to reduce fieldwork costs. This latter reason is particularly true in the United Kingdom where face-to-face interviewing is the predominant mode of data collection for large scale national social surveys (Betts and Lound, 2010). Sequential use of different modes can encourage sample members to participate who would otherwise have failed to do so, thereby increasing response (Millar and Dillman, 2011), and if a sufficient number of respondents use less expensive modes, this should reduce the overall costs of data collection.

However, modes differ in terms of access (coverage error), non-response bias (non-response error) and responses obtained (measurement error), which may lead to non-equivalence between data collected in different modes (see de Leeuw, 2005) and lessen the apparent advantages of mixing modes. This paper focuses on non-equivalence between modes due to measurement error. We investigate the role of the interviewer in causing differences in measurement between modes, contrasting modes in which an interviewer is present (face-to-face and telephone) with self-completion (web).

A major role of interviewers is motivation, encouraging respondents to make sufficient effort in processing the survey item, and reducing task difficulty by offering support and additional explanations of what is needed. In this context, the role of the interviewer will be more important if the survey task is difficult. Task difficulty and motivation are two of the main causes of satisficing, or giving a less than optimal response (Krosnick, 1991, 1999).

Satisficing has been argued to be at the root of many response effects in surveys, such as primacy and recency, acquiescence, no opinion responses, problems with rating tasks, and middle category selection (Krosnick, 2000; Krosnick and Fabrigar, 1997). Krosnick (1991) and Narayan and Krosnick (1996) make a distinction between 'weak' and 'strong' satisficing. In weak satisficing, respondents take shortcuts but do not abandon any of the major response processes, for example selecting the first response option that constitutes a reasonable answer or agreeing with items that make an assertion. In strong satisficing, respondents miss out whole components of the response process (such as the retrieval stage), for example selecting 'don't know' when an answer is known; choosing the middle category or neutral response on an attitude scale when an opinion is held;

selecting the same response for every item (non-differentiation or 'straight-lining') , or answering randomly.

Interviewers also reduce the privacy of the reporting situation, which may have an impact on respondents' 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. This would be moderated by rapport which is better in face-to-face interviewing in contrast to the greater social distance in telephone interviewing. The literature on this topic is somewhat mixed as to which method of data collection yields more truthful answers. There are papers showing no difference between face-to-face and telephone interviewing on sensitive questions (Aneshensel et al, 1982, Feldman-Naim et al, 1997), face-to-face being better than telephone (e.g., Aquilino, 1994, Johnson, Hougland and Clayton, 1989) and telephone being better than face-to-face (e.g., Pless and Miller, 1979, Sykes and Collins, 1988). But the majority of the literature suggests that either face-to-face is better or there is no difference.

The extent to which the interviewer affects responses, and hence contributes to differences between modes in measurement, will depend on the characteristics of the item, such as sensitivity, visual layout, and difficulty due to item format and/or wording.[1] Each of these will be discussed in turn.

Item sensitivity reflects whether or not the survey items used are likely to be seen as sensitive to respondents. There is a large body of evidence that shows respondents are more likely to give socially desirable answers when asked questions by an interviewer than when completing a questionnaire on their own, even if it is a self-completion questionnaire within an interview survey (evidence reviewed by Tourangeau, Rips and Rasinski, 2000; see also Tourangeau and Yan, 2007 and Kreuter, Presser and Tourangeau, 2008).

Aural versus visual presentation reflects whether the respondent hears the item or sees it. The research evidence suggests that there are important differences, and that respondents find the task of answering an item easier when it is are presented visually (see Dillman, Smyth and Christian, 2009

[1]Sensitivity, visual layout, and item difficulty have been labelled as characteristics of the item. But these obviously interact with the characteristics of the respondents (for example, sensitive items are not sensitive to those who do not exhibit the sensitive behaviour or attitude; and poor visual layout may affect some respondents but not others).

for a review; also Couper, Traugott and Lamias, 2001; Tourangeau, Couper, Conrad, 2004; Ciochetto, Murphy, and Agarwal, 2006).

Task difficulty due to item format is usually only discussed in the context of mode specific formats (see Christian, Dillman and Smyth, 2008). For example, tick-all-that-apply cannot be administered over the telephone and instead a list of 'yes/no' statements is often used instead. But at the same time, there is evidence that certain item formats are intrinsically more difficult for respondents to complete than others:

- A ranking task is more difficult than a series of rating items. A ranking task requires the respondent to understand the nature of ranking a list of options and 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 to rank (Alwin and Krosnick, 1985; Rokeach, 1973; Feather, 1973, Fowler, 1995). Another indicator of difficulty is that ranking takes longer to complete than rating (McIntyre and Ryans, 1977; Reynolds and Jolly, 1980; Taylor and Kinnear, 1971).
- Agree /disagree statements are more difficult than comparable rating scales: the formulation of agree/disagree statements is not intuitive, as identified in the following example: "*disagreeing that one is seldom depressed is a complicated way of saying one is often depressed*" (Fowler, 1995: p. 56). In addition, the standard 5-point agree/disagree scale (strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree) contains two dimensions: respondents' attitudinal position (i.e., agree or disagree) and the intensity of their feeling (e.g., strongly agree versus agree; Fowler, 1995).[2]
- Long scales are more difficult than short scales. The cognitive complexity of the task increases with the number of scale points (Fowler, 1995; Tourangeau, Rips and Raskinski, 2000). Fowler (1995) suggests "*5 to 7 categories is probably as many categories as most respondents can use meaningfully for most rating tasks*" (Fowler, 1995: p. 53).
- End-labelled scales are more difficult than fully-labelled scales; a label on every scale point clarifies the meaning of each point (Peters & McCormick 1966; Krosnick and Berent, 1993; Krosnick and Fabrigar, 1997).

[2] In addition, the agree/disagree format has been found to be less reliable and valid than comparable rating scales (Saris et al, 2010). As discussed later, the agree/disagree format is also prone to acquiescence bias (agreeing to a statement regardless of its content).

Difficulty due to item wording (sometimes called inherent difficulty) relates to survey items with unfamiliar concepts, ambiguous wording or extensive recall tasks. For example, respondents may proceed with their best guess of what the survey item means or make do with using only what is easily accessible from their memory (Tourangeau, Rips and Raskinski, 2000).[3]

This paper makes several contributions to the extensive literature on mode effects. Firstly, we exploited a unique combination of experimental data from a general population sample. Secondly, cognitive interviews were used after the survey experiment to shed light on unusual quantitative findings and explore possible causes of mode differences in responses. Thirdly, the experiment was designed to assess the impact of interviewers on items with the different formats outlined above. Note that item sensitivity, task difficulty due to format and aural versus visual transmission of information were held constant across mode, where possible.

Section 2 sets out the hypotheses we tested; Section 3 documents the methodology used for the experimental survey and the cognitive interviews; Section 4 presents the results and Section 5 provides a discussion and conclusion.

2. HYPOTHESES

Based on the review above, we hypothesised the interviewer would have an impact both on the extent of satisficing and social desirability bias, leading to potential differences in measurement between modes.

We expected interviewers to motivate respondents to undertake more difficult tasks due to item format; and if the task was undertaken, we expected interviewers to help respondents understand how to carry out the task properly. We therefore predicted that more satisficing behaviour would be exhibited by web questionnaire respondents[4] compared to interviewed respondents (CAPI – computer-assisted personal interviewing, or CATI – computer-assisted telephone interviewing). More specifically, we hypothesised that the following satisficing behaviours would be less likely to occur when an interviewer was present:

[3] Although the research team designed the experiment to include a contrast of easy versus difficult item wording (inherent difficulty), this aspect is not formally included in this paper. This is because what the researcher judges as a difficult item may not match a respondent's perception of item difficulty (see Hunt, Sparkman and Wilcox, 1982; Sangster and Fox, 2000; Nicolaas et al, 2011)

[4] Sometimes called CAWI (computer-assisted web interview).

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

2. *Acquiescence response bias on agree/disagree items*[5]*:* Without the presence of an interviewer, respondents may satisfice by simply agreeing to statements, regardless of their content. We therefore expected web respondents to be more likely than respondents in interviewer administered modes to satisfice by agreeing to statements.

3. *Middle category satisficing:* Without the motivation of an interviewer, respondents may satisfice by choosing the middle option when in fact they do have a positive or negative opinion. We therefore expected that respondents would be more likely to satisfice by selecting the middle category in self-completion than interviewer administered modes.

4. *Primacy and recency effects on items with 5 or more categories:* Without the motivation of the interviewer, we firstly expected 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. Secondly, we expected the extent of primacy effects in the visual web mode to be larger than the extent of recency effects in the aural interviewer administered modes.

5. *Item non-response in items with difficult formats:* Without the motivation and help of the interviewer on difficult item formats (in this experiment, ranking and end-labelled scales[6]), we expected more missing data in web than in the interviewer administered modes.

---

[5] Acquiescence response bias is defined as the propensity to agree to a statement regardless of its content. Although acquiescence falls under Krosnick's list of satisficing behaviours, there are alternative explanations: ambiguity of the agree/disagree statement itself (see Peabody, 1966; McBride and Moran, 1967), characteristics of the respondent (e.g., less educated - see Schuman and Presser, 1981; Landsberger and Saavedra, 1967), deference to the interviewer (see Carr, 1971; and Lenski and Leggett, 1960; Javeline, 1999) and category fallacy (i.e, choosing a 'safe' category because of a concern about looking foolish or ignorant) (see Warnecke, et al, 1997; Jackman, 1973).

[6] We specifically excluded the scales using the agree/disagree format and long response scales. Although these formats are considered difficult, we expected satisficing behaviour to show up in a

We further hypothesised the interviewer to have a bearing on the respondent's willingness to disclose potentially sensitive information.

Social desirability: We expected a pattern of results indicative of socially desirable responding to be more prevalent in the two interview modes than web, as the presence of the interviewer reduces privacy of reporting.

3. METHODS

3.1 Experimental design

The mixed modes experiment was conducted on a follow-up sample of respondents from two waves of the 2008 NatCen quarterly Omnibus survey and from the 2008-09 wave of the British Household Panel Study (BHPS). Those who agreed to be re-contacted, were randomly allocated to one of three modes for the Omnibus sample (CAPI, CATI and web questionnaire) and for the BHPS either to CATI or web. The CAPI component for the BHPS sample was from the main BHPS survey (now subsumed under the UK Household Longitudinal Study)[7]. The web sample for both studies was restricted to respondents who had access to and used the internet. Although this restriction did not hold for CAPI and CATI respondents, the analyses for all modes proceeded only with respondents who had access to and used the internet.

In the overall project 15 items were taken from the BHPS and 67 other items were selected from other surveys, or newly designed to test our hypotheses. The twenty-four items relevant to this paper are listed in Appendix A, with some items used to test more than one hypothesis. The items included two ranking tasks (items 19 and 20) which were expected to show greater non-differentiation in web; twelve agree/disagree items thought to lead to more acquiescence bias in web (items 7 to 18); long scales (with 5 or more categories – items 1 to 24), thought to engender primacy effects with visual presentation and recency effects with oral presentation, with more order effects in web; and end-labelled scales (items 21 to 24) and ranking tasks (items 19 and 20), expected to show more missing data in self-completion. In addition, all of the subjective items with middle categories were analyses for 'middle category' satisficing (items 1, 2, 7-18, 21 and 22). One

---

different way than item nonresponse, i.e., acquiescence on agree/disagree items and primary/recency effects on long scales.

[7] The CAPI data from the UK Household Longitudinal Survey were not available at the time of the writing of this paper.

of the agree/disagree series was based on a sensitive topic to investigate the presence of more socially desirable answering in the interview modes (items 15-18).

The items were identical in each of the three modes, except where the item format was considered to be too difficult to administer in CATI and therefore not included (for example, ranking tasks) or where a split ballot design was used to test format differences within a mode (for example, the use or not of showcards in CAPI).

Table 1 documents the sample sizes and response rates for the different modes in the NatCen Omnibus and BHPS samples.

Table 1: Mixed mode experiment sample sizes

	NatCen Omnibus		BHPS	
	N	Response rate	N	Response rate
CAPI	282 ♦	78%	Not available	Not available
CATI	314 ♦	68%	421 ♦	70%
Web	349	47%	334	37%

♦ Excludes respondents without internet access/who do not use the internet.

3.2 Analysis methods

As documented in Table 1, response rates varied between modes. To adjust for the potential effects of differences in the characteristics of the responding samples between modes, we controlled for socioeconomic variables in all analyses.[8] Selection of the controls involved logistic regression models estimated simultaneously, forwards stepwise and backwards stepwise, with a mode pairing as the dependent variable and the independent variables comprising the socio-demographic variables available in the Omnibus dataset. This approach identified variables that differed between any two modes. If a variable was significant at $p<0.10$ in any of the logistic regression models, it was selected as a control. The control variables included in all subsequent analyses were: age, sex, ethnicity, economic status and marital status. The significant control variables for the BHPS data

[8] Adjustment for control variables was chosen in preference to standard weighting to the population or propensity score weighting as the most suitable approach for analyses, given that the comparisons involved three modes and the samples for the experiment were drawn from existing survey respondents rather than the general population.

were sex, age, and marital status, but ethnicity and economic status were added to replicate NatCen Omnibus survey controls.

To test the hypotheses, we estimated logistic regressions for the relevant indicators of satisficing or social desirability, including mode and the socio-demographic controls as explanatory variables. Because of the relatively small sample size for item formats within modes, findings at the $p<.10$ level are reported.

3.3 The cognitive interviewing methodology

Cognitive interviewing is traditionally thought of as a pretesting method. In contrast, we pre-planned a cognitive interviewing follow-up study, designed to gain a greater understanding of how mode effects happen, even if they were not directly observed, and to seek explanations for any unusual quantitative findings. Thirty seven respondents were recruited for the cognitive interviewing phase from respondents who had participated in the NatCen Omnibus mixed modes experiment. Although mode differences are typically detected at an aggregate level, we found that certain respondent 'satisficing' behaviours differed by mode (i.e., acquiescence through agreeing to opposite agree/disagree statements and non-differentiation in a ranking task). Specific quotas were set up to contrast respondents who had displayed satisficing versus those who had not.

The cognitive interview respondents first experienced a selection of survey questions from the mixed modes questionnaire in three modes of data collection (CAPI, CATI and CAWI) lasting about 10 minutes. (Note that all 37 respondents experienced all three modes). This was followed by the actual cognitive interview component in the form of retrospective think-alouds and pre-specified probes which lasted approximately 50 minutes. Sets of questions with a particular format (e.g., agree/disagree), were divided into two parts based on the results from the quantitative analysis with the goal of creating two equivalent groups of questions which could be used in different modes. This ensured that no respondent was asked the same question more than once. The cognitive interviews were carried out without reference to respondents' previous answers in the mixed mode experiment interview because at least 5 months had passed since that interview.

The survey questions were administered in standard quantitative fashion and mimicked as closely as possible in the three modes. This involved the interviewer sitting with the respondent face-to-face (for the CAPI component), being in a different room in the respondent's home and talking over a landline/mobile phone (for the CATI component) and having the respondent use the interviewer's

laptop completely on his/her own (for the CAWI component).[9] The retrospective think-alouds proceeded by reminding the respondent of the survey question, data collection mode, his or her answer and any behavior displayed whilst answering e.g., hesitation. The respondent then talked through how he or she had gone about answering the question and how he or she had decided on the answer. For some of the question format experiments, pre-scripted structured open probes were used to explore specific aspects of the response process such as non-differentiation in a rating task.

All cognitive interviews were then transcribed and the data introduced into the qualitative charting programme, "Framework", for analysis. The themes behind respondent's answers were explored as is typically done in qualitative analysis. The next level was to see if respondents' answering processes differed at all by mode. For a full description of the cognitive interviewing methodology used and some of its differences and innovations compared to standard cognitive interviewing, see Gray, Blake and Campanelli (2014).

4. RESULTS

4.1 INTERVIEWER EFFECTS ON SATISFICING

4.1.1 Duplicates and non-differentiation in ranking tasks – Hypothesis 1

To assess Hypothesis 1 (that the extent of duplicates and non-differentiation in ranking tasks would be greater in web than interviewer administered modes), we used two ranking tasks (items 19 and 20 in Appendix A).[10] The first ranking task asked which changes to the respondent's neighbourhood would be most important; the second task asked which geographical unit of the respondent's address (from street, city etc. to UK and European level) was most important to them. It was phrased in terms of an address game that children may play.[11]

[9] Which items were asked in which mode were varied across version of the protocol, but the mode order (CAPI, CATI, and web) remained constant.

[11] Non-differentiation can, in principle, be prevented in web surveys through program edits and error messages to respondents. The concern is that such measures may irritate respondents and lead to survey drop-out. For this study we purposely did not use any such program edits to facilitate comparison with CAPI (where there were also no programme edits). We wanted to understand the types and magnitude of errors that respondents make when responses are unconstrained. Also web without edits mirrors a paper self-completion form.

[11] The ranking tasks were carried out in CAPI and web but not in CATI as they would be impossible to complete without information provided visually. Consequently, only Omnibus data are reported here as the BHPS CAPI data were not yet available.

First we examined an indicator of duplicate ranks which took a value 1 if the respondent had assigned any of the options the same ranking, and 0 otherwise. In the Omnibus data, non-differentiation was significantly more prevalent among web respondents than CAPI respondents. For the children's address game item it was 18.0% in CAPI and 49.2% in web (OR=5.11, p<.001) and for the improvements to the neighbourhood item it was 16.3%in CAPI, 29.3% in web (OR=2.22, p<.01).

Second we focused on an extreme form of duplicate ranking, non-differentiation, where the indicator took value 1 if the respondent had picked the same ranking throughout or the same ranking for all but one of the items, and 0 otherwise. The significantly greater level of non-differentiation in web was confirmed for the children's address game item (5.2% in CAPI and 13.7% in web (2.91, p<.05). Although a similar pattern of non-differentiation was seen for the improvements to the neighbourhood item (0.7% in CAPI and 3.6% in web), the difference did not reach significance at the 10% level.

Third, we used cognitive interviewing to investigate how respondents understood the ranking task in web using the children's address game item (item 20 in Appendix A). None of the six respondents who were assigned the ranking task did the task correctly. Most had duplicate ranks and the others ranked the first item as 1 and left the rest blank. Both the task of ranking and the nature of the address game item confused respondents. This latter point could explain why more errors of duplication and non-differentiation in the children's address game item were identified in the survey data. An alternative explanation is that choosing some or many duplicates could be seen as a valid response to the address item if one felt all levels of the address were equally important. The most key point, however, is that Hypothesis 1 was supported by the data, with more errors in web than CAPI.

4.1.2 Acquiescence response bias on agree/disagree items – Hypothesis 2

To test Hypothesis 2, that acquiescence would be more prevalent in web than in the interviewer administered modes (CATI/CAPI), we used 12 agree/disagree items from three multi-item scales (items 7-18 in Appendix A). The first scale contained items on the quality of the neighbourhood (items 7- 10); the second scale contained items on the thoroughness of preparation for a financial decision (items 11-14), and the third scale had items about mental patients and former prisoners living in the respondent's neighbourhood (items 15-18) and was designed to be sensitive. Each of the three scales contained both positive and negative statements.

First we analysed acquiescence at the item level. Here we focused on the 8 non-sensitive statements (items 7-14). The sensitive items were excluded, as both socially desirable answering to the positive statements item (items 16 and18) and potentially more truthful answers to the negative statements (items 15 and 17) could be confused with acquiescence.[12] The acquiescence indicator took the value 1 if the respondent answered 'strongly agree' or-'agree', and 0 otherwise.[13] Hypothesis 2 was not supported in the first analysis: web respondents were not more likely to acquiesce with any of the items; in fact the opposite pattern was found. As shown in Table 2, there were significantly higher levels of agreement in the interviewer modes on four of the items in the Omnibus data and 6 of the items in the BHPS data.

[12] 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.

[13] Agree/disagree scales are often analysed in this way because research shows 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, p66; see also 'response contraction bias' in Tourangeau, Rips and Raskinski, 2000)

Table 2: Acquiescence as measured by choice of 'strongly agree' or 'agree' categories

Quality of neighbourhood items

		Omnibus data				BHPS data			
Item ref number	Strongly Agree & Agree	CAPI	CATI	Web	Results	CAPI	CATI	Web	Results
7	%	90.4	88.9	85.4	CAPI>web, OR=1.69, p<.05 CATI>web, OR=1.58, p<.10	NA	92.9	87.7	CATI>web, OR=1.76, p<.05
	Base	282	314	349			421	334	
8	%	9.7	12.6	9.8	No significant differences by mode	NA	11.6	6.1	CATI>web, OR=2.45, p<.05
	Base	134	159	183			198	180	
9	%	59.9	56.7	56.4	No significant differences by mode	NA	68.3	64.7	No significant differences by mode
	Base	282	314	349			419	334	
10	%	75.5	73.3	76.5	No significant differences by mode	NA	80.3	76.7	No significant differences by mode
	Base	135	161	183			198	180	

Thoroughness of preparation before financial decision items

Item ref number	Strongly Agree & Agree	CAPI	CATI	Web	Results	CAPI	CATI	Web	Results
11	%	35.2	43.0	28.7	CAPI>web, OR=1.36, p <.10 CATI>CAPI, OR=1.36, p<.10 CATI>web, OR=1.85, p<.001	NA	42.0	32.3	CATI>web, OR=1.59, p<.01
	Base	281	314	349			419	334	
12	%	89.7	90.8	85.1	CAPI>web, OR=1.63, p<.10 CATI>web, OR=1.89, p<.05	NA	88.3	80.8	CATI>web, OR=1.74, p<.01
	Base	282	314	349			419	333	
13	%	41.3	47.5	39.0	CATI>web, OR=1.40, p<.05	NA	45.2%	39.3	CATI>web, OR=1.34, p<.10
	Base	281	314	349			418	333	
14	%	70.9	70.8	64.5	No significant differences by mode	NA	75.8	65.6	CATI>web, OR=1.81, p<.05
	Base	134	161	183			198	180	

Table 3: Differences in acquiescence as measured by agreement to opposite statements

	Omnibus data				BHPS data			
Scale	CAPI %	CATI %	Web %	Results	CAPI %	CATI %	Web %	Results
Neighbourhood scale	3.7	4.4	2.2	No significant differences by mode	NA	5.6	2.2	CATI>web, OR=3.12, p<.10
Financial decisions scale	42.5	52.8	39.3	CATI>web, OR= 1.60, p=.05	NA	49.2	40.2	CATI>web, OR=1.48, p<.10
Sensitive scale	35.6	35.2	27.7	CAPI>web, OR=1.52, p<.10	NA	35.3	27.5	No significant differences by mode

Second we examined acquiescence at the scale level.[14] In multi-item scales, acquiescence behaviour is typically identified by respondents agreeing to opposite statements[15] (DeVellis, 2012). Focusing on respondents who agreed to all four statements and those who had agreed to a pair of opposite statements, the results in Table 3 again suggested that acquiescence was higher in the interview modes compared to web (contrary to Hypothesis 2).

Third, we used the cognitive interviews to look for evidence of mode differences across respondents who had agreed to opposite statements. Thirty two instances of agreement to opposite statements were found, but 9 of these could be excluded due to confusion over the word 'rarely' in two of the 'financial decision' statements (items 11 and 13) and to the absence of a not applicable category for all the items in the financial decision scale.

Surprisingly, of the remaining instances of agreement to opposite statements the majority could not be attributed to acquiescence. Respondents gave clear, justifiable reasons for why they chose the answer they did. For example, one respondent strongly agreed to item 8 in Appendix A (more properties in poor repair) because some houses could do with some work and agreed to item 10 (more properties well kept up) because *"in this village . . . it's like half and half. There is a bit*

[14] We tested the reliability of each scale. In the Omnibus sample, the internal consistency for the difficult scale (items 7-10) was alpha=0.75 and for the sensitive scale (items 15-18) was alpha=0.71, with similar results for each of the modes. However, the third scale (items 11-14) had poor internal consistency overall (alpha=0.33), a result that was repeated for each mode separately. Principal component analysis showed that the difficult and the sensitive scales were both unidimensional, while the third scale was not. These results were replicated in the BHPS dataset. The cognitive interview data suggest that poor scale statistics for the third scale could be due to the problematic word 'rarely' in the 2 negatively words statements of the scale (items 11 and 13) and a lack of a not applicable category for all of the statements in the scale.

[15] For example, "Compared to other neighbourhoods, this neighbourhood has more properties that are in a poor state of repair" as opposed to "Compared to other neighbourhoods, this neighbourhood has more properties that are well-kept". "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". "I would rarely read all the small print before making important financial decisions" as opposed to "I would do a lot of research before making an important financial decision".

[that] . . . wants doing up and there's the other part which doesn't" (Female, 50 to 59, no qualifications, employed, very low income, White British).

Only two cases showed evidence of acquiescence. A clear case was apparent when a respondent was unable to justify her answer to the survey items in CATI. This was a Pakistani female respondent who did not understand the item and chose 'agree' as her choice. In the cognitive retrospective think alouds she explained: "*I think I don't understand that, I just say agree"* (Female, 30 to 39, no qualifications, low income, with poor English as rated by the interviewer). This respondent's behaviour is in line with cultural norms where 'agree' reflects politeness (see Javeline, 1999) and the 'category fallacy' theory (see Warnecke, et al, 1997). A second, possible case of acquiescence involved a respondent answering in web who had ambivalent feelings and commented that it was hard to choose agree or disagree. Interestingly, respondents with similar views to the respondent chose the middle category; thus her choice of 'agree' could be a type of acquiescence (Female, 40 to 49, higher education below degree level, employed, low income, White British). Ability to justify answers to the agree/disagree items did not differ by original data collection mode for any of the items.

4.1.3 Middle category satisficing – Hypothesis 3

Hypothesis 3 suggested that web respondents who are without the presence of the interviewer as opposed to CAPI and CATI respondents would be more likely to choose a middle category option. We first explored this hypothesis with the difficult item formats, including long scales (satisfaction with street cleaning and satisfaction with waste and recycling collection, items 1 and 2 in Appendix A), agree/disagree items (items 7-18), and end-labelled scales (satisfaction with the economy and satisfaction with democracy and personal freedom, items 21and 22).[16] If the respondent selected the middle category, the indicator took a value of 1, and 0 otherwise.

The results for both the Omnibus and BHPS data are shown in Table 4a. These indicated that web respondents were more likely to choose the middle category than CAPI and/or CATI respondents on the long satisfaction items (items 1 and 2 with one of the two items showing significant differences in the Omnibus data and both items showing significance in the BHPS data) and the agree/disagree items (items 7-18, with 7 of the twelve items showing significance in the Omnibus

[16] The other difficult items (items 3-6, 19-20 and 23-24 were excluded as these items did not have 'sensible' middle categories.

data and 6 of twelve items showing significance in the BHPS data).[17] Thus both of these item formats show support for Hypothesis 3. In contrast, there is no evidence for more middle category endorsing among web respondents on the end-labelled questions (items 21 and 22 with no significant results in the Omnibus data and only one in the BHPS data).

We then investigated Hypothesis 3 using the relevant easy category formats. As shown in Table 4b, these included the short scale versions of satisfaction items 1 and 2 and the fully labelled versions of satisfaction items 21 and 22. Both the Omnibus and BHPS data showed that web respondents were significantly more likely to select the middle category on the three category satisfaction scales, but not for the fully-labelled 7 category satisfaction items (items 21 and 22).

Table 4a: Evidence for web middle category effects in difficult formats

Item Ref Number	Item Topic	Omnibus Data	BHPS Data
1	7-point satisfaction with street cleaning	web>CATI, OR=1.89, p < .10	web>CATI, OR=1.84, p<.10
2	7- point satisfaction with waste and recycling collection	No significant differences by mode	web>CATI, OR=1.51, p<.05
7	Neighbourhood not a bad place	web>CAPI, OR=2.67, p<.01 web>CATI, OR=1.83, p<.05	No significant differences by mode
8	More properties in bad state of repair	web>CATI, OR=1.86, p<.10	web>CATI, OR= 2.77, p<.01
9	Not suffer from litter, dog mess and graffiti	web>CAPI, OR=1.75, p<.05 web>CATI, OR=1.58, p<.05	web>CATI, OR= 1.44, p<.10
10	More properties that are well kept	No significant differences by mode	No significant differences by mode
11	Financial decision: Rarely read the small print	web>CAPI, OR=2.44, p<.01 web>CATI, OR= 2.21, p<.01	web>CATI, OR= 2.09, p<.01
12	Financial decision: Do a lot of research	web>CAPI, OR=4.15, p<.001 web>CATI, OR=3.09, p<.001	web>CATI, OR= 2.74, p<.001
13	Financial decision: Rarely talk to financial advisor	web>CAPI, OR=2.03, p<.01	web>CATI, OR= 2.21, p<.001
14	Financial decision: Definitely talk to family and friend	No significant differences by mode	No significant differences by mode
15	Would worry if mental health patients lived in	CATI>web, OR=1.43, p<.05	CATI>web, OR=1.44, p<.05

[17] Note that this does not include item 15 in the Omnibus data where there are more middle category endorsements in CATI than web.

	neighbourhood		
16	Mental health patients have just as much right to live in neighbourhood	web>CAPI, OR=2.01, p<.001 web>CATI, OR=2.37, p<.001	web>CATI, OR= 2.21, p<.001
17	Would worry if former prisoners lived in neighbourhood	No significant differences by mode	No significant differences by mode
18	Former prisoners have just as much right to live in neighbourhood	No significant differences by mode	No significant differences by mode
21	End-labelled satisfaction with the economy	No significant differences by mode	No significant differences by mode
22	End-labelled Satisfaction with democracy and personal freedom	CAPI>web, OR1.53, p<.10 CAPI>CATI, OR=2.24, p<.01	No significant differences by mode

Table 4b: Evidence for CAWI middle category effects in easy formats

1	3-point satisfaction with street cleaning	web>CAPI, OR=1.66, p<.10 web>CATI, OR=2.26, p<.01	web>CATI, OR=2.22, p<.01
2	3- point satisfaction with waste and recycling collection	web>CAPI, OR=2.08, p<.05 web>CATI, OR=3.71, p<.001	web>CATI, OR=2.62, p<.01
21	Fully-labelled satisfaction with the economy	No significant differences by mode	No significant differences by mode
22	Fully-labelled satisfaction with democracy and personal freedom	No significant differences by mode	No significant differences by mode

However, choosing a middle category may or may not be an act of satisficing. The cognitive interviews were able to make a distinction between 'clear' satisficing, 'possible' satisficing and no satisficing. Any cases which were not obvious (or 'clear' satisficing) were categorized as 'possible' satisficing.[18] The distinctions can be seen more clearly with the examples given in Figure 1.

Figure 1: Examples of 'clear' and 'possible' satisficing

Examples of 'clear' satisficing:

- "*I'll be truthful, I just answered that, with no thought in my head*" (Male, no qualifications, low income, White British)

[18] The categorizations of the three researchers doing the analysis were reviewed by the leader of the cognitive interviewing project to ensure reliability.

- "*To tell you the truth, I just clicked it*" (Female, no qualifications, very low income, White British)
- "*I'm not too sure, I think you have me on that one"* (Male, high school equivalent, on incapacity benefit, White British)

Examples of 'possible' satisficing:

- Chose 'neither nor' because not that bothered about the state of repair of properties (Female, higher education below degree, medium income, White British)
- Admitted this is not something she things about (Female, first degree, high income, White British)
- "*Is slightly satisfied the middle one? I'll go for the middle one"* (Female, first degree, high income, White British)

For this hypothesis, the cognitive interviews first focused on an investigation of CATI versus web respondents for items 1 and 2 in Appendix A (satisfaction with street cleaning and satisfaction with waste and recycling collection). In the survey questions administered at the beginning of the cognitive interviews, there was a slightly higher number of endorsements of middle categories for web respondents than CATI respondents (e.g. 9 of 25 web versus 5 of 25 CATI respondents chose the middle category on at least one of the questions). The cognitive interviews showed that almost all of the respondents who chose the middle category on a survey question did so for sensible and justifiable reasons. The few cases of 'possible' satisficing were found in the web mode of data collection.

Second, the cognitive interviews explored the issue of middle category satisficing across all 3 modes for items 7-18 (the 12 agree/disagree questions). Here once again, only a few of the respondents who chose the middle category were classified as cases of 'clear' and 'possible' satisficing. The remainder of the respondents had validly chosen the middle category. The instances of satisficing occurred in both CAPI and web, but the 'clear' satisficing was only found for web respondents. These results of the cognitive interviews suggested that not all middle category endorsements represent satisficing. In particular, the results suggested the possibility of more satisficing in web[19], [20] and thus are aligned with the quantitative findings.

[19] As part of the larger grant project, reason for category choice was explored across all categories, not just middle categories and across more items than just the 24 explored in this paper. The general finding was that there was more satisficing in web as compared to CAPI and more

In summary, there was evidence to confirm that middle category endorsement was more common for web respondents, thus supporting Hypothesis 3. Curiously, middle category endorsement was linked to some but not all of question formats, with evidence on both the long and short items and the agree/disagree items. However, there was no support for the hypothesis from the end-labelled versus fully-labelled experiment.

4.1.4 Primacy and recency effects on items with long lists of categories – Hypotheses 4

We explored primacy effects with all items of 5 or more response categories. This included 7 and 8 category scales from a long versus short scale experiment (items 1-6 in Appendix A), 5 category agree/disagree scales (items 7-14) [21], 6 and 7 category ranking tasks (items 19 and 20) and 7 category end-labelled scales from an end-labelled versus fully labelled experiment (items 21-24). In all cases, the primacy indicator took the value 1 if the respondent selected the first response option, and 0 otherwise; the recency indicator took value 1 if the respondent selected the last option, and 0 otherwise.

First, we explored the six items from the long versus short scale experiment (items 1 – 6) because this experiment was crossed with a showcard/no showcard experiment in CAPI. This allowed us to isolate the effects of the interviewer's presence, holding the visual presentation of the response options constant across modes. For example, we tested for differences in the extent of primacy effects between web and CAPI respondents who were given showcards. This analysis used only Omnibus data as BHPS CAPI data were not yet available. As shown in Table 5a (Omnibus data, findings in bold), only one of the six items examined (satisfaction with street cleaning, item 1),

---

satisficing in CATI as compared to CAPI. See Campanelli et al 2010 for details about this part of the cognitive interviewing results.

[20] Cognitive interviewing is often seen as a qualitative method which would preclude any kind of quantification. But to try to understand mode differences, which are usually manifested at the aggregate level, it was difficult to avoid looking at the magnitude of the differences across modes. A compromise was to use quantifiers like 'a few'. And the reader needs to remember that these quantified amounts are unique to this sample of respondents and that the same prevalences may not be replicated with a different sample.

[21] Agree/disagree items 15-18 were excluded. These were sensitive items, so a primacy effect would be confounded with a socially desirable answer when the first category was a positive category and a recency effect when the last category was a socially desirable answer.

showed the expected relationship with web respondents more likely to choose the first category than CAPI respondents with a showcard. There were primacy effects shown on two other items, but not in the expected direction: CATI with more primacy effects than CAPI (item 1) and web with more primacy effects than CATI (item 4).

Using the same six questions, we then tested for differences between web and CATI/CAPI without showcards to investigate whether response order effects differed between interviewer administered and self-completion modes. When response options were presented visually (web), we expected response order effects in the form of primacy; when response options were presented orally (CATI/CAPI without showcards), we instead expected recency effects. Regardless of the type of order effects, we expected the extent to be larger with self-completion (web) than interviewer assisted modes (CATI/CAPI). As shown in Table 5b, there was only one instance of a web primacy effect for any of the six questions in comparison to CAPI and CATI for the Omnibus data or in comparison to CATI for the BHPS data. This was on Omnibus item 4 (amount spent on leisure activities) with more web respondents than CATI respondents choosing the first category (odds ratio = 1.72). Interestingly, this item also showed the expected recency effect, with CATI respondents more likely than web respondents to choose the last category (odds ratio = 5.91). These findings fit part, but not all, of the expectations our hypothesis. We would have expected a larger odds ratio from the web primacy effect than the CATI recency effect, but this was not the case. There were two other instances where there was a recency effect in the Omnibus data, with CAPI or CATI having larger values for the last category than web (items 1 and 2), but there was no corresponding primacy effect. .

Next we investigated primacy and recency effects on all the other items with 5 or more categories. The ranking tasks (items 19-20 in Appendix A) and 7 category end-labelled scales and fully-labelled scales (items 21-24) all had showcards. As shown in the upper part of Table 5c ('CAPI with showcards') there were no examples where web respondents had a higher endorsement of the first category than CAPI respondents. The lower part of Table 5c ('CAPI with no showcard') shows the 5 category agree/disagree scales (items 7-14) which were asked without showcards. There were no instances of primacy effects, where web had higher endorsements of the first category in either the Omnibus or BHPS data. In the BHPS data there was one instance of a recency effect. This was on item 8 (agreeing/disagreeing that the neighbourhood has more properties that are in a poor state of repair), with CATI respondents more likely to endorse the last category than web respondents. But as described for Table 5b, this only supports a part of Hypothesis 4.

Primacy and recency effects were also explored for the items with less than 5 categories. This included the short scale versions of items 1 through 6 and rating scales versions of items 19 and 20 (see Appendix A). The results are found in Table 5d. There were two examples of web primacy effects in the expected direction. These were item 4 in the BHPS data (amount spent on leisure activities) and item 19a in the Omnibus data (importance of less crime in neighbourhood). Item 19a also showed the expected recency effect, with CAPI more likely than web respondents to select the last response category. For this item, the odds ratio for the primacy effect was only slightly larger than that for the recency effect (Odds ratio=1.99 versus 1.91, respectively), thus possibly supporting Hypothesis 4. Item 19a was one of 7 items from the Omnibus that showed expected recency effects, with CAPI (without a showcard) and/or CATI respondents more likely to choose the last category than web respondents. The others were items 2, 19b, 19c, 19f, 19g and 20f. There were 4 similar instances between CATI and web respondents in the BHPS data (items 19b, 19c, 19d and 19f), but with no corresponding primacy effects these findings only provide partial support for Hypothesis 4. The prevalence of so many recency effects on a short 4 category scale was surprising given the small number of options for the respondent to attend to before selecting a response.

In summary although 45 questions were examined with the Omnibus data and 39 with the BHPS data, there were very few instances of a traditional primacy effect in the visual modes or recency effects in the oral modes. Exceptions were the several unexpected recency effects on the four category rating questions (Table 5d). However, there is little evidence to support Hypothesis 4, as only one of the comparisons between CAPI with a showcard and web showed the expected primacy effects (Table 5a) and only one of the comparisons between CAPI without a showcard / CATI and web showed the expected primacy effects in web to be slightly more pronounced than the expected recency effects in CAPI/CATI. Unexpectedly, this was on a 4 category rating scale rather than the hypothesised long scales. Note that the cognitive interviews did not specifically address primacy and recency effects.

In contrast, over the 43 Omnibus and 37 BHPS items[22] in Tables 5a-d, there was a pattern of CATI respondents being more likely to select the first response option than respondents in other modes. As shown in italics, 14 of the 43 items in the Omnibus data and 14 of the 37 items in the BHPS data

[22] There are fewer BHPS items because the BHPS is excluded from the 6 items in Table 5a. In addition, the ranking version of items 19 and 20 in Table 5c are excluded from both and Omnibus and BHPS data as ranking was not administered in CATI.

show this pattern, which was most prevalent on the satisfaction scales (items 1-2 – both long and short versions, tables 5a, 5b and 5d respectively) and the agree/disagree scales (items 7-14, lower part of table 5c). There was little evidence for this pattern on the factual items from the long versus short scales experiment (items 3-6, Tables 5a, 5b and 5d), either the end-labelled or fully labelled scales (items 21-24, top of Table 5c), the rating tasks (items 19-20, Table 5d). Finding primacy effects for CATI respondents was surprising as primacy effects are mainly expected when response categories are presented visually.

Table 5a: Evidence of increased likelihood of web respondents choosing the first category compared to CAPI respondents with showcard (Other primacy effects also shown)

Item Ref Number	Item Topic	Type of order effect and Item Format	Omnibus data	BHPS data
CAPI with showcard				
1	Satisfaction with street cleaning	Primacy effects: 7-category satisfaction item	**web>CAPI, OR=2.62, p<.05** *CATI>CAPI, OR=3.96, p<.01*	NA
2	Satisfaction with waste and recycling collection	Primacy effects: 7-category satisfaction item	No significant differences by mode	NA
3	Length lived in area	Primacy effects: 7-category item	No significant differences by mode	NA
4	Amount spent on leisure activities	Primacy effects: 8-category item	web>CATI, OR=1.86, p<.05	NA
5	Type of dwelling	Primacy effects: 8-category item	No significant differences by mode	NA
6	Locations nearest to home	Primacy effects: 7-category item	No significant differences by mode	NA

Note that this table uses bold for findings in the expected direction, italics for CATI positivity bias and grey shading of other statistically significant effects.
'NA' refers to BHPS CAPI data not being available for this comparison.

Table 5b: Evidence of increased likelihood of web respondents choosing the first category compared to CAPI respondents without a showcard and CATI respondents more likely to choose the last category (Other primacy and recency effects also shown)

Item Ref Number	Item Topic	Item Format	Omnibus data	BHPS data
CAPI with no showcard				
1	Satisfaction with street cleaning	Primacy effects: 7 category satisfaction item	*CATI>CAPI, OR=2.01, p<.10*	No significant differences by mode
				No significant differences by

		Recency effects	**CAPI>web, OR=3.36, p<.10**	mode
2	Satisfaction with waste and recycling collection	Primacy effects: 7 category satisfaction item	*CATI>CAPI, OR=1.94, p<.10*	*CATI>CAWI, OR=1.51, p<.10*
		Recency effects	**CAPI>web, OR=3.64, p<.10**	No significant differences by mode
3	Length lived in area	Primacy effects: 7 category item	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
4	Amount spent on leisure activities	Primacy effects: 8 category item	CAPI>CATI, OR=2.61, p<.05 **web>CATI, OR=1.72, p<.05**	No significant differences by mode
		Recency effects	**CATI>web, OR=5.91, p<.01** CATI>CAPI, OR=3.15, p<.10	No significant differences by mode
5	Type of dwelling	Primacy effects: 8 category item	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
6	Locations nearest to home	Primacy effects: 7 category item	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode

Note that this table uses bold for findings in the expected direction, italics for CATI positivity bias and grey shading of other statistically significant effects.

Table 5c: Evidence of primacy and recency effects on all other scales with more than 5 categories (Other primacy and recency effects also shown)

Item Ref Number	Item Topic	Item Format	Omnibus data	BHPS data
CAPI with showcard				
19	Improve-ments to the neigh-bourhood	Primacy effects (% endorsement of first category of ranking task): 7 category ranking task	No significant differences by mode	NA
20	Children's address game	Primacy effects (% endorsement of first category of ranking task): 6 category ranking task	No significant differences by mode	NA
21	Satisfaction with the economy	Primacy effects: 7 category end-labelled	No significant differences by mode	No significant differences by mode

		Recency effects	No significant differences by mode	No significant differences by mode
		Primacy effects: 7 category fully labelled	No significant differences by mode	No significant differences by mode
		Recency effects	CAPI>web, OR=1.54, p<.10	No significant differences by mode
22	Satisfaction with democracy and personal freedom	Primacy effects: 7 category end-labelled	*CATI>CAPI, OR=5.97, p < .01* *CATI>web, OR=4.13, p < .01*	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
		Primacy effects: 7 category fully labelled	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
23	Frequency of grocery shopping	Primacy effects: 7 category end-labelled	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
		Primacy effects: 7 category fully labelled	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
24	Purchases of hot beverages	Primacy effects: 7 category end-labelled	No significant differences by mode	*CATI>web, OR=1.53, p<.10*
		Recency effects	No significant differences by mode	No significant differences by mode
		Primacy effects: 7 category fully labelled	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
CAPI with no showcard				
7	Neigh-bourhood not a bad place	Primacy effects: 5 category agree / disagree	*CATI>CAPI, OR=1.55, p<.01* *CATI>web, OR=1.84, p<.001*	*CATI>web, OR=1.92, p<.001*
		Recency effects	No significant differences by mode	No significant differences by mode
8	More properties in bad state of repair	Primacy effects: 5 category agree / disagree	No significant differences by mode	*CATI>web, OR=2.77, p<.01*
		Recency effects	No significant differences by mode	**CATI>web, OR1.57, p<.10**

9	Not suffer from litter, dog mess and graffiti	Primacy effects: 5 category agree / disagree	*CATI>web, OR=1.74, p<.01*	*CATI>web, OR=1.44, p<.10*
		Recency effects	No significant differences by mode	No significant differences by mode
10	More properties that are well kept	Primacy effects: 5 category agree / disagree	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
11	Financial decision: Rarely read the small print	Primacy effects: 5 category agree / disagree	CAPI>web, OR=1.73, p<.05 *CATI>CAPI, OR=1.65, p<.05* *CATI>web, OR=2.86, p<.001*	*CATI>web, OR=4.08, p<.001*
		Recency effects	No significant differences by mode	No significant differences by mode
12	Financial decision: Do a lot of research	Primacy effects: 5 category agree / disagree	CAPI>web, OR=1.48, p<.05 *CATI>web, OR=1.84 p<.001*	*CATI>web, OR=2.06, p<.001*
		Recency effects	No significant differences by mode	No significant differences by mode
13	Financial decision: Rarely talk to financial advisor	Primacy effects: 5 category agree / disagree	*CATI>web, OR=1.83, p<.05*	*CATI>web, OR=1.89, p<.05*
		Recency effects	No significant differences by mode	No significant differences by mode
14	Financial decision: Definitely talk to family and friend	Primacy effects: 5 category agree / disagree	CAPI>web, OR=1.56, p<.10	*CATI>web, OR=3.12, p<.001*
		Recency effects	No significant differences by mode	No significant differences by mode

Note that this table uses bold for findings in the expected direction, italics for CATI positivity bias and grey shading of other statistically significant effects.

Table 5d: Evidence of primacy and recency effects on all other scales with fewer than 5 categories (Other primacy and recency effects also shown)

Item Ref Number	Item Topic	Item Format	Omnibus data	BHPS data
CAPI with no showcard				
1	Satisfaction with street cleaning	Primacy effects: 3 category satisfaction item	CAPI>web, OR=1.78, p<.05 *CATI>CAPI, OR=1.58, p<.10* *CATI>web, OR=2.82, p<.001*	*CATI>web, OR=2.11, p<.01*
		Recency effects	web>CATI, OR=2.07, p<.05	No significant differences by mode
2	Satisfaction with waste and recycling collection	Primacy effects: 3 category satisfaction item	*CATI>CAPI, OR=1.61, p<.10* *CATI>web, OR=1.82, p<.05*	*CATI>web, OR=2.53, p<.001*
			CAPI>web, OR=1.83, p<.10	No significant differences by

		Recency effects		mode
3	Length lived in area	Primacy effects: 3 category item	*CATI>web, OR=2.39, p<.05*	*CATI>web, OR=2.54, p<.10*
		Recency effects	No significant differences by mode	No significant differences by mode
4	Amount spent on leisure activities	Primacy effects: 3 category item	No significant differences by mode	**web>CATI, OR=1.82, p<.01**
		Recency effects	No significant differences by mode	No significant differences by mode
5	Type of dwelling	Primacy effects: 3 category item	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
6	Locations nearest to home	Primacy effects: 3 category item	No significant differences by mode	*CATI>web, OR=1.68, p<.05*
		Recency effects	No significant differences by mode	No significant differences by mode
19a	Improve-ments to the neigh-bourhood – less traffic	Primacy effects: 4 category rating task	**web>CAPI, OR=1.99, p<.05**	No significant differences by mode
		Recency effects	**CAPI>web, OR=1.91, p<.05** CAPI>CATI, OR=1.64, p<.10	No significant differences by mode
19b	Improve-ments to the neigh-bourhood – less crime	Primacy effects: 4 category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	**CAPI>web, OR=2.27, p<.10** **CATI>web, OR=2.53, p<.05**	**CATI>web, OR=3.28, p<.01**
19c	Improve-ments to the neigh-bourhood – more/better shops	Primacy effects: 4 category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	**CAPI>web, OR=1.61, p<.10** **CATI>web, OR=1.82, p<.05**	**CATI>web, OR=2.54, p<.001**
19d	Improve-ments to the neigh-bourhood – better schools	Primacy effects: 4 category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	**CATI>web, OR=1.65, p<.05**

19e	Improve-ments to the neigh-bourhood – more/better leisure facilities	Primacy effects: 4category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
19f	Improve-ments to the neigh-bourhood – better transport links	Primacy effects: 4 category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	**CAPI>web, OR=2.42, p<.01** CAPI>CATI, OR=1.69, p<.10	**CATI>web, OR=1.94, p<.05**
19g	Improve-ments to the neigh-bourhood – more parking spaces	Primacy effects: 4 category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	**CAPI>web, OR=1.57, p<.10**	No significant differences by mode
20a	Children's address game	Primacy effects: 4 category rating task	No significant differences by mode	*CATI>web, OR=1.78, p<.05*
		Recency effects	No significant differences by mode	No significant differences by mode
20b	Children's address game	Primacy effects: 4category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
20c	Children's address game	Primacy effects: 4 category rating task	*CATI>web, OR=1.63, p<.10* CAPI>web, OR=1.70, p<.05	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
20d	Children's address game	Primacy effects: 4 category rating task	*CATI>web, OR=1.64 p<.05*	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
20e	Children's address game	Primacy effects: 4 category rating task	No significant differences by mode	No significant differences by mode
		Recency effects	No significant differences by mode	No significant differences by mode
20f	Children's address game	Primacy effects: 4 category rating task	CATI>CAPI, OR=1.98, p<.10	No significant differences by mode
		Recency effects	**CAPI>web, OR=2.03, p<.05** **CATI>web, OR=1.75, p<.10**	No significant differences by mode

Note that this table uses bold for findings in the expected direction, italics for CATI positivity bias and grey shading of other statistically significant effects.

4.1.5 Item non-response – Hypothesis 5

To test the hypothesis that there would be more satisficing in the form of item non-response with self-completion compared to interviewer administered modes, we focused on the more difficult formats of ranking (items 19-20 in Appendix A) and end-labelled scales (items 21-24). As shown in Table 6, there was very little item non-response. Although the expected pattern of more item-nonresponse in web than the interviewer administered modes was observed for items 1 and 2, there were no significant differences between modes.

Table 6 shows that missingness was higher in web compared to CAPI for both ranking tasks. Although both finding are in the expected direction, the differences were small and do not reach significance. There was virtually no missing data on the end-labelled scales. High levels of complete data suggest that, in this study at least, missingness cannot be considered to be a proxy for interviewer help and motivation for respondents to undertake the difficult task of completing end labelled items. Thus Hypothesis 5 is not supported due to the small amounts of item missing data. The cognitive interviews did not address item non-response.

Table 6: Percent item non-response by mode and item

	Omnibus data						BHPS data		
Item Ref Number	Item Topic	Item format	CAPI %	CATI %	Web %	Results	CATI %	Web %	Results
19	Improvements to the neighbourhood	7 category ranking task	4.1	NA	5.4	No significant differences by mode	NA	NA	NA
20	Children's address game	6 category ranking task	1.5	NA	2.2	No significant differences by mode	NA	NA	NA
21	Satisfaction with the economy	7 category end-labelled scale	0.0	0.0	0.0	No significant differences by mode	0.5	0.0	No significant differences by mode
22	Satisfaction with democracy and personal freedom	7 category end-labelled scale	0.0	0.0	0.0	No significant differences by mode	1.0	0.0	No significant differences by mode
23	Frequency of grocery shopping	7 category end-labelled scale	0.0	0.6	0.0	No significant differences by mode	1.5	0.6	No significant differences by mode
24	Purchases of hot beverages	7 category end-labelled scale	0.0	0.0	0.0	No significant differences by mode	0.5	0.0	No significant differences by mode

4.1.6 Interviewer effects on Social desirability – Hypothesis 6

We tested the impact of interviewer presence on socially desirable responding using a scale of sensitive agree/disagree statements about people with mental health problems/prisoners living in the respondent's neighbourhood (items 15-18 in Appendix A). Responses to the sensitive statements were clearly in the direction of socially desirability in the interviewer modes but not in the web survey. This held true for each of the 4 sensitive questions with both the Omnibus and BHPS data. These results support Hypothesis 6.

Note that the differences between 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. There were no differences in the level of socially desirable reporting between face-to-face and telephone interviewing. No showcards were used in the face-to-face interviewing, removing-aural versus visual distinctions that could confound this result. The cognitive interviews did not address social desirability bias.

Table 7: Social desirability measured by choice of 'strongly agree' or 'agree' categories to the sensitive scale

		Omnibus data				BHPS data			
Item ref number	Direction of socially desirable answer	CAPI %	CATI %	Web %	Results	CAPI %	CATI %	Web %	Results
15	Lower percentage	37.9	36.1	53.0	web>CAPI, OR=1.78, p<.001 web>CATI, OR=1.64, p<.05	NA	32.3	52.7	web>CATI, OR=2.36, p<.001
	Base	282	313	349			421	332	
16	Higher percentage	64.8	69.0	43.8	CAPI>web, OR=2.30, p<.001 CATI>web, OR=1.80, p<.05	NA	65.3	45.6	CATI>web, OR=2.38, p<.001
	Base	281	313	349			421	333	
17	Lower percentage	70.1	60.1	72.9	CAPI> CATI, OR=1.67, p<.05 web>CATI, OR=1.78, p<.05	NA	64.3	72.5	web>CATI, OR=1.51, p<.10
	Base	147	153	166			221	153	
18	Higher percentage	45.9	51.0	33.1	CAPI>web, OR=1.83, p<.05 CATI>web, OR=2.09, p<.01	NA	43.2	30.1	CATI>web, OR=1.78, p<.05
	Base	146	153	166			222	153	

5. DISCUSSION

The results of our research illustrate different ways in which interviewers may or may not influence responses, contributing to differences in measurement between self-completion modes (such as web) and interviewer administered modes. We had hypothesised that on non-sensitive items the presence of an interviewer would motivate respondents to generate an accurate answer and to reduce the difficulty of the task by offering support and providing explanations of what is needed. This in turn should reduce the likelihood of the respondents demonstrating satisficing behaviour in difficult item formats. In contrast, we hypothesised that on sensitive items, the presence of an interviewer would reduce the privacy of the reporting situation, which can have an impact on respondents' willingness to answer truthfully.

There was evidence that interviewers helped respondents carry out complicated tasks. Web respondents were more likely to complete ranking tasks incorrectly (by assigning the same rank to more than one item), than CAPI respondents (Hypothesis 1). The cognitive interviews further suggested a general confusion among respondents about how to complete ranking tasks.

Similarly, there was evidence that interviewers did motivate respondents to fully consider an item and the response options, reducing the extent of satisficing. Web respondents were more likely to select middle response categories than CAPI or CATI respondents (Hypothesis 3). The cognitive interviews supported this finding and were further able to distinguish satisficing from justified reasons for selecting a middle response category.

It is less clear why middle categories were chosen. Web respondents were more likely to choose the middle category than interviewed respondents for items that represented both easy and difficult item formats. Web respondents, as opposed to interview respondents, were also more likely to choose a middle category on some types of items (agree/disagree scales and some satisfaction scales). The satisfaction questions from the long versus short scales experiment clearly showed middle category satisficing, whereas the satisfaction questions from the end-labelled versus fully-labelled ones did not.[23]

[23] This could be due to the items themselves. Item 22 (satisfaction with democracy and personal freedom in end-labelled format behaves like the two satisfaction questions from the long/short scale experiment with both more middle category endorsement in web and CATI primacy effects. The other satisfaction question (item 21 – satisfaction with the economy) showed a very different pattern

Contrary to expectations for Hypothesis 2, the extent of acquiescence was larger in CATI and CAPI than web. This is in accord with other research showing higher levels of acquiescence in telephone mode compared to postal mode (Dillman and Tarnai, 1991). Due to the many proposed reasons for acquiescing behaviour (as described in footnote 5), it is less clear acquiescence should be considered as an indicator of satisficing. If it is thought of as an indicator of politeness, then it would make sense for it to be more prevalent in interview modes. The cognitive interviews suggested that there were few instances where agreement to opposite statements could be taken to indicate acquiescence and the one clear example was actually one of cultural politeness.

Finally, we found little evidence of traditional primacy and recency effects (Hypothesis 4), which mirrors the results from previous studies of mail and telephone modes (e.g., Dillman et al., 1995). Instead, CATI respondents were found to be more likely to choose the first category (which was also the positive category) on the various scales. This finding makes a useful contribution to a controversy in the literature. Summarising over two decades of studies, Dillman, Smyth and Christian (2009) described a telephone positivity bias and suggested this was due to an aural versus visual effect. They found a *"substantial difference in responses to scalar items when asked by telephone versus visual modes"* (which includes face-to-face with showcard, mail and web), with telephone respondents providing more extreme positive answers. Similarly in a meta-analysis, Ye, Fulton and Tourangeau (2011), found that telephone respondents were *"more likely to endorse the most extreme positive response category (than in mail, IVR [Interactive Voice Response], or web surveys)" (p. 358).* But they found that face-to-face was like telephone and concluded that it was caused by a 'Mum about Undesirable Messages' (MUM) effect. That is, where respondents adjust their answers if they are undesirable for the receiver, in this case the interviewer. Curiously Ye, Fulton and Tourangeau's findings were based on 3 of the same 6 studies reviewed by Dillman, Smyth and Christian (2009), but Ye, Fulton and Tourangeau (2011) reached different conclusions. Unfortunately, Ye, Fulton and Tourangeau do not report whether or not a showcard was used in face-to-face mode. Our findings offer a new viewpoint which questions the findings of both other studies. Telephone respondents were more likely to give extreme positive answers, but there was

---

from most of the items in the questionnaire, with most respondents choosing the negative end of the scale and no differences by mode. This is understandable, since during the year of data collection, the UK was in economic recession, which is likely to have created more uniform, negative views among respondents.

no evidence of this for face-to-face. Most importantly this was true for face-to-face with a showcard (a visual mode) or without a showcard (an aural mode).

A key, but unexpected, finding was that patterns of satisficing can differ by item 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 items in the agree/disagree format.

Finally our study examined the impact of interviewer presence on respondents' answers to sensitive statements (Hypothesis 6). In contrast to the complex pattern of results for satisficing behaviour between modes, the evidence clearly showed that more socially desirable answers were provided in the interview modes, reinforcing the standard practice of including sensitive items within a self-completion module. There were no differences between the results for CAPI and CATI, which is in line with what has been found in some of the published literature.

Strengths and limitations

Our study had the advantages of (1) being based on a probability sample of the adult population (2) using random assignment to mode, and (3) including a cognitive interview follow-up study. Some mixed mode studies are based on special populations such as students (Smyth et al, 2008), and many do not have the opportunity for full randomisation because of its expense or the hope that assigning respondents to their preferred mode will increase response rates (Vannieuwenhuyze, Loosfelt and Molenberghs, 2010). Unfortunately the latter design confounds selection bias with mode effects.

Our use of cognitive interviews as a follow-up study provided useful insights. For example, the cognitive interviews highlighted the difficulties respondents can have with ranking tasks, a finding that reflects the typical use of cognitive interviewing. However, the cognitive interviewing conducted in this study extended the technique to provide a deeper understanding of patterns of quantitative results. For example, it allowed a distinction to be drawn between respondents who chose the middle category of an item as a satisficing response from those for whom it was a valid answer, the former being a much smaller proportion of the respondents. In addition, it demonstrates that most respondents do try to answer survey items as best they can, and that patterns of responses that would appear to indicate satisficing may not necessarily reflect this at the level of individual respondents. For example, the cognitive interviewing findings on acquiescence to agree/disagree scales showed that almost all respondents who had agreed to opposite statements did so for valid

reasons. Thus, the psychometric practice of eliminating such respondents from the analysis may be too harsh a solution.

One of the limitations of our study was the use of many pre-existing survey items from reputable sources that had unexpected problems, such as the use of the word 'rarely' in the financial decision multi-item scale (items 11-14) and the extreme negative responses generated by the satisfaction with the economy scale (item 21). Ideally all of the items should have been tested for our experiment before use.

A second limitation was that the very low levels of item non-response limited item-nonresponse as an indicator of satisficing. Due to this we were unable to conduct a proper investigation of Hypothesis 5.

In Summary

Our expected and unexpected findings make a useful contribution to the extant research literature. We have provided evidence that interviewers do motivate and help respondents while at the same time causing respondents to give socially desirable answers. We have also shown that findings vary by item format, and that even on the same item satisficing can be manifested in different ways. Our contribution to the interview survey positivity debate, suggests that more research is needed on this topic.

References

Alwin, D. and Krosnick, J. (1985). The Measurement of Values in Surveys: A Comparison of Ratings and Rankings. *Public Opinion Quarterly*, 49(4), 535-552.

Aneshensel, C. S., Frerichs, R. R., Clark, V. A. and Yokopenic, P. A. (1982). Measuring Depression in the Community: A Comparison of Telephone and Personal Interviews. *Public Opinion Quarterly,* 46(1), 110-121.

Aquilino, W. S. (1994). Interview Mode Effects in Surveys of Drug and Alcohol Use: A Field Experiment. *Public Opinion Quarterly*, 58(2), 210-240.

Betts, P. and Lound, C. (2010). The Application of Alternative Modes of Data Collection in UK Government Social Surveys: Review of Literature and Consultation with National Statistical Institutes', Office for National Statistics.

Biemer, P.P. and Lyberg, L.E. (2003). *Introduction to Survey Quality*. New York: Wiley.

Campanelli, P., Gray, M., Blake, M. and Hope, S. (2010). Mixed Modes and Measurement Error: Using Cognitive Interviewing to Explore the Results of a Mixed Modes Experiment. Presented at the annual meetings of the American Association for Public Opinion Research.

Carr, L. (1971), The Srole Items and Acquiescence. *American Sociological Review*, 36, 287-293.

Christian, L. and Dillman, D. (2004). The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions. *Public Opinion Quarterly*, 68(1), 57-80.

Christian, L., Dillman, D. and Smyth, J. (2008). Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys. In *Survey Advances in Telephone Methodology,* J. Lepkowski, et al. (eds). Hoboken, New Jersey: Wiley.

Ciochetto, S. Murphy, E. and Agarwal, A. (2006). Usability Testing of Alternative, Design Features for the 2005 National Census Test (NCT) Internet Form: Methods, Results, and Recommendations of Round-2 Testing. *Human-Computer Interaction Memorandum* #85, Washington, DC: U.S. Census Bureau (Usability Laboratory).

Couper, M., Traugott, M. and Lamias, M. (2001). Web Survey Design and Administration. *Public Opinion Quarterly*, 65(2), 230-253.

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.F. (2012). Scale Development: Theory and Applications, 3rd Edition. Thousand Oaks, CA: Sage.

Dillman, D. Brown, T., Carlson, J., Carpenter, E., Lorenz, F., Mason, R., Saltiel, J. and Sangster, R. (1995). Effects of Category Order on Answers to Mail and Telephone Surveys. *Rural Sociology*, 60(4), 674-687.

Dillman, D., Smyth, J. and Christian, L.M. (2009). *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method*, 3rd edition. Hoboken, NJ: Wiley.

Dillman, D. and Tarnai, J. (1991). Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys. In *Measurement Errors in Surveys*, P. Biemer, et al. (eds). New York: Wiley.

Feather, N. T. (1973). The measurement of values: Effects of different assessment procedures. *Australian Journal of Psychology*, 25, 221- 231.

Feldman-Naim, S., Myers, F.S., Clark, C. H., Turner, E. H., and Leibenluft, E. (1997). Agreement Between Face-to-Face and Telephone-Administered Mood Ratings in Patients with Rapid Cycling Bipolar Disorder. *Psychiatry Research,* 71, 129-132.

Fowler, F.J. Jr. (1995). *Improving Survey Questions: Design and Evaluation.* Thousand Oaks, California: Sage.

Gray, M., Blake, M., and Campanelli, P. (Forthcoming in 2014). The Use of Cognitive Interviewing Methods to Evaluate Mode Effects in Survey Questions, *Field Methods*, 26(2).

Hunt, S., Sparkman, R. and Wilcox, J. (1982). The Pretest in Survey Research: Issues and Preliminary Findings. *Journal of Marketing Research*, 19(2), 269-273.

Jackman, M. (1973). Education and Prejudice or Education and Response Set? *American Sociological Review*, 38, 327-339.

Javeline, D. (1999). Respond Effects in Polite Cultures: A Test of Acquiescence in Kazakhstan, *Public Opinion Quarterly*, 63(1), 1-28.

Johnson, T.P., Hougland, J.G. and Clayton, R.R. (1989). Obtaining Reports of Sensitive Behavior: A Comparison of Substance Use Reports from Telephone and Face-to-Face Interviews. *Social Science Quarterly*, 70, 174-183.

Kreuter, F., Presser, S. and 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, 213-236.

Krosnick, J. (1999). Survey Research. *Annual Review of Psychology*, 50, 537–567.

Krosnick, J. (2000). The Threat of Satisficing in Surveys: The Shortcuts Respondents Take in Answering Questions. *Survey Methods Newsletter*, 20(1). London: NatCen Social Research.

Krosnick, J. and Berent, M. (1993). Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format. *American Journal of Political Science*, 37(3), 941-964.

Krosnick, J. and Fabrigar, L. (1997). Designing Rating Scales for Effective Measurement in Surveys. In *Survey Measurement and Process Quality*, L. Lyberg, et al. (eds). Hoboken, New Jersey: Wiley.

Landsberger, H. and Saavedra, A. (1967). Response Set in Developing Countries. *Public Opinion Quarterly*, 31, 214-229.

Lenski, G., and Leggett, J. (1960). Caste, Class, and Deference in the Research Interview. *American Journal of Sociology*, 65, 463-467.

McIntyre S., and Ryans, A. (1977). Time and Accuracy Measures for Alternative Multidimensional Scaling Data Collection Methods: Some Additional Results. *Journal of Marketing Research*, 14, 607-610.

McBride, L. and 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, 115-118.

Millar, M. and Dillman, D. (2011). Improving Response to Web and Mixed-Mode Surveys. *Public Opinion Quarterly*, 75 (2), 249-269.

Narayan, S. and Krosnick, J. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly,* 60(1), 58-88.

Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A. and Lynn, P. (2011). Is It a Good Idea to Optimise Question Format for Mode of Data Collection? Results from a Mixed Modes Experiment. ISER Working paper, no. 2011-31, ISER, University of Essex.

Peabody, D. (1966). Authoritarianism Scales and Response Bias. *Psychological Bulletin*, 65, 11-23.

Peters, D.L., and McCormick, E.J. (1966). Comparative Reliability of Numerically Anchored Verusu Job-task Anchored Rating Scales. *Journal of Applied Psychology*, 50, 92-96.

Pless, I.B. and Miller, J.R. (1979). Apparent Validity of Alternative Survey Methods. *Journal of Community Health,* 5, 22-27.

Reynolds, T., and Jolly, J. (1980). Measuring Personal Values: An Evaluation of Alternative Methods. *Journal of Marketing Research*, 17, 37-80.

Rokeach, M. (1973). The Nature of Human Values. New York: Free Press.

Sangster, R. and Fox, J. (2000). *Housing Rent Stability Bias Study*. Washington, DC: U.S. Bureau of Labor Statistics, Statistical Methods Division.

Saris, W., Revilla, M., Krosnick, J. and Shaeffer, E. (2010). Comparing Questions with Agree / Disagree Response Options to Questions with Item-Specific Response Options. *Survey Research Methods*, 4(1), 61-79.

Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Quesiton Form*, Wording, and Context. New York: Academic Press.

Smyth, J. D., Christian, L.M. and Dillman, D. A. (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.

Sykes, W. and Collins, M. (1988). Effects of Mode of Interview: Experiments in the UK. In *Telephone Survey Methodology*, R. M. Groves, et al (eds). New York: Wiley.

Taylor, J., and Kinnear, T. (1971). Empirical comparison of alternative methods for collecting proximity judgments. *American Marketing Association Proceedings*. Fall Conference, 547.50.

Tourangeau, R., Couper, M., and Conrad, F. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68(3), 368-393.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Tourangeau, R. and Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin*, 133(5), 859-883.

Warnecke, R., Johnson, T. Chavez, N, Sudman, S, O'Rourke, D. Lacey, L. and Horm, J. (1997). Improving Question Wording in a Survey of Culturally Diverse Populations. *Annals of Epidemiology*, 7, 334-342.

Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-Mode Surveys. *Public Opinion Quarterly*, 74(5), 1027-1045.

Ye, C., Fulton, J. and Tourangeau, R. (2011). More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice. *Public Opinion Quarterly*, 75(2), 349–365.

Appendix A: Wording and source of items analysed

Item Formats	Showcard in CAPI?	Item Reference Number	Item Topic and Wording	Response Options for the Long Version of the Scale	Source
Long scales (from a long versus short scale experiment) crossed with showcard/no showcard in CAPI	A random half of CAPI respondents received a showcard and others did not	1	SATISFACTION WITH STREET CLEANING: And how satisfied or dissatisfied are you with street cleaning?	Very satisfied, Moderately satisfied, Slightly satisfied, Neither satisfied nor dissatisfied, Slightly dissatisfied, Moderately dissatisfied Very dissatisfied	Citizenship Survey, 2007
		2	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. . .	Very satisfied, Moderately satisfied, Slightly satisfied, Neither satisfied nor dissatisfied, Slightly dissatisfied, Moderately dissatisfied Very dissatisfied	Citizenship Survey ,2007 (modified to make item more difficult)
		3	LENGTH LIVED IN AREA: How long have you lived in this area?	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	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?	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	TYPE OF DWELLING: Which of these best describes your home?	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 Behaviours Towards the Environment, 2007
		6	LOCATIONS NEAREST TO HOUSE: Which of the following is closest to where you live?	A primary school, A secondary school, A 6th form college, A river, A lake, A cinema, A theatre	New

Item Formats	Showcard in CAPI?	Item Reference Number	Item Topic and Wording	Response Options	Source
Set of Four Agree / disagree statements	No showcards	7-10	**QUALITY OF NEIGHBOURHOOD**: The next few items are about the extent to which you agree or disagree with statements about your neighbourhood. Here is the first statement. • This neighbourhood is not a bad place to live. • Compared to other neighbourhoods, this neighbourhood has more properties that are in a poor state of repair. • Compared to other neighbourhoods, this neighbourhood does not suffer from things like litter, dog mess and graffiti. • Compared to other neighbourhoods, this neighbourhood has more properties that are well kept.	Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree	Modified and extended from a Southern Housing Association questionnare
Set of Four Agree / disagree statements	No showcards	11-14	**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.	Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree	Modified and extended from two statements from 2006 British Social Attitudes survey
Set of Four **SENSITIVE** Agree / disagree statements	No showcards	15-18	**MENTAL HEALTH PATIENTS AND FORMER PRISONERS IN R'S NEIGHBOURHOOD:** 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 neighbourhood 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 neighbourhood as any other people.	Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree	Extended from the Attitudes to Pensions Survey

Item Formats	Showcard in CAPI?	Item Reference Number	Item Topic and Wording	Response Options	Source
Ranking task (from a rating versus ranking experiment)	Showcards used	19	**IMPROVEMENTS TO THE NEIGHBOURHOOD:** What would you consider most important in improving the quality of your neighbourhood? Please rank the following 7 items from 1 (meaning most important) to 7 (meaning least important).	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 rating/ranking)
		20	**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 items from 1 (meaning most important) to 6 (meaning least important).	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
End labelled scales (from an end-labelled versus fully-labelled experiment)	Showcards used	21	**SATISFACTION WITH THE ECONOMY:** And on the whole, how satisfied are you with the present state of the economy in Great Britain?	7 Categories with end labels Very Satisfied and Very dissatisfied	European Social Survey, 2006
		22	**SATISFACTION WITH DEMOCRACY AND PERSONAL FREEDOM:** On the whole, how satisfied are you with the way democracy and personal freedom work in Great Britain?	7 Categories with end labels Very Satisfied and Very dissatisfied	New
		23	**FREQUENCY OF GROCERY SHOPPING:** The next item is about grocery shopping which includes food, drinks, cleaning products, toiletries and household goods. How often do you personally do grocery shopping?	7 Categories with end labels Every day and Never	New
		24	**PURCHASES OF HOT BEVERAGES:** In the last two weeks, how many teas, coffees and other hot beverages have you purchased outside the home? Please look at this card and tell me your answer.	7 Categories with end labels None and More than 25	New