ISER Working Paper Series

Mixed Modes and Measurement Error: Using Cognitive Interviewing to Explore the Results of a Mixed Modes Experiment

Pamela Campanelli
The Survey Coach

Margaret Blake
Ipsos MORI

Michelle Mackie (Gray)
Ipsos MORI

Steven Hope
University College London

No. 2015-18
September 2015

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

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

www.iser.essex.ac.uk

Non-technical summary

Surveys can be presented to participants in a variety of ways (survey modes). This includes interviewing face-to-face or via telephone, or a self-completion questionnaire on paper or on the web (internet). Previous research has shown that levels of survey participation differ by mode, but less is known about how participants' answers to survey questions differ by mode. Knowing more about the latter is important because researchers are increasingly including several different modes in the same survey and it is a problem if the answers given by participants in different modes are not totally comparable.

An experiment was carried out to look in more detail at mode differences in survey questions. Participants in a face-to-face survey of the general public were invited to take part in a follow-up experimental survey and were randomly assigned to three mode groups: telephone, face to face and web using the same questionnaire. After excluding eligible people who did not take part, there were 380 (73 percent) face-to-face, 409 (69 percent) telephone and 349 (47 percent) web participants. Following the experiment, a small group of participants from the experiment (37) were invited to take part in a cognitive interview (a methodology normally used for pre-testing survey questions). During each interview, survey questions were administered in three different modes (face to face, telephone and web). Participants were then asked to 'think aloud' about how they had come to each answer, to understand more about how the mode a question was presented in had affected how they answered. These participants were selected because they had answered questions in the experiment in ways which were related to mode effects. This is an unusual use of cognitive interviewing.

The paper presents four examples which discuss results for both the experiment and the cognitive interviews along with evidence from the literature. In the last part of the paper we examine participants' mode preferences through their comments at the end of the cognitive interviews.

Our overall conclusion is that cognitive interviewing can uncover important findings which are hidden in more traditional quantitative approaches to looking at survey methodology, while at the same time providing more detail on findings that support traditional quantitative approaches.

Mixed Modes and Measurement Error:

Using Cognitive Interviewing to Explore the Results of a Mixed Modes Experiment

Pamela Campanelli, The Survey Coach

Margaret Blake, Ipsos MORI

Michelle Mackie (Gray), Ipsos MORI

Steven Hope, UCL Institute of Child Health

Abstract

This paper explores the use of cognitive interviewing as a pre-planned follow-up to a quantitative mixed modes experiment. It describes both the quantitative and cognitive interview phases and results. The goal for both was to explore measurement error differences between (computer-assisted personal interviewing - CAPI, computer-assisted telephone interviewing - CATI and computer-assisted web interviewing - CAWI). The cognitive interviewing produced evidence that in particular circumstances, supported or challenged the quantitative results. This is illustrated through the use of five examples. In conclusion, this novel application of cognitive interviewing was useful, with implications for survey design and interpretation of quantitative findings.

Keywords: Cognitive interviewing, mixed modes, measurement error, satisficing, mode preferences, polar point, yes/no list

JEL classification: C00

Acknowledgments: 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 also wish to acknowledge the core members of the Mixed Modes and Measurement Error team: Gerry Nicolaas (Ipsos MORI) and Peter Lynn and Annette Jäckle (Institute for Social and Economic Research). Thanks must also go to the other cognitive interviewers who worked on this study and last, but not least, to all of the participants who kindly took part in the research. Without their help, none of this would have been possible.

Contact: Pamela Campanelli, The Survey Coach, 33 Military Road, Colchester, Essex CO1 2AD, United Kingdom. E-mail: dr.pamela.campanelli@thesurveycoach.com.

1. Introduction

Cognitive interviewing is traditionally thought of as a pretesting method. However, this paper explores the use of cognitive interviewing as a pre-planned follow-up to a quantitative survey that was used to collect data for a mixed modes experiment from an ESRC funded grant on 'Mixed Modes and Measurement Error'.

The ESRC funded Mixed Modes and Measurement Error study had four main components: (1) a review of the literature and development of a theoretical framework to assess the susceptibility of different types of survey questions to various mode effects, (2) the collection of new experimental quantitative data, (3) cognitive interviewing to provide a deeper understanding of the results of the new experiment, and (4) the development of a set of principles for designing questions that are portable across data collection modes. This paper focuses on how part three (the cognitive interviewing) was used as a way to explore the findings of part two (the quantitative mixed modes experiment). To meet this goal, five examples are used. The first four examples cover quantitative findings which showed mode effects and the subsequent cognitive findings which shed further light on these mode effects. These included: (1) more endorsements in a 'yes/no' list format (as opposed to a ' mark all that apply' format) in the interviewer modes than web, (2) issues with end-labelled scales (difficulty for participants in remembering the direction of the scales when there is no visual depiction of the scale, (3) issues with end-labelled scales (fewer middle categories selected by telephone participants than face-to-face or web participants) and (4) more use of middle categories for satisfaction and agree/disagree questions in web than in the interviewer modes. The fifth example looks at participants' mode preferences through an examination of their comments and opinions on the three modes of survey data collection they experienced at the beginning of the cognitive interviews.

The use of cognitive interviewing in this project is novel both from the results identified and the methodology used. This paper focuses on the results of the cognitive interviewing and a companion paper by Gray, Blake and Campanelli (2014) focused on the methodology used.

This Introduction section continues with a discussion of the different ways the general practice of cognitive interviewing has been used in survey research. This is followed by the Methods section with a description of the data collection and analysis methods for both the quantitative mixed modes experiment and the cognitive interviewing. This is followed by the five examples. All of the results are then brought together and reviewed in the Discussion section.

1.1 Background on cognitive interviewing

Since its birth in the mid-1980s, cognitive interviewing has become a well-known and commonly-practiced method for testing survey questions. It is a collection of techniques used in a semi-scripted qualitative interview, with the most popular ones being thinking aloud and verbal probing. Think aloud protocols require participants to tell the cognitive interviewer what they are thinking while they are answering a survey question.[1] Verbal probing involves pre-planned and spontaneous questions which the cognitive interviewer asks of participants in order to assess how well participants understood the questions and concepts being measured, the accuracy of their recall and judgment strategies, and how they mapped their responses into answer categories (if these are present). Both think alouds and probing can be conducted either concurrently with the administration of the survey questions, or retrospectively, after the survey itself (or sections of it) have been completed. Cognitive interviewing therefore enables researchers to examine in greater detail the question and answer process, helping to identify problems with questions and the thought processes participants go through as they answer survey questions. For a more detailed explanation of cognitive interviewing see Willis (2005) and Collins (2015).

Since the development of cognitive interviewing in the mid-1980s, it has progressed from simply implementing cognitive interviews to evaluating the ability of cognitive interviews to predict question problems in the field (see, for example, Willis and Schechter 1997; Dillman and Redline 2004; and Forsyth, Rothgeb and Willis 2004). Evaluation has also focused on implementation decisions and the techniques used during the cognitive interview itself (see, for example, Beatty, 2004; DeMaio and Landreth, 2004).

In addition, there has been a focus on extending the cognitive interview. For example, the cognitive interview has been broadened to cover the collection of background information about the participant. Cosenza and Fowler (2000) discuss a "prospective probing" procedure in which subjects talked about their situation with respect to the general topic of the interview prior to the cognitive interview. They argue that this helped the cognitive interviewer detect and resolve inconsistencies that emerged between the prospective story and how the subjects were interpreting the survey questions. Gerber (1999) and Ainsworth (2000) discuss a variation called, 'the ethnographic approach'. Ainsworth (2000) argued that survey questions may be cognitively clear but not fit the situations in participants' lives and stressed the importance of cultural variables.

[1] This is an outgrowth of the work of Simon and Ericsson (1980). In their 'protocol analysis', subjects were asked to think aloud while performing a problem solving task.

Cognitive interviewing methodology has also been extended across languages. By conducting a comparative analysis, cognitive interviews can identify problems in questions that have been incorrectly translated or that convey even subtle meaning differences in other languages (Harkness, van de Vijver and Mohler, 2003). For example, Padilla (2007) used cognitive interviewing to explore 'construct overlap' across different linguistic and cultural groups. Similarly Carrasco (2003) and Willis (2004) looked at ways to achieve equivalency across English and Spanish questionnaires and across multiple cultural groups, respectively. Other researchers have looked at the adaptations needed to use cognitive interviewing in other languages (e.g., Goerman, 2006; Pan, 2004).

Finally, the usefulness of cognitive interviewing has been shown beyond the pretesting stage of the survey process. Miller (2008) suggests that well-documented cognitive interview findings from the pretesting stage can be used after the survey fieldwork to assist in the interpretation of quantitative analytic results. Or in more rare instances, cognitive interviewing can be used purposely to understand existing quantitative findings. For example, Jakwerth, Stancavage and Reed (1999) used standardised probes developed by their cognitive laboratory staff to explore why students were leaving certain questions blank in the National Assessment of Educational Progress assessment questionnaire and Davis, Nicolas, Waters, Cook, Gibbs, Gosch, and Ravens-Sieberer (2007) used think aloud to try to understand the discrepancies between the way parents and children answered a health related quality of life questionnaire.

1.1.1 Aims of the cognitive interviewing phase of the mixed modes study

Similar to the work of Jakwerth et al (1999) and Davis et al (2007), the cognitive interviewing phase of the ESRC funded Mixed Modes and Measurement Error study followed a quantitative survey. However in contrast to these studies, it was a pre-planned phase of the project. In addition, the cognitive interviewing follow-up was intended to provide a greater understanding of how mode effects happen even if they are not directly observed and to serve as a follow-up for unusual findings from the quantitative mixed modes experiment.

2. Methods

2.1 Data collection for the quantitative mixed modes experiment

For the quantitative mixed modes experiment which preceded the cognitive interviewing phase, initial data were collected using NatCen Social Research's Omnibus survey. This survey used a probability sample of adults aged 16 and over in Great Britain and clients were able to buy questionnaire space. The survey was administered quarterly to a fresh sample of participants and 1,600 face-to-face

interviews were completed using Computer-Assisted Personal Interviewing.[2]

Prior to the quantitative mixed modes experiment, data from 15 questions were collected over two waves of the NatCen Omnibus (July to November, 2008).[3] All participants who agreed to be re-contacted were recruited for the quantitative mixed modes experiment and randomly assigned to one of three modes (face-to-face using Computer Assisted Personal Interviewing - CAPI; telephone using Computer Assisted Telephone Interviewing - CATI; and web using Computer Assisted Web Interviewing - CAWI). In the experiment, separate surveys for each of these three modes were collected by NatCen and conducted between January and June of 2009. After excluding eligible people who did not take part, there were 380 (73 percent) face-to-face, 409 (69 percent) telephone and 349 (47 percent) web participants (response rate RR5, American Association for Public Opinion Research, 2011).

The questionnaire repeated the original Omnibus survey module of 15 questions and included an additional 67 questions that had been designed to test a set of hypotheses about the causes and consequences of mixed mode effects. These 67 additional questions were classified according to type of question (including satisfaction, other attitudinal, behavioural and other factual), task difficulty and sensitivity of the question. In addition, seven different question format comparisons were tested: (1) short versus long scales, (2) rating versus ranking, (3) agree/disagree statements versus balanced forced choice questions, (4) 'yes/no' list versus 'mark all that apply', (5) branched versus non-branched scales, (6) fully-labelled versus end-labelled scales and (7) showcards versus no showcards on long lists in CAPI.

This paper uses some of the question formats which showed mode effects: 'yes/no' list, end-labelled scales, short and long satisfaction questions and agree/disagree statements. A full list of the questions used in this paper, including their actual wording and their sources, are listed in the Appendix.

2.1.1 Analysis methods for the quantitative mixed modes experiment

The quantitative analysis included only participants who had access to and used the internet. The resulting sample sizes were 282 in CAPI, 314 in CATI and 349 in CAWI. The analysis also used five

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

[3] These were to form the beginning of a small longitudinal component to the ESRC funded Mixed Modes and Measurement Error study.

control variables to correct for differential non-response bias across CAPI, CATI and CAWI.[4]

Analysis of variance with the control variables was used when the dependent variable was continuous and estimated marginal means were of interest and logistic regression with the control variables was used when the dependent variable was dichotomous.

2.2 Data collection for the cognitive interviewing phase

This section provides a condensed description of the cognitive interviewing methodology used in this project. A full discussion is found in Gray, Blake and Campanelli (2014).

Participants for this study were selected from the quantitative mixed modes experiment sample. Mode effects are typically detected at the aggregate rather than individual level. However, two aspects of participant behaviour that indicated mode differences could be identified at an individual level in the quantitative mixed modes experiment (1) agreeing to opposite statements (a typical indicator of acquiescence behaviour) and (2) misunderstanding a ranking task and giving the same ranking to all items or to all but one of the items (i.e., non-differentiation). The participants exhibiting these behaviours were more likely to have lower levels of education and income, not to be employed (or if working, in lower level occupations), to be a social renter, and to be non-white. In addition, a comparison group of participants were also included in the sample. They had higher levels of education and income, were in higher level occupations, of white ethnicity and house owners. Thirty seven participants were recruited and interviewed.[5]

Cognitive participants were first interviewed with a subset of survey questions from the quantitative mixed modes experiment.[6] Survey questions associated with unusual quantitative findings were

[4] Adjustment with 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 participants rather than directly from the general population.

[5] Thirty six interviews had been planned (eighteen who showed acquiescence behaviour; nine who had misunderstood the ranking task, and nine contrasting participants). But one interviewer had conducted six interviews but still needed a particular type of person to fulfil her quota, so an additional interview was conducted.

[6] The survey questions were administered at the beginning of the cognitive interviews because the original quantitative survey had been conducted five months before.

selected, as well as survey questions related to other findings that required further investigation. These questions were incorporated into a questionnaire with six versions because there were more issues to investigate than would fit into the preferred timing of a single cognitive interview (i.e., 10 minutes for standard survey question administration and 50 minutes for cognitive interviewing). Question issues that were considered important by the research team were explored in all six questionnaires. Question issues that were of lesser importance were present in most but not all of the six questionnaires and were randomly assigned to a questionnaire version. Each of the six questionnaire versions consisted of questions in CAPI, CATI and CAWI with mode selection based on the need to follow up on particular findings. Each participant was exposed to questions in each of the modes in that order.[7] [8] In all versions, the plan was to administer the survey question in standard quantitative fashion and mimic the modes as closely as possible. This involved the interviewer sitting with the participant face-to-face (for the CAPI component), being in a different room in the participant's home and talking over a phone (for the CATI component) and having the participant use the interviewer's laptop completely on his/her own (for the CAWI component). After participants experienced the survey questions in the three modes, there was a switch to face-to-face cognitive interviewing, using retrospective thinking aloud. A few standardised probes were used when there was a need to focus on a particular question issue. It is important to note that these cognitive interviews differed from traditional cognitive interviews. Participants were not asked about their understanding of the survey questions, but rather how they came up with their answers. This is because the purpose of this cognitive interviewing follow-up phase was to gain a greater understanding of how mode effects happen even if they are not directly observed.

The cognitive interviewing was conducted by both researchers and survey interviewers trained and experienced in using cognitive interviewing methods and together they interviewed participants in their homes in London; Essex; Manchester and Lancashire; Leeds and Yorkshire; Nottinghamshire; and Edinburgh, Scotland.

[7] This was done to ease the burden on the cognitive interviewers as the design was already complex.

[8] Although every participant experienced the three modes, participants were only asked a given survey question once. This was accomplished by taking a set of questions with a particular format (e.g., agree/disagree), level of sensitivity and level of difficulty and administering some in one mode and the rest in a different mode across the different versions of the questionnaire. In a few instances, newly written questions designed to be equivalent to the original survey questions (in terms of format, sensitivity, difficulty and type of question) were used in one mode and the original question in a different mode.

2.2.1 Analysis method for the Cognitive Interviewing phase

All cognitive interviews were audio-recorded with participant consent and then transcribed. The transcribed information was summarised to capture key points from participants and the summarised information was entered into the qualitative data management program, 'Framework', which was used for analysis. The analysis was divided among the three researchers with a fourth researcher overseeing the project. Each researcher read through all of the summarised transcript information for his/her survey questions, checking the verbatim transcripts where needed. The Framework program allowed analysis of different sub-groups of participants based on further variables, such as mode and the participant's answer to the survey question at the start of the cognitive interview. The goal was to look for anything in the response process that was seen to differ by mode, including participant satisficing[9]. The themes that were identified were then written down and compared across modes.

As a qualitative method, the use of quantification in reporting cognitive interview findings is usually discouraged. However, 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 the use of occasional vague quantifiers like 'a few' or 'most'.[10]

3. The five examples

In this section we present five examples to illustrate the cognitive interviewing insights into the quantitative findings. The first four draw from the quantitative findings of mode differences and each sub-section includes (1) a review of the literature, (2) how the question issue was explored in the quantitative mixed modes experiment, (3) the results of the quantitative mixed modes experiment, (4) the specific plan for the cognitive interviewing and (5) the cognitive interviewing results. The last example was unique to cognitive interviewing with no quantitative component and thus only includes a review of the literature, the cognitive interview design and the cognitive interviewing results.

3.1 Example 1: The 'yes/no' list format

3.1.1 Literature review on 'mark all that apply' versus 'yes/no' list

It is common practice for the 'mark all that apply' format to be used in web and mail surveys. However by necessity, telephone surveys usually assign a 'yes/no' option to each category. It is common in the

[9] A review of the literature on satisficing is in Section 3.4.1 and a discussion of the cognitive interview identification of satisficing is in Section 3.4.5.

[10] The issue with any kind of quantification is how the cognitive interviewing results might generalise to the bigger population and how the results would differ with a different sample.

UK to use ‘mark all that apply’ in face-to-face interviews (mainly with showcards and interviewer probing), as it is generally argued that a series of ‘yes/no’ questions would take longer and be tedious for interviewers and participants.

Research suggests the two formats are not functionally equivalent. Sudman and Bradburn (1982) were the first to suggest avoiding the ‘mark all that apply’ format. This was because of problems in interpreting the results. “*With [mark all that apply], it is difficult to interpret what the absence of a check mark means. While the presence of a check mark indicates a positive instance, the omission of it might indicate that in fact the adjective does not apply, or that participants did not notice [the response option] because they were hurrying over the list, or that they were not sure whether it would apply*” (p. 168). And they suggested a better alternative would be to have a separate answer category for each option such as “*’yes/no,’ ‘applies/does not apply’, ‘true for me/not true for me,’ and the like*” (p. 149).

Mitofsky and Edelman (1993) mention that the ‘yes/no’ format produced a higher endorsement of behaviours than ‘mark all that apply’ in their 1993 AAPOR conference paper, but this was never published. Rasinski, Mingay, and Bradburn (1994) conducted the first published experimental test on the topic. They compared ‘yes/no’ versus ‘mark all that apply’ formats with a paper questionnaire given to high school seniors as part of a field test for round three of the National Education Longitudinal Study. In addition to the two formats, they also varied the order of the answer lists (switching the first and last halves). They found that for the same item, the percentage of ‘yes’ responses in the ‘yes/no’ format was higher than the percentage choosing the item in a ‘mark all that apply’ format. While they were unable to identify whether the increased number of responses was valid or reflected over-reporting, they reported a tendency for the set of options at the top of the list to be more likely to be accepted and this was particularly true for the ‘mark all that apply’ format.

Using Washington State University students as participants, Smyth, Dillman, Christian and Stern (2006) expanded on the work of Rasinski et al (1994) by using web as well as paper self-completion and attitudes as well as behaviours. Recruiting participants from the same population, Smyth, Christian and Dillman (2008) extended the research further to include telephone surveys. Compared to ‘mark all that apply’, Smyth et al (2006) found higher endorsements in the ‘yes/no’ format and evidence of better quality through more thoughtful answers and reduced primacy effects and no evidence for ‘yea-saying’ or increased missing data. Smyth et al (2008) found the percentage of endorsements in the ‘yes/no’ format were the same for web and telephone modes.

3.1.2 How the literature issues were explored in the quantitative mixed modes experiment

One of the seven question experiments in the ESRC funded Mixed modes and Measurement Error study was a comparison of 'mark all that apply' and a 'yes/no' list. The 'yes/no' versions of these questions included an eight question series on 'ways to reduce poverty' and an eight question series on 'things you like about your neighbourhood'. The purpose of these questions were to replicate the research by Smyth et al (2006) and Smyth et al (2008) on their 'mark all that apply' versus 'yes/no' research and extend it by including a comparison with face-to-face interviewing in addition to telephone and web and through using a probability sample of the general adult population rather than university students.[11] More details on the specific findings of Smyth et al (2006, 2008) and the eight specific hypotheses addressed in the mixed mode experiment are given in Nicolaas, Campanelli, Hope, Jäckle and Lynn (under review).

3.1.3 Results of the quantitative mixed modes experiment

This paper focuses on just two findings from the eight hypotheses of Nicolaas et al (under review). First, in contrast to Smyth et al (2008), Nicolaas and her colleagues did find a difference in the mean number of items endorsed in the 'yes/no' format between CAWI and CATI. Second, the mean number of items endorsed in the 'yes/no' format for CAPI (not used by Smyth et al (2008) also differed from CAWI. More specifically, Table 1 shows that for the poverty series of questions, both CAPI and CATI participants had higher mean endorsements of the 'yes' category than CAWI participants. For the neighbourhood series of questions, CATI participants had higher mean endorsements than CAPI or CAWI participants.

Prior to the cognitive interviewing, the research team discussed several possible reasons for these findings which showed unexpected mode differences. (1) Could it be because CAWI participants were seeing all the 'yes/no' items at once? No, as the web set up was with only one item per page. (2) Could this be due acquiescence (meaning 'yea saying') in the interview modes, or to unexpected social desirable answering in the interviewer modes? It was not clear.

[11] Smyth et al (2006) do not mention how students were selected. Smyth et al (2008) used a random sample of students.

Table 1: Mean number of 'yes' answers in the 'yes/no' format and mode of data collection[§]

Question Series	Format	Mode of Data Collection	n	Estimated mean number of items endorsed♦	Mean Difference of Pairwise Comparisons	T test statistic
Poverty	Yes/No	CAPI	110	5.9	0.8	3.9***
		CATI	113	5.6	0.5	2.8**
		CAWI	178	5.1		
Neighbourhood	Yes/No	CAPI	141	5.7	0.4	1.9
		CATI	142	6.1	0.5	2.5*
		CAWI	166	5.6		

♦ Adjusted for control variables

* $p < .05$, **$p < .01$, ***$p < .001$.

§ This table is originally Table 4 from Nicolaas et al (under review).

3.1.4 Specific plan for the cognitive interviewing phase for these questions

As discussed in Section 2.2, only some of the quantitative questions could be used in the cognitive interviewing. It was decided by the research team to focus on the poverty question series because CAPI was an extension to the work of Smyth et al (2008) and CAPI differed from CAWI on the poverty question series results. Six of the eight questions from the series were used. These are highlighted in grey in the Appendix and include 'education for children', 'reducing discrimination' and 'increasing income support' which showed mode differences between CAPI and CAWI participants, and 'increasing pensions', 'redistribution of wealth' and 'investing in job creation' which showed no mode differences. The 'mode difference' versus 'no mode difference' questions were allocated across the questionnaire versions. Two versions had 'mode difference' questions in CAPI and the 'no mode difference' questions in CAWI, two versions had the reverse, and two versions (not discussed in the paper) explored other issues.

3.1.5 Cognitive interviewing results

For each of the six questions, themes were placed under one of four heading based on how the participant had answered the survey questions in the cognitive interviews. These consisted of CAPI-YES, CAPI-NO, CAWI-YES and CAWI-NO. Figure 1 shows a reduced example of this using the 'reducing discrimination' item.[12]

[12] Initial analysis looked for any differences in themes between the three 'mode difference' questions and the three 'no mode difference' questions and by CAPI and CAWI, but no clear patterns emerged.

Using the example of Figure 1, firstly it can be seen that a lot of similar themes appeared in both modes. One exception was the instance of 'possible' and 'clear' satisficing[13] in the CAWI mode. Looking across think alouds for all of the poverty questions, satisficing was more common in CAWI than CAPI. Curiously, almost all of these satisficing responses were in the 'yes' category. This could indicate that, in the absence of a middle category, 'yes' is an easy answer. As shown in Figure 1, one participant specifically said she *"erred on the side of 'yes'"* (JF05: Female, 30 to 39, postgraduate degree, high income, White British).

Secondly, the think aloud data themes in Figure 1 show that although participants chose a 'yes' or 'no' to the survey question, some were really in a middle ground (saying "possibly yes", "it depends", "hopefully yes"). Looking across the poverty questions there was a clear pattern that participants who described their views during the cognitive debriefing as in the 'middle ground' were much more likely to have chosen a 'yes' to the survey question than a 'no'. Of these participants in the middle ground, more were in CAPI than were in CAWI.

Thirdly, (not shown in Figure 1), participant comments suggested that two questions ('income support' and 'redistribution of wealth') could be sensitive. For example, on the 'increasing income support' question, one CAPI participant commented, "*it's a hard one to say 'no' . . . what's somebody going to think me saying no*". (SA05: Female, 40 to 49, CSE, O level or A level, low income, White British). On the 'redistribution of wealth' question, a CAWI participant commented, "*I don't feel that those that are out and earning money at a decent level should be the ones to pay to support that, and that sounds really awful. It's an awful viewpoint, but I think there is part of that in there*" (AR06: Female, 30 to 39, first degree, high income, White British). Both participants gave the non-socially desirable response of 'no'.

So overall, these findings raise questions about what a 'yes' response means in the 'yes/no' format. The 'yes' answers (as opposed to the 'no' answers) included more participants who gave a satisficing response and in contrast, more participants who had thought about the question and therefore would have liked an option in between 'yes' and 'no' such as "it depends" or "possibly, yes". The 'yes' answers for the two questions which could be sensitive may result from participants giving a socially desirable answer. The cognitive interview findings also suggest that 'yes' answers due to satisficing may be more likely to occur in CAWI, while 'yes' answers from thoughtful participants who expressed

[13] For how the cognitive interviews identified satisficing, see Section 3.4.5.

Figure 1: Themes of participants' think alouds: Justifications for answers to the poverty item on reducing discrimination

Survey Question Coded As	Participants' thoughts during cognitive debriefing	CAPI	CAWI
YES	Clear reasons for choosing 'yes'	With respect to workplace • General about getting a job • Women still suffer inequality General comments about discrimination: • Discrimination due to race and colour in general – they could lose out	With respect to workplace • If job goes to someone else *"person who's discriminated against might be left in poverty"* Personal experience • Yes because self-experience of being an agency worker
	Qualified or dependent reasons for choosing 'yes'	Possibly • Possibly, but there are other ways which are more effective It depends • Depends on what's being done and how • Would have liked a category between 'yes' and 'no'. Hope it will • Not sure but hopes it will No discrimination exists • Yes, but people are not *"allowed to discriminate as lots of laws to prevent"*	
	Problems with response process		R didn't understand what was meant by discrimination or what type of discrimination Possible satisficing * • Less easy to answer but erred on the side of 'yes' Clear satisficing * • *"I'm not very sure really how, how I've come to that answer"* • *"To tell you the truth, I just clicked it"*

Figure 1 continued

Survey Question Coded As	Participants' thoughts during cognitive debriefing	CAPI	CAWI
NO	Clear reasons for choosing 'no'	Can't see any connection between discrimination and poverty • Can't see why it would • No influence on poverty at all • Have never connected discrimination with poverty No discrimination exists • Doesn't see gender discrimination in jobs • No discrimination in benefit systems	Can't see any connection between discrimination and poverty • *"I don't think discrimination has got anything to do with poverty*" • *"If you're a different race or ethnic background, I don't think it automatically means that you are poverty-stricken*" No discrimination exists • Don't honestly believe there is discrimination in the work place nearly as much as there used to be • A 'no' answer is sensitive
	Problems with response process		R didn't understand what was meant by 'discrimination'

the desire to have an option between 'yes' and 'no' during the cognitive debriefing may be more likely to occur in CAPI. Although the two participants who pointed out the sensitivity of two questions did not give a socially desirable answer, the literature suggests that the socially desirable answers of 'yes' should be more common in CAPI than CAWI (see, for example, Kreuter, Presser and Tourangeau, 2008).

Interestingly the latter two findings are in line with the quantitative mixed mode results for the poverty questions where more 'yes' answers were seen in CAPI than CAWI, whereas the first more general finding about satisficing is at odds with the quantitative mode results as it suggests more 'yes' answers in CAWI. The effect of these different types of participant behaviour patterns on the quantitative data would depend on the relative frequency of occurrence of each behaviour pattern in the population under study.

The cognitive interview findings spurred further quantitative analysis as reported in Nicolaas et al (under review). As discussed in Section 3.1.3, it was unclear from the quantitative analysis whether the higher endorsements of the 'yes' category in the interviewer modes were due to acquiescence ('yea-saying') or socially desirable answering. The cognitive interviewing found no evidence for

acquiescence ('yea-saying') on these questions.[14] Although the cognitive interviewing suggested two of the poverty questions could be sensitive, more importantly it also suggested some participants answering 'yes' to the poverty questions were being more thoughtful. Without the cognitive interviewing, the research team would not have considered this latter finding. Further quantitative analysis also ruled out social desirability as a cause in favour of participant's more thoughtful answers.

3.2 Example 2: Visual presentation of end-labelled scales and confusion over scale direction

Where participants are given a visual response scale, practice tends to differs in relation to whether only the two ends of the scale are labelled or whether the scale is fully-labelled (i.e. all points are named). Practice tends to differ according to the mode being used, with fully-labelled or end-labelled scales being used for CAPI and CAWI and end-labelled scales tending to be used for CATI questions with five or more categories.

Krosnick and Fabrigar (1997, p.149) suggest that fully-labelled scales, as opposed to end-labelled scales, are easier to answer *"because people rarely express complex conceptual meaning in everyday conversation via numbers"*. In addition, Krosnick and Presser (2010, p 271) suggest that answering a number scale requires the participant to *"first generate a verbal definition for each point and then match these definitions against their mental representation of the attitude of interest"*. In contrast, the verbal labels of a fully labelled scale clearly help to clarify the meaning of the scale categories. Krosnick and Fabrigar also conclude that *"fully labelled scales are more reliable and valid than end-labelled scales"* (p. 152).[15]

If one does use an end-labelled scale, visual presentation can vary. Christian and Dillman (2004) used three pairs of questions that had the same stem for the question, but varied how the answer spaces were displayed for a five-point number scale with end labels in a paper self-completion task. In what they labelled as the 'graphical' version, the answer categories could be seen (i.e., the categories had a linear layout with numbers and boxes for each category). For the other version, the

[14] It is worth noting that cognitive interviewing is able to detect acquiescence. Campanelli, Gray, Blake and Hope (2015) discuss how cognitive interviews can identify clear and possible acquiescence in agree/disagree questions.

[15] They also point out that this assumes that researchers have selected labels with *"relatively precise meaning"* and that reflect *"equal intervals along the continuum of interest"* (p. 152).

participant was only given a number box in which to insert his or her preferred answer. The latter would simulate a telephone interview where typically no visual aids are used.

3.2.1 Literature review

Christian and Dillman (2004) found that participants had confusion with the number box version of the scale. In 10 percent of cases, they saw erasure marks and answer changes where participants had changed their answers from one to five, two to four, etc. suggesting that participants had forgotten the direction of the scale. This was only observed in 1 percent of cases who received the graphical version. They also investigated whether participants had confused the direction of the scale, but had not changed their answers. To do this they took both versions of the three questions and looked at correlations with 13 other questions on the same topic in the questionnaire. The expectation for this analysis of the three questions with the 13 other questions, was that each of the resulting correlations would be lower for the number box version than the graphical version if participants had reversed the direction of the scale and not been aware of their error when answering. Christian and Dillman (2004) found evidence of further 'direction of scale' confusion as the mean correlation for the number box version was 0.14 as opposed to 0.24 for the graphical version. Examining individual correlations between questions in both formats showed that in 89 percent of the comparisons, the correlation for the graphical version was higher than the respective one for the answer box version.

Through the use of client-side 'paradata'[16], Stern (2008) focused on detecting when and where participants change their answers to survey questions. As part of his paper, he replicated the work of Christian and Dillman (2004) by including a graphically displayed end-labelled scale with numbers and a number box version. He extended their research by also including a fully-labelled scale and a graphically displayed end-labelled scale without numbers. He used the phrase 'reciprocal change' for instances where it was clear that the change was due to the participant forgetting the direction of the scale. He concluded that "*receiving the fully labelled version significantly reduced the odds that the participant would make a reciprocal change (exp^B 0.31; p =0.001), whereas receiving the number box version increased the odds of making a reciprocal change (exp^B 1.68; p =0.05)* (p. 386)."

Stern's research also brought an interesting twist regarding the direction of code numbers. Christian and Dillman (2004) had believed that the greater number of participants who changed their answers

[16] See Heerwegh (2002; 2003).

to the number box version, as opposed to the visual version, was due to non-visual versus visual scale differences. But Stern's (2008) results suggested that this may not be a full interpretation. He had two versions of the visually (graphically) presented end-labelled scales; one with code numbers and one without code numbers. Although these two versions were not statistically different, the version with numbers had more reciprocal changes than the one without as seen by the higher odds ratios (exp^B =1.36 and exp^B = 1.05, respectively). He speculated that this may be due to the fact that the number version had the numbers displayed in a counter intuitive way with '1' used as an indicator of the highest level of satisfaction and '5' as the lowest.[17]

3.2.2 How the issues in the literature were explored in the quantitative mixed modes experiment

The quantitative mixed mode experiment included a small experiment to explore further the results of Christian and Dillman (2004) and Stern (2008). On a question about the 'importance of money', the experiment compared a 'graphical' 1 to 10 scale with a 'number box' version of the same question. In line with Stern (2008) and Dillman and Christian (2005), the question was set up so that 10 was the strong positive category 'very important'. (It is also traditional for 0 to 10 and 1 to 10 scales to start with the lowest value as the most negative). The experiment was only set up in CAWI as Christian and Dillman (2004) and Stern (2008) were all exploring the aspect of visual design changes in self-completion. If reciprocal changes are purely caused by the direction of the numbering, then there should be no reciprocal changes in this experiment.

The experiment also included four questions in both end-labelled and fully-labelled format (both with code numbers; two of the types used by Stern, 2008). These included two satisfaction questions ('satisfaction with democracy and personal freedom' and 'satisfaction with the state of the economy') and two behavioural frequency questions ('frequency of grocery shopping' and 'amount

[17] Dillman and Christian (2005) did a further experiment where they varied whether '1' was most positive or '5' was most positive (Christian and Dillman, 2004, had used one as most positive). Although the percentage of reversals in not indicated, answers in the number box format were significantly more positive when '5' was the most positive category rather than '1'. But this is in contradiction to the heuristics given by Tourangeau, Couper, and Conrad (2004) which suggest that the leftmost or top item is seen as the first and that 'up means good'. Also in many aspects of life 1st is regarded as the best (e.g., getting a 1st in your degree, being the 1st across the finish line, etc.)

of hot beverages purchased outside the home')[18]. The closest replication of Christian and Dillman (2004) is to look at the end-labelled version of the scales and to compare CATI participants who were only asked to give a number (i.e., similar to using the number box format) with CAPI with showcard and CAWI participants who could see a 'graphical' version.

3.2.3 Results of the quantitative mixed modes experiment

Reciprocal changes measured with the correlation between the 'importance of money' and other related questions

The experiment had collected paradata to look at the length of time it took participants to answer each question in CAPI, CATI and CAWI, but we did not have paradata to look at reciprocal changes that participants may have made. We could assess whether participants had forgotten the direction of the scale and failed to change their answer by using the method Christian and Dillman (2004) had employed. For the format experiment on 'importance of money', reciprocal changes were studied by taking both versions of the question and looking at correlations with nine other questions expected to correlate with it.

Table 2 shows the results for the 'importance of money' question. There is some limited evidence that despite the direction of the scale being intuitive, there may still be reciprocal changes on the number box version. Four of the nine questions are significant and in the direction expected if reciprocal changes are happening (see the four items with a 'plus') and two are significant and not in the expected direction (see the two items with a 'minus'). The cognitive interviews showed that these latter two questions were confusing to participants. This was because of the use of the negative word 'rarely' in the statement part of these agree/disagree formatted questions (see Hope, Campanelli, Nicolaas, Lynn and Jäckle, 2014), suggesting that there are other factors which make a scale difficult to understand.

[18] In survey practice it would be rare for behavioural frequency questions to appear in end-labelled format, but a larger part of the research grant was ensuring that both behavioural and factual questions were tested as well as satisfaction and attitudinal questions.

Table 2: Differences between Spearman rank correlation coefficients for non-visual and visual formats for 'importance of money' by other related questions[§]

	Non-visual correlation coefficient n=183	*Visual correlation coefficient n=166*	*P value for test of the difference between the two correlation coefficients*
How well managing financially *(Expect positive correlation)*	0.11	0.02	Not significant
Financial expectation a year from now *(Expect negative correlation)* +	-0.06	-0.16	*p<.05*
Monthly spend on eating out *(Expect negative correlation)*	0.11	0.09	Not significant
Would rarely read small print before a financial decision *(Expect positive correlation)* -	-0.23	-0.01	*p<.01*
Would do a lot of research before a financial decision *(Expect negative correlation)*	0.01	-0.01	Not significant
Would rarely talk to a financial advisor before a financial decision *(Expect positive correlation)* -	0.02	-0.12	*p<.05*
Would definitely talk to family/friends before a financial decision *(Expect negative correlation)* +	0.09	-0.10	*p<.05*
Housing type *(Expect positive correlation)* +	-0.21	0.17	*p<.01*
Monthly spend on leisure activities *(Expect negative correlation)* +	0.22	-0.09	*p<.01*

\+ Significant and showing expected pattern (i.e., if reciprocal changes are present for the non-visual version)

\- Significant and not showing expected pattern

§ This table is originally Table 6 from Lynn, Hope, Jackle, Campanelli and Nicolaas (2012), but title, headings and footnotes have been edited to be in keeping with this paper

Reciprocal changes measured with the correlation between four end-labelled scales and other related questions

In the questionnaire, there were very few other questions which might be related to the four end-labelled scales. As shown in Table 3, only the pattern of correlations between 'satisfaction with the economy' and 'how managing financially these days' shows the expected pattern (i.e, lower correlations for CATI because of no visual aid and the possibility of reciprocal changes). A further examination of the table suggests that other factors such as mode differences could be at work. For example, the sensitive question 'people in Britain are too ready to criticise' has the highest correlations in CAWI where question sensitivity due to interviewers' presence is removed.

Table 3: Differences between Spearman rank correlation coefficients for CATI (non-visual), CAWI (visual) and CAPI (visual) across four end-labelled scales by other related questions

	How well managing financially			People in Britain are too ready to criticise		
	CATI n=161	CAWI n=183	CAPI n=135	CATI n=161	CAWI n=183	CAPI n=135
Satisfaction with democracy and personal freedom *(Expect positive correlation)*	.22	.10	.02	.07	.15	.08
Satisfaction with the economy *(Expect positive correlation)*	.07	.15	.16	-.01	.16	-.12
	How well managing financially			How frequently go to pub		
	CATI n=160/ 161	CAWI n=183	CAPI n=135	CATI n=160/ 161	CAWI n=183	CAPI n=135
Frequency of grocery shopping *(Expect positive correlation)*				-.01	.00	.16
Amount of hot beverages purchased outside the home *(Expect negative correlation)*	-.16	-.12	-.05	-.28	-.09	-.24

3.2.4 Specific plan for the cognitive interviews for these questions

The cognitive interviews focused just on the four end-labelled scales and were set up specifically to explore the findings of Christian and Dillman (2004) and Stern (2008) with respect to participants being more likely to forget the direction of the scale in the number box version, in this case the telephone version. In the initial 'survey question' part of the cognitive interview, participants either received 'satisfaction with the economy' and 'frequency of grocery' questions in CAPI with showcards and the 'satisfaction with democracy and personal freedom' and 'amount of hot beverages' questions in CATI (in three versions of the questionnaire) or the reverse (in three versions of the questionnaire). The cognitive interviewers were instructed to ask the CATI participants, *"How confident or unconfident were you that you remembered the direction of the scale, that is, which label went with which number?"*

3.2.5 **Cognitive interviewing results**

Although not all participants were asked ask the confidence probe described in Section 3.2.4[19], the general answer from those who were probed was 'confident'.[20] A more useful approach came from reviewing the cognitive think alouds themselves. Here it was clear that a few participants actually did confuse the direction of scale on the end-labelled questions.

3.3 Middle category issues in aural end-labelled scales

3.3.1 Literature review

Dillman and Christian (2005) show that end-labelled and fully-labelled scales do not produce the same results. Their findings suggest that participants give more positive answers to fully-labelled scales than to end-labelled scales. In addition, although not discussed in their paper, it can be seen from their results Table 1c (p. 48 of their paper) that there is greater use of the middle response option in end-labelled scales (with a difference of 21 percent between the two formats for question one and 22 percent for question two).

Aural questions are more difficult than visual ones (Schwarz, Strack, Hippler and Bishop, 1991). *"Visual display of response options provides cues and allows participants to review the options at their own speed, in the order they choose and to re-read options. If participants take full advantage*

[19] This was due to cognitive interviewer error.

[20] Asking about confidence may not have been an optimal probe. Oksenberg and Cannell (1989, p.26) noted that: *"Respondents did not appear to doubt their own, often mistaken, interpretations" of a survey question."*

of these opportunities, then they should be better able to understand the response task as a result" (Lynn et al, 2012, pp. 1-2).

3.3.2 How the issues in the literature were explored in the quantitative mixed modes experiment

The four end-labelled scales ('satisfaction with democracy and personal freedom', 'satisfaction with the economy', 'frequency of grocery shopping' and 'amount of hot beverages purchased outside the home') were compared to their fully-labelled counterparts in terms of 'positiveness' and middle category effects.

To look at the visual versus aural issues, we compared the response distributions of the four end-labelled scales between CAPI with a showcard and CATI, thus keeping the presence of the interviewer constant.

3.3.3 Results of the quantitative mixed modes experiment

Results of the comparison of the end-labelled and fully-labelled scales show that 'satisfaction with democracy and personal freedom' replicates the work of Dillman and Christian (2005). The fully-labelled scale shows significantly more positive responses and the end-labelled scale shows significantly more middle category choices.[21] The two behavioural frequency questions: 'frequency of grocery shopping' and 'amount of hot beverages' showed the same pattern as the 'satisfaction with democracy and personal freedom' question. In both cases, the participants in the fully-labelled version were significantly more likely to choose categories from the beginning, rather than the end, of the scale and participants in the end-labelled version were significantly more likely to choose the middle category. These findings are documented in Campanelli, Gray, Blake and Hope (2015) [22]

[21] The pattern is not replicated with 'satisfaction with the economy' question. At the time the question was asked, the UK economy was experiencing an economic downturn and the majority of responses were clustered around the negative side / end of the scale (i.e., dissatisfied). There were no differences between formats in the selection of the middle category for this question.

[22] Using cognitive interviewing, Campanelli et al (2015) showed that despite the similar response distribution between the 'satisfaction on democracy and personal freedom' and the behavioural frequency questions, the ways in which participants were processing the two types of tasks were very different.

The results of the visual/aural comparison between CAPI with a showcard and CATI using end-labelled scales are in Table 4. The results show that there was significantly less use of the middle category in CATI than CAPI for two of the four questions ('satisfaction with democracy and personal freedom' and 'frequency of grocery shopping'), perhaps suggesting CATI participants found it more difficult to choose a middle category without having a visual aid.

Table 4: Effect of Visual Presentation of End-labelled Scales on Response Distribution §

	Middle categories		Chi-square and level of significance ♦
	CAPI (with show card)	CATI	
Satisfaction with democracy and personal freedom	34.5	18.4	$\chi 2(1)=12.9$; $p<.001$
Satisfaction with the economy	9.0	7.2	Not significant
Frequency of grocery shopping	29.6	13.0	$\chi 2(1)=16.7$; $p<.001$
Amount of hot beverages purchased outside the home	13.3	13.7	Not significant

♦ The same conclusions were made when logistic regression with the control variables was used to adjust for differences in nonresponse bias across modes.

§ This table is originally from Table 9, Lynn et al, 2012.[23]

3.3.4 Specific plan for the cognitive interviews for these questions

As described in Section 3.3.1, we had assumed that end-labelled scales would be more difficult in CATI than CAPI due to the lack of a visual aid. The 'survey question' part of the cognitive interviews focused on CAPI and CATI. Within the four end-labelled scales, two of the four questions were administered in CAPI and two in CATI and this was reversed for half of the participants.

[23] Note that the figures for 'amount of hot beverages purchased outside the home' do not correspond to Table 9, Category 4 (Lynn et al, 2012). This question ranges from zero to six whereas the other items range from one to seven so the wrong category has been designated as the middle category. The entry in Table 4 of this paper is correct.

3.3.5 Cognitive Interviewing Results

In both CAPI and CATI and on both 'frequency of grocery shopping' and 'satisfaction with the economy', there were participants who spontaneously mentioned that they would have liked to have had labels on all the categories. It is interesting that this is true for both an easy satisfaction question as well as for a difficult behavioural frequency one. Here is an example:

- *"I was a bit stuck, to be fair, on the number thing . . . they could have written something other than giving it scores of numbers, they could have written, you know, very satisfied, dissatisfied, blah, blah, blah, would have been easier than picking a number because I don't understand, where does four, what's four, what's three.*" (SA01: female, 30 to 39, secondary school, employed, low income, White British).

In addition during the cognitive debriefing, participants spontaneously mentioned issues when completing the survey questions in the telephone mode. This was true across all of the questions. The majority of participants commented that it was much more difficult without visual aids. For example:

- *"I think I was just trying to work it out in me head there and then, it, it can be quite hard on the telephone I think."* (JF04: Female, 50 to 59, higher education below degree level, medium income, White British)
- *"I'm visual, so I like to see things in front of me."* (PC03: female, 20 to 29, postgraduate degree, employed, medium income, White British)

The cognitive interviewing also specifically asked all participants about the CATI questions: *"How easy or difficult would it be for you to choose the middle category if you wanted to?"* Roughly equal numbers said that finding the midpoint would be difficult as opposed to easy. However, when then asked to give the actual number of the midpoint, about two-thirds of the cognitive interview participants did this incorrectly.

Thus the cognitive interview findings suggest that participants do have difficulty in using end-labelled scales, particularly without visual aids. This is particularly true in the case of finding the middle category.

3.4 Example 4: Choosing middle categories for less that optimal reasons

3.4.1 Literature Review

Satisficing

Since Krosnick (1991) introduced Herbert Simon's term, 'satisficing', to survey research, it has become a popular way to understand response effects in surveys. It is based upon the assumption that optimal question answering involves doing a great deal of cognitive work. So depending on the participant's ability, the participant's motivation and the difficulty of the task, participants may decide to take a less than optimal shortcut to their answer.

Krosnick (1991) also makes a distinction between 'weak' and 'strong' satisficing. Weak satisficing is where participants take shortcuts but do not abandon any of the major response processes (e.g., comprehension, retrieval, judgment and response from Tourangeau, 1984). Strong satisficing is where participants select an answer *"without reference to any internal psychological cues specifically relevant to the attitude, belief, or event of interest. Instead, the participant may look to the wording of the question for a cue, point to a response that can be easily selected and easily defended if necessary. If no such cue is present, the respondent may select an answer completely arbitrarily"* (Krosnick and Presser, 2010)

Differential satisficing by mode

Some research on mode effects suggests that the likelihood of satisficing is greater in self-administration modes than in interviewer-administered modes. For example, the internet allows participants to multitask and quickly skip from one topic to the next. *"This in turn may lead to more superficial cognitive processing, more top of the head answers, and more satisficing in responding to survey questions"* (de Leeuw, 2005, p. 244). In contrast interviewers control the survey process and more channels of communication are available in interviewer modes. A well-trained interviewer can also motivate participants to produce complete and accurate answers (e.g., through probing) and reduce the difficulty of the task by offering support and providing explanations of what is needed (Hope et al 2014; Skjak and Harkness, 2003). Furthermore, there is evidence that satisficing is more likely to occur in telephone interviews than face-to-face interviews (Jordan, Marcus and Reeder, 1980; Holbrook, Green & Krosnick, 2003).

Middle category satisficing

Kalton et al (1980), summarising across various studies, concluded that 15 to 49 percent of participants chose the middle category when offered, but would not have volunteered it if it was not offered. Krosnick (1991) suggests that choosing a middle category when one has a different opinion is a type of strong satisficing. Krosnick and Presser (2010) suggest that the conditions thought to foster satisficing are low education (a proxy for low cognitive skills) and low attitude strength. Reviewing several authors they suggest that attraction to middle alternative was unrelated

to educational attainment and the results were mixed with respect to low attitude strength. However, this does not rule out strong satisficing on middle categories for those who have low motivation to do the survey, or who have other reasons to satisfice.

3.4.2 How the issues in the literature were explored in quantitative mixed modes experiment

The mixed modes experiment was designed to look at several sources of satisficing. Hope et al (2014) explored satisficing which may occur in self-completion modes without the positive aspects of the presence of the interviewer. Here we look at the choice of the middle category. Without the motivation of an interviewer, participants may satisfice by choosing the middle option when they have a positive or negative opinion. All of the questions with middle categories were explored including 12 five category agree/disagree questions, and two satisfaction questions in three versus seven category format.[24]

3.4.3 Results of the quantitative mixed modes experiment

Table 5 shows how the decision to choose a middle category varied by mode. These results indicated CAWI participants were more likely to choose the middle category than CAPI and/or CATI participants on both the long and short satisfaction questions (with three of the four questions showing significant results at $p < .10$ or less) and on the 12 agree/disagree items (with seven of the 12 showing significant results at $p < .05$ or less).

[24] Participants' use of middle categories on end-labelled scales was covered in Example 3.

Table 5: Evidence for more use of the middle category by CAWI as opposed to CAPI/CATI participants [§]

Question Number	Item Topic	Mode comparison, odds ratio from logistic regression and level of significance
N43	3- point satisfaction with waste and recycling collection	CAWI>CAPI, OR=1.66, p<.10 CAWI>CATI, OR=2.26, p<.01
N43	7-point satisfaction with waste and recycling collection	CAWI>CATI, OR=1.89, p < .10
N44	3-point satisfaction with street cleaning	CAWI>CAPI, OR=2.08, p<.05 CAWI>CATI, OR=3.71, p<.001
N44	7- point satisfaction with street cleaning	No significant differences by mode
N35	Agree/Disagree question: Neighbourhood not a bad place	CAWI>CAPI, OR=2.67, p<.01 CAWI>CATI, OR=1.83, p<.05
N36	Agree/Disagree question: More properties in bad state of repair	CAWI>CATI, OR=1.86, p<.10
N37	Agree/Disagree question: Not suffer from litter, dog mess and graffiti	CAWI>CAPI, OR=1.75, p<.05 CAWI>CATI, OR=1.58, p<.05
N38	Agree/Disagree question: More properties that are well kept	No significant differences by mode
FM64	Agree/Disagree question: Financial decision: Rarely read the small print	CAWI>CAPI, OR=2.44, p<.01 CAWI>CATI, OR= 2.21, p<.01
FM65	Agree/Disagree question: Financial decision: Do a lot of research	CAWI>CAPI, OR=4.15, p<.001 CAWI>CATI, OR=3.09, p<.001
FM66	Agree/Disagree question: Financial decision: Rarely talk to financial advisor	CAWI>CAPI, OR=2.03, p<.01
FM67	Agree/Disagree question: Financial decision: Definitely talk to family and friend	No significant differences by mode
N52	Agree/Disagree question: Would worry if mental health patients lived in neighbourhood	CATI> CAWI, OR=1.43, p<.05
N53	Agree/Disagree question: Mental health patients have just as much right to live in neighbourhood	CAWI>CAPI, OR=2.01, p<.001 CAWI>CATI, OR=2.37, p<.001
N54	Agree/Disagree question: Would worry if former prisoners lived in neighbourhood	No significant differences by mode
N55	Agree/Disagree question: Former prisoners have just as much right to live in neighbourhood	No significant differences by mode

[§] This table is originally from Tables 4a and 4b, Hope et al (2014).

3.4.4 Specific plan for the cognitive interviews for these questions

Although not part of the original cognitive interviewing analysis plan, it became clear that cognitive interviewing was an excellent tool for identifying instances of satisficing and so we made detecting instances of satisficing an additional goal.

3.4.5 Cognitive interviewing results

In the analysis of the cognitive interview think alouds, we made a distinction between clear satisficing, possible satisficing and no satisficing on middle category choices.[25] The following quotes illustrate our distinctions between clear and possible satisficing.

Examples of clear satisficing included participants' replies to the verbal probing such as the following:

- "*I'll be truthful, I just answered that, with no thought in my head*" (OM02: Male, 60 or older, no qualifications, low income, White British)
- "*To tell you the truth, I just clicked it"* (SA04: Female, 50 to 59, no qualifications, very low income, White British)
- "*I think I just picked something round about the middle. I didn't give it that great a thought*" (SA03: Female, 40 to 49, higher education below degree level, employed, low income, White British)
- The interviewer asked how that was connected to the participant's answer to the survey question and the participant said, "*I'm not too sure, I think you have me on that one."* (PC02: Male, 40 to 49, CSE, O level or A level, on incapacity benefit, White British)

Examples of possible satisficing included participants' replies such as the following:

- "*Very Important*" and then was a bit caught out "*I think the tubes are quite good but buses can always be improved, I guess*" (MG03: Female, 50 to 59, postgraduate degree, high income, White British)
- Participant chose 'neither agree nor disagree' because she was not that bothered about the state of repair of properties (JF04: Female, 50 to 59, higher education below degree level, medium income, White British)

[25] The same process was used to identify satisficing on other categories, but the results are not given in this paper

- Admitted this is not something she thinks about (JC01: Female, 30 to 39, first degree, high income, White British)
- Participant answered 'satisfied' but then after the interviewer probed the participant asks "*is slightly satisfied the middle one? I'll go for the middle one"* (SA07: Female, 20 to 29, first degree, high income, White British)

The cognitive interviewing first focused on CATI versus CAWI participants for the two satisfaction questions ('satisfaction with street cleaning' and 'satisfaction with waste and recycling collection').[26] In the survey questions administered at the beginning of the cognitive interviews, there was a slightly greater number of endorsements of middle categories for CAWI participants compared to CATI participants. The cognitive interviews showed that almost all of the participants who chose the middle category on a survey question did so for justifiable reasons. The few cases of 'possible' satisficing were found in the CAWI mode of data collection.

The cognitive interviews also explored the issue of middle category satisficing across all three modes for the 12 agree/disagree questions.[27] Once again, only a few of the participants who had chosen the middle category were classified as cases of 'clear' or 'possible' satisficing. The remainder of the participants had chosen the middle category because it reflected the attitudes they actually had. There were instances of satisficing which occurred in both CAPI and CAWI, but the 'clear' satisficing was only found for CAWI participants.

The cognitive interviews results suggested that most middle category endorsements did not indicate satisficing, thus making a distinction that cannot be made with quantitative analysis. In contrast, the cognitive interview findings reinforced the quantitative findings and the survey research literature

[26] The cognitive interviews used the three category versions of these questions and created two alternative versions ('satisfaction with local shops for toiletries, clothing and food' and 'satisfaction with local green grocers'). Two of the questionnaire versions used the alternative questions in CATI and the original questions in CAWI, two of the versions did the reverse, and two of the versions explored other issues not discussed in this paper.

[27] The 12 agree/disagree questions were used on all versions of the questionnaire, but mode was varied. In three versions the four items scales were split across modes and in three versions the four item scales were kept together within a mode and compared to a different four item scale in a different mode.

by suggesting there may be slightly more satisficing in CAWI than in CAPI or CATI.[28] Further evidence to support more satisficing in CAWI than CAPI and how participants reacted to the different modes in general is found in Section 3.5.

3.5 Example 5: Mode preferences

3.5.1 Literature review

Rather than focusing solely on measurement differences between modes, some researchers have investigated whether participants prefer to use particular modes over others. Groves & Kahn (1979), for example, asked participants about mode preference at the end of a telephone and a face-to-face survey of households. The choices provided were a face-to-face interview, a telephone interview or a mailed questionnaire and a category was left for don't know and mixed opinion answers. Among telephone participants, the most preferred mode was telephone and among face-to-face participants, it was face-to-face. It could also be seen that telephone was less popular than face-to-face, with only 39 percent of those who experienced the telephone preferring telephone compared to 78 percent of those who experienced the face-to-face mode preferring face-to-face. The preference for the mode they were answering in was also found by Tarnai and Paxson (2004). They showed that businesses returning a completed mail questionnaire were more likely than those completing the survey on the web to prefer a mail survey (75 percent versus 17 percent) and those completing the survey on the web were more likely to prefer the web survey than those completing it by mail (77 percent versus 18 percent). In a school-based study, Brener, Eaton, Kann, Grunbaum, Gross, Kyle and Ross (2006) showed that students experiencing the "*CASI mode (computer-assisted self-interview) were more likely than students in the PAPI mode (paper and pencil interview) to indicate a preference for the computer, whereas students in the PAPI mode were more likely than students in the CASI mode to either prefer PAPI or have no preference*" (p. 369).

In an attempt to get around the possible bias of participants choosing the mode they had just experienced, Miller et al (2002) asked internet participants how they would find it to complete the same survey by mail, by telephone or face-to-face. Here, the majority of participants preferred mail (67 percent in the first group and 71 percent in the second group). Although internet was not asked about, there still may be some effect of 'mode experienced' as internet and mail are both self-completion modes with visually presented questions and answer categories.

[28] This general finding is also supported by the cognitive interviewing results on satisficing on question categories other than the middle one.

Millar, O'Neill, and Dillman (2009) extended the mode preference literature to see if participants could be 'pushed' into doing a web-survey by not offering a mail option. They showed that 87 percent of those who did the survey by mail said they preferred mail and 90 percent of those who did the survey by web, preferred web. The mode that participants experienced was also the best predictor of mode preference in logistic regression when analyses controlled for demographic variables and various variables to account for internet familiarity and use. They found that the strong relationship between mode completed and mode preference also held for those participants they 'pushed' into doing a web-survey.

Smyth and Olson (2010) looked at the longitudinal stability of mode preferences. At the end of a telephone survey, they asked participants their mode preferences and not surprising, telephone was the preferred mode. For the second wave of the survey, participants were randomly assigned to web and mail conditions (no telephone condition). Smyth and Olson found that mode preference was not very stable as participants were strongly influenced by the mode completed. They found that participants were 26 times more likely to report a preference for mail when they are given a mail survey and 16 times more likely to report a preference for web when they were given a web survey.

All of the literature discussed above shows a strong consistent pattern between mode completed and mode preference. However, the literature lacks a clear indicator of mode preference not associated with mode received. One possibility would be to ask mode preference after participants had experienced all modes. Unfortunately, in most cases this would be cost prohibitive.

3.5.2 Specific plan for cognitive interviews

The cognitive interviewing work is well placed to extend the literature on mode preference as the participants in this study had had experiences analogous to all three modes of data collection: CAPI (face-to-face), CATI (telephone) and CAWI (web). Some participants' comments about mode suggested mechanisms by which poorer quality data could be collected in certain modes.

As described in Section 2.2, participants first experienced survey questions in CAPI, then in CATI and then in CAWI. This was followed by cognitive interviewing retrospective think alouds and probes. At the end of the cognitive interviewing, cognitive interviewers were instructed to ask the participants, "*Overall what was it like to answer the questionnaire in person versus over the phone versus by yourself on the computer?*" If not covered in the responses to the previous question, cognitive interviewers were also to ask, "*What did you like or dislike about these three different ways you were asked the questions?"*

3.5.2 Cognitive interviewing results

Figure 2 summarises participants' comments and each bullet point represents a theme. Some participants gave multiple comments which are divided and put under the appropriate themes. It should be noted that there may be one or many participants for a given theme.

As can been seen from Figure 2, CATI was the least preferred mode. It was clear from the individual data before aggregation in the figure that no participants picked it as their first choice and only two picked it as their second choice. Participants had several reasons for disliking CATI, falling under the themes of 'distractions', 'giving less attention', 'pressured for time', 'having to work harder' and 'hearing and seeing issues'. There were also comments about how taking part in the telephone interview was not as good an experience as the face-to-face interview, and there were negative reactions to the telephone in general. In summary, these findings suggest that telephone responses may be of a lower quality than face-to-face or web responses, either because the participant pays less attention, the mode introduces difficulties on the participant's part (such as ability to hear) or the lack of any visual stimuli. Some illustrative quotes are given below.

- *"On the telephone I'm starting to look at things on my desk . . ."* (OM04: Female, 20 to 29, first degree, employed, medium income, White British)
- *"Now I have done some phone ones before . . . but I'll be honest with you if it's a question I don't understand I wouldn't pursue it. I would just give them an answer that I thought was relevant. . . . if I didn't understand it, I wouldn't say could you just repeat that again, you know. I would just go 'oh yes."* (JC05: female, 40 to 49, first degree, not working, low income, White British)
- *"Not the most effective way of asking questions 'cos you, you're trying to end the conversation quick, this, for me I just try to end the conversation quickly."* (MG02: female, 20 to 29, foreign qualification, employed, low income, White other)
- *[On the telephone] "I'm expected to retain all that even just for one question and I might get it wrong."* (OMO6: Female, 50 to 59, post grad degree, employed, high income, White British)
- *"I thought the telephone was difficult, 'cos it was difficult to hear everything you were saying."* (SA07: Female, 20 to 29, first degree, high income, White British)
- *"Generally speaking I don't like that sort of thing [a survey] on the telephone. Mainly because it's not there for me to see and I'm trying to think. . . Unless the answers are something that I can answer absolutely straightforwardly and don't have to think about, then I would rather have it in print."* (AR03: Male, 60 or older, first degree, retired, medium income, White British)

- *"I still think it's [the telephone] more intimidating than an eye to eye."* (OM05: male, 60 or older, secondary school, retired, very low income, White British)
- *"I don't like talking on the phone. Never have done."* (OM02: Male, 60 or older, no qualifications, low income, White British)

In contrast, participant preferences for CAPI versus CAWI were mixed. There were two basic tradeoffs between these modes:

(1) A tradeoff of valuing the personal nature of face-to-face versus the privacy of CAWI:

- *"Face to face without a doubt is the best. I would say that the computer is just so impersonal that you can't ask the computer the interpretation of the question. So, consequently you could receive a totally wrong answer by the misinterpretation, i.e. waste of time asking the question."* (OM05: male, 60 or older, CSE, O level or A level, retired, very low income, White British)
- *"Face-to-face, you feel like a bit on the spot."* (SA04: Female, 50 to 59, no qualifications, very low income, White British)
- *"You always have a feeling people, there's always an element of you're trying to please the person [laughs] you're always trying to answer the question to please them, I'm sure that's a well-known phenomenon and it does, you are a bit like that, you know, but with the web based thing it's black and white." (JC03: male, 50 to 59, CSE, O level or A level, employed, low income, White British)*

(2) A debate about whether the pace in CAWI is fast or slow:

- *"On the computer, it's a lot quicker, which I suppose could be taken as a good thing, but at the same time, I don't, me personally, I think face-to-face is a lot better. . . . [On the computer] I'm just doing the answers as they came into my head. I don't like it, no. I prefer to think about my answers."* (SA06: male, 20 to 29, no qualifications, not working, income not given, White British)
- *"I read it a bit too quick and just clicked". "* (SA04: Female, 50 to 59, no qualifications, very low income, White British)
- *[On the computer] "I'm given plenty of time to do it..."* (JC05: female, 40 to 49, first degree, not working, low income, White British)

Figure 2: Mode preferences and reasons given for these by first and second preferred modes and least preferred mode

	CAPI	CATI	CAWI
First choice mode	• Encourages better, considered, genuine answers (Rs more engaged, committed, had more concentration) • Allows opportunity to receive feedback from interviewer • Provides opportunity to ask interviewer questions (can check, ask things, encourages discussion) • More personal and comfortable (likes seeing the person) • Likes visual aids • Likes the cues given by the interviewer • It's traditional		• Familiarity, ease and like for using computers (problematic for less literate and less computer literate) • Allows you to take your time, go at you own pace • Allows completion in one's own space (i.e., when it's convenient, at their leisure) • Provides the opportunity to think and give better answers • Can visually see questions • Can visually see and read answers • Can go back to previous questions • It's quicker, one can go quickly through it • Is private and personal (don't feel judged, no one is listening and judging nor telling you if your answers were right or wrong) • Makes one feel less like they have to please the interviewer

Figure 2 continued

	CAPI	CATI	CAWI
Second choice mode	• OK, but has to have showcards • Even with cards, there's a need to keep questions in mind • Have to think more (compared to the computer) • It puts you on the spot / under pressure to give an answer • Doesn't like interviews (but liked this one)	• One can blank out one's surroundings on the phone	• Gives you time to think • Can visually see questions and answers • Able to go back • Answers are less influenced because interviewer is not there • OK, as long as it is not too long • OK, with help from someone (i.e., a son) • Can misread things • Makes you go quickly without thinking and choosing a middle answer • Sometimes unsure of what to do next on the computer (i.e., assuming you have to tick something to get to next question)

Figure 2 continued

	CAPI	CATI	CAWI
Least preferred mode	• Confusing and difficult to understand • Makes you think more and wonder if giving the right answer • Distractions (can blot this out on the phone)	• Distractions (household distractions, doing other things simultaneously, difficult to concentrate) • Give less attention to questions on the phone (need for questions to be repeated, rushed to a judgment, feeling it doesn't matter how you answer) • Pressured for time on the phone, makes you want to finish quickly • More misunderstandings • Have to work harder to answer (need to think more, hard to think about question and answer at same time and keep information in your head, hard to remember long lists of answer options) • Hard to hear (phone is fuzzier) • Lack of visual aids (can't see) • Harder to ask interviewer question (compared with face-to-face) • Less personal than face-to-face • Intimidating, uncomfortable • Dislike talking on the phone • A phone appointment would be on R's mind all day	• Disengaging • Distractions • Makes you want to get through it quickly and make quick decisions • Computer can take you through too fast • Inability to ask questions • Problematic for non-internet households

CAPI and CAWI both offer the possibility of visual stimuli for participants as even in face to face interviewing, answers can be presented on showcards. Participants with a preference for CAPI mentioned that they appreciated the showcards and participants with a preference for CAWI commented on how nice it was to see both the questions and answers. In addition, Figure 2 shows that participants had mentioned specific negative aspects of CAWI and specific positive aspects of CAPI. Some problems which participants mentioned for CAWI included: finding it less engaging, easier to get distracted, and an inability to ask questions about the survey questions. These were also mentioned as problems with CATI. There were also the web specific concerns of poor computer literacy and no home internet access. Specific positive comments about CAPI may suggest better quality data in this mode.

- *"Erm, I think for me you would get a more honest answer face-to-face. [Interviewer asks: What do you mean by honest?] Erm, more, an answer that I've thought about a bit more, rather than like I say on the computer sometimes you just want to get on the next page."* (OM04: female, 20 to 29, first degree, employed, medium income, White British)
- *"If it's like face-to-face then you're like committed, that you're going to answer it properly."* (PC01: female, 39 to 39, higher education below degree level, employed, medium income, other ethnicity)
- *"You'll get more personally from me, you'll get more out of me on a one-to-one interview than you would probably on a laptop or on the phone."* (JC05: female, 40 to 49 first degree, not working, low income, White British)

Participants' comments, which could indicate better or poorer quality answers, were useful. For example, participants' lack of preference for CATI (although more extreme than expected) was in line with the literature. For example, the figures from Groves and Kahn (1979) showed a much lower preference for telephone over face-to-face mode, even when this was the mode experienced. Comments about face-to-face showed potentially good quality data, but not with respect to privacy as would be expected. Comments about web surveys showed a mix of pros and cons with respect to quality. Although some participants clearly prefer the computer, some participants will give poorer quality answers from being less engaged and/or answering too quickly.

These findings suggest strengths and weaknesses of the different modes which could impact on how participants who are uncertain of their answers or who are faced with tricky answer

formats deal with this when completing questionnaires. Thus the information provided by participants about their mode preferences and the reasons for them offers insights into some of the root causes of mode effects.

4. Discussion and conclusions

This project was novel in the sense that cognitive interviewing was (1) pre-planned to take place after, rather than before a quantitative survey and (2) focused on uncovering and exploring mode differences rather than the standard aim of investigating the cognitive processes of comprehension, recall, judgment and response associated with survey answers.

Cognitive interviewing uncovered the cognitive processes involved in understanding and responding to questions that are the building blocks of mode effects. It identified issues not evident from quantitative analyses (e.g., which cases represented satisficing and which did not). Cognitive interview data, which exist in the form of participants' accounts or 'narratives', helped the research team explain expected, as well as unusual patterns in the quantitative mixed mode experiment. Thus, the cognitive interview results provided evidence that both supported and contradicted quantitative results and the survey research literature. In addition, it provided new ways to look at particular findings.

In the case of the 'yes/no' list (Example 1), the quantitative mixed mode experiment results showed unexpected differences across modes. The cognitive interview findings suggested the reasons for this pattern were complex, i.e., there was more going on than the aggregate quantitative findings suggested. Cognitive interviewing implied that the 'yes/no' format was not generally free of problems as suggested by Smyth et al (2006). At the same time the cognitive interview findings stimulated further quantitative analysis. Before the cognitive interviews, no member of the research team would have considered 'more thoughtful' answers as a factor in the CAPI higher levels of endorsements of the 'yes' category on the poverty question series.

For the aural end-labelled scales (Example 2), the quantitative mixed modes experiment results showed that there was confusion of scale direction on non-visual end-labelled scales. The cognitive interviewing confirmed that confusion of scale direction did indeed take place. This complements the findings of Dillman and Christian (2004) who were able to detect this

through participants scratching out and changing their answers in paper self-completion surveys and the evidence for reciprocal changes gathered with paradata by Stern (2008).

For the aural end-labelled scales (Example 3), the quantitative mixed modes experiment results found more middle category selection in end-labelled versus fully-labelled format, but less so for CATI participants. The cognitive interviewing confirmed the difficulty that CATI participants had in general without the use of visual aids. It also challenged the conclusion made by Tarnai and Dillman (1992). They had thought the greater use of the middle category in visual modes rather than aural modes was due to the fact that the middle category was more prominent. Evidence from the cognitive interviewing suggested participants had difficulty finding the middle category on an aural end-labelled scale. Therefore, it is not so much that a visual mode makes the middle category prominent, but that aural modes make it more difficult to find.

In the case of middle category satisficing (Example 4), the quantitative mixed modes experiment results showed more middle category endorsement in satisfaction and agree/disagree questions in CAWI than in the interviewer modes, suggesting more satisficing. The cognitive interviewing showed that a majority of participants who had chosen middle categories gave reasonable and justifiable answers and these did not appear to be instances of satisficing. In addition, the cognitive interviews allowed us to highlight just those participants who were satisficing. Having done this, the cognitive interviews could then confirm what would be expected in the survey literature, i.e., that CAWI participants may be more likely satisfice than CAPI participants. The cognitive interviewing work on mode preferences (Example 5) strengthens this conclusion. Although CAWI participants can work at their own pace, some mentioned completing the survey too quickly on the computer. Other participants mentioned that CAWI gives no opportunity for asking questions as well as commenting on distractions and paying less attention to the question. The mode preference work also showed preconditions for potential satisficing in CATI. These included issues such as problems hearing, no visual aids, distractions, paying less attention and for some, a general dislike of using the telephone.

Like any study, both the quantitative mixed modes experiment and the cognitive interviewing phases have limitations. In terms of the quantitative mixed modes experiment, some of the questions discussed in this paper were not optimal; they were taken from other surveys and

were not re-tested before use in this research. For example, the satisfaction question on 'state of the economy' was a poor choice of a 'typical' satisfaction question, as a majority of participants were dissatisfied with the economy in 2009. The financial decision scale used the term 'rarely' which was confusing when combined with an answer scale including 'disagree' since it becomes a double negative. We are also aware that the end-labelled behavioural frequency questions would be unlikely to be used in practice and were only included to complete the larger quantitative mixed modes experiment's hypotheses.

In terms of the cognitive interviewing some possible limitations arise with the design and implementation of the interviewing and interviewer presence. (1) There could be a concern that the survey component of the cognitive interviews did not fully replicate true CAPI, CATI and CAWI conditions. Although not a complete replication, we do however feel that the three modes were convincingly implemented, which was supported by participant comments. (2) There could be concern that cognitive interviewers more experienced in exploring mode effects may have followed up on the occasional ambiguous think aloud statement to get more definitive information. (3) Some of the standardised probes (such as the standardised probe about confidence of scale direction for the end-labelled scales in CATI) and newly designed questions could have been improved. (4) There could be a concern that the retrospective think aloud and probing could lead to post hoc rationalisations. This is impossible to detect completely, but generally participants appeared to have thought about the issues initially and remembered these during the retrospective work. Out of all the participants and answers to all the questions, there was only one clear example of a post hoc rationalisation. However, a slight exception to this is the CAWI mode where participants were sometimes surprised when their CAWI answer was read back to them later in the cognitive interview. But this could also suggest a much lower level of cognitive processing on questions in the CAWI mode. (5) Interviewer presence could have played a role as participants were asked about their mode preference in front of a face-to-face cognitive interviewer. But given the candid comments made by participants and the better rapport which develops in the cognitive interview as opposed to standard face-to-face interviewing, this may only be a small concern. (6) The sample size was limited and it would have been useful to have a bigger sample size so that sub-themes could be explored and to potentially capture a larger set of instances of participant issues (such as the confusion of the direction of end-labelled scales in CATI).

The final set of limitations relate to the participants. (a) Were cognitive interview participants on their 'best behaviour' and not answering as normal survey participants? We will never know for sure, but these participants did show less than optimal behaviour by satisificing and interrupting interviewers. In addition, their comments throughout about mode clearly suggested that some gave less than optimal answers in CATI and/or CAWI. Similarly, some participants talked with pride about telling their true feelings and others confessed to the interviewers about how bad their opinions must seem. (b) In contrast, it must be remembered that the sample of participants used for the cognitive interviews purposively over-sampled participants who showed less than optimal behaviour on the mixed modes experiment. (c) All of the cognitive interview participants had been interviewed twice previously and still agreed to taking part in the cognitive interviews. These kinds of people are likely to be different to both the people who never take part in surveys and the people who one might draw into a one-off survey.

Overall, we feel the use of cognitive interviewing to explore the quantitative findings from a mixed modes study has been a very useful exercise and we hope that it will set a trend for more cognitive interviewing work like this.

References

Ainsworth, B. (2000), "Issues in the Assessment of Physical Activity in Women," *Research Quarterly for Exercise and Sport*, 71(Supplement 2): 37-42.

American Association for Public Opinion Research (2011), "Standard definitions: Final dispositions of case codes and outcome rates for surveys", 7th edition, Retrieved September 8, 2014, from http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/StandardDefinitions2011_1.pdf.

Beatty, P. (2004), "The Dynamics of Cognitive Interviewing," in Methods for Testing and Evaluating Survey Questionnaires, eds. S. Presser et al, Hoboken, NJ: Wiley.

Brener, N., Eaton, D., Kann, L., Grunbaum, J., Gross, L., Kyle, T., and Ross, J. (2006), "The Association of Survey Setting and Mode with Self-Reported Health Risk Behaviour Among High School Students," *Public Opinion Quarterly*, 70(3): 354–374.

Campanelli, P., Gray, M., Blake, M., and Hope, S. (2015), "Cognitive Interviewing as Tool for Enhancing the Accuracy of the Interpretation of Quantitative Findings," *Quality & Quantity*, published online: 07 April 2015.

Carrasco, L. (2003), "The American Community Survey (ACS) en Español: Using Cognitive Interviews to Test the Functional Equivalency of Questionnaire Translations," Statistical Research Division Study Series (Survey Methodology #2003-17), Washington, DC: U.S. Census Bureau.

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.

Collins, D. (ed) (2015), *Cognitive Interviewing Practice*, Thousand Oaks, CA: *Sage.*

Cosenza, C., and Fowler, F.J. (2000), "Prospective questions and other issues in cognitive testing," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 994-997.

Davis, E., Nicolas, C., Waters, E., Cook, K., Gibbs, L., Gosch, A., and Ravens-Sieberer, U. (2007), "Parent-Proxy and Child Self-Reported Health-Related Quality of Life: Using Qualitative Methods to Explain the Discordance," *Quality of Life Research*, 16(5): 863-71.

de Leeuw, E. (2005), "To Mix or Not to Mix Data Collection Modes in Surveys," *Journal of Official Statistics, 21*(2):233-255.

DeMaio, T.J., and Landreth, A. (2004), "Do Different Cognitive Interview Techniques Produce Different Results?," in Methods for Testing and Evaluating Survey Questionnaires, eds. S. Presser et al, Hoboken, NJ: Wiley.

Dillman, D., and Christian, L. (2005), "Survey Mode as a Source of Instability in Responses Across Surveys," *Field Methods*, 17(1): 30-52.

Dillman, D., and Redline, C. (2004), "Testing Paper Self-administered Questionnaires: Cognitive interview and Field Test Comparisons," in Methods for Testing and Evaluating Survey Questionnaires, eds. S. Presser et al, Hoboken, NJ: Wiley.

Forsyth, B., Rothgeb, J., and Willis, G. (2004), "Does Question Pretesting Make a Difference? An empirical test," in Methods for Testing and Evaluating Survey Questionnaires, eds. S. Presser et al, Hoboken, NJ: Wiley.

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

Fowler, F.J., Roman, A., and Di, Z. (1998), "Mode Effects in a Survey of Medicare Prostate Surgery Patients," *Public Opinion Quarterly*, 62(1): 29-46.

Gerber, E. (1999), "The View from Anthropology: Ethnography and the Cognitive Interview," in Cognition and Survey Research, eds. M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur and R. Tourangeau, New York: Wiley.

Gray, M., Blake, M., and Campanelli, P. (2014), "The Use of Cognitive Interviewing Methods to Evaluate Mode Effects in Survey Questions," *Field Methods*, 26(2): 156-171.

Groves, R., and Kahn, R. (1979), *Surveys by Telephone*, New York: Academic Press.

Goerman, P. (2006), "Adapting Cognitive Interview Techniques for Use in Pretesting Spanish Language Survey Instruments," Statistical Research Division Research Report Series (Survey Methodology #2006-3), Washington, DC: U.S. Census Bureau.

Harkness, J.A., Van de Vijver, F.J.R., and Mohler, P. (eds) (2003), *Cross-Cultural Survey Methods*, Hoboken, NJ: Wiley.

Heerwegh, D. (2002), Describing Response Behaviour in Web Surveys Using Client-Side Paradata, paper presented at the International Workshop on Web Surveys, Mannheim, Germany, October 17-19.

Heerwegh, D. (2003), "Explaining Latency and Changing Answers Using Client-Side Paradata from a Web Survey," *Social Science Computer Review*, 21(3), 360-373.

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., and Jäckle, A. (2014), "The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration," ISER Working paper, no. 2014-20, ISER, University of Essex.

Jakwerth, P., Stancavage, F.B., and Reed, E. (1999), "An Investigation of Why Students Do Not Respond to Questions," Working Paper No. 2003-12, National Centre for Education Statistics, U.S. Department of Education: Washington, DC.

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., and Fabrigar, L. (1997), "Designing Rating Scales for Effective Measurement in Surveys." in Survey Measurement and Process Quality, eds. L. Lyberg et al, Hoboken, NJ: Wiley.

Krosnick, J., and Presser, S. (2010), "Question and Questionnaire Design," in Handbook of Survey Research, eds. P. Marsden and J. Wright, 2nd edition, Bingley, UK: Emerald.

Lynn, P., Hope, S., Jackle, A., Campanelli, P., and Nicolaas, G. (2012), "Effects of Visual and Aural Communication of Categorical Response Options on Answers to Survey Questions," ISER Working paper, no. 2012-21, ISER, University of Essex.

Millar, M., O'Neill, A., and Dillman, D. (2009), "Are Mode Preferences Real?" Technical Report 09-003, Pullman, Washington: Social & Economic Sciences research center (SESRC), Washington State University.

Miller, K. (2008), *Results of the Comparative Cognitive Test Workgroup Budapest Initiative Model*, Hyattsville, MD: National Center for Health Statistics.

Miller, T., Miller-Kobayashi, M., Caldwell, E., Thurston, S., and Collett, B. (2002), "Citizen Surveys on the Web: General Population Surveys of Community Opinion," *Social Science Computer Review*, 20(2): 124-136.

Mitofsky, W., and Edelman, M. (1993), "A Review of the 1992 VRS Exit Polls," paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Charles, IL.

Nicolaas, G., Campanelli, P., Hope, S., Jackle, A., and Lynn, P. (under review), "Revisiting 'Yes/No' versus 'Check all that Apply': Results from a Mixed Modes Experiment."

Oksenberg, L., and Cannell, C. (1989), "New Methods for Pretesting Survey Questionnaires," paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Padilla, J.L. (2007), "The Use of Cognitive Interviewing to Assess "Construct Overlap" in Cross-Cultural Surveys," paper presented at 2nd Conference of European Survey Research Association, Prague.

Pan, Y. (2004), "Cognitive Interviews in Languages Other Than English: Methodological and Research Issues," *2004 American Statistical Association Proceedings of the Joint Statistical Meetings*, Statistical Computing Section [CD-ROM], (pp. 4859-4865). Alexandria, VA: American Statistical Association.

Rasinski, K., Mingay, D., and Bradburn, N. (1994), "Do Respondents Really 'Mark All That Apply' on Self-Administered Questions?," *Public Opinion Quarterly*, 58: 400–408.

Schwarz, N., Strack, F., Hippler, H.J., and Bishop, G. (1991), "The Impact of Administration Mode on Response Effects in Survey Measurement," *Applied Cognitive Psychology*, 5: 193-212.

Skjak, K.K., and Harkness, J. (2003). "Data Collection Methods," in Cross-cultural Survey Methods, eds. J. Harkness, F. van de Vijver and P. Mohler, New York: Wiley.

Smyth, J.D., Christian, L.M., and Dillman, D. (2008), "Does "Yes or No" on the Telephone Mean the Same as "Check-All-That-Apply" on the Web?," *Public Opinion Quarterly*, 72(1): 103-113.

Smyth, J.D., Dillman, D., Christian, L.M., and Stern, M.J. (2006), "Comparing Check-all and Forced-Choice Question Formats in Web Surveys," *Public Opinion Quarterly*, 70(1): 66-77.

Smyth, J. and Olson, K. (2010), "Measuring Mode Preference: Question Wording and Question Order Experiments," paper presented at the 65th Annual Conference of the American Association for Public Opinion Research, Chicago, IL.

Sudman, S., and Bradburn, N.M., (1982), *Asking Questions - A Practical Guide to Questionnaire Design*, San Francisco, CA: Jossey-Bass.

Stern, M. (2008), "The Use of Client-side Paradata in Analyzing the Effects of Visual Layout on Changing Responses in Web Surveys," *Field Methods,* 20(4): 377-398.

Tarnai, J., and Dillman, D. (1992), "Questionnaire Context as a Source of Response Differences in Mail vs Telephone Surveys," in Order Effects in Social and Psychological Research, eds. N. Schwarz, H. J. Hippler, and S. Sudman, New York: Springer-Verlag.

Tarnai, J., and Paxson, M.C. (2004), "Survey Mode Preferences of Business Respondents," paper presented at the 59th Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Tourangeau, R. (1984), "Cognitive Science and Survey Methods," in Cognitive Aspects of Survey Design: Building a Bridge between Disciplines, eds. T. Jabine, M. Straf, J. Tanur, and R. Tourangeau, R., Washington, DC: National Academies Press. pp. 73-100.

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.

Willis, G. (2004), "Overview of Methods for Developing Equivalent Measures Across Multiple Cultural Groups," paper presented at the Eighth Conference on Health Survey Research Methods, eds. S.B. Cohen and J.M. Lepkowski, Hyattsville, MD: National Center for Health Statistics.

Willis, G. (2005), *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Thousand Oaks, CA: Sage.

Willis, G. and Schechter, S. (1997), “Evaluation of Cognitive Interviewing Techniques: Do the Results Generalize to the Field?,” *Bulletin de Methodologie Sociologique*, 55: 40-66.

Appendix: Wording and source of questions analysed

For Example 1					
Item Formats	Showcard in CAPI?	Question Number	Topic of question and actual wording Note that only questions highlighted in grey were used in the cognitive interviewing	Response Options	Source
'Yes/no' list (from a 'mark all that apply' versus 'yes/no' experiment)	No showcards		**WAYS TO REDUCE POVERTY** I am now going to ask you a number of questions about different methods for reducing poverty. In your opinion, which of the following would be effective?	Two categories: Yes and No	Poverty and Social Exclusion Survey of Britain, 1999
		GB21.	**INCREASING PENSIONS:** Would increasing pensions reduce poverty?		
		GB22.	**EDUCATION FOR CHILDREN:** Would investing in education for children reduce poverty?		
		GB23.	**IMPROVING ACCESS TO CHILDCARE:** Would improving access to childcare reduce poverty?		
		GB24.	**REDISTRIBUTION OF WEALTH:** Would the redistribution of wealth reduce poverty?		
		GB25.	**INCREASING TRADE UNION RIGHTS:** Would increasing trade union rights reduce poverty?		
		GB26.	**REDUCING DISCRIMINATION:** Would reducing discrimination reduce poverty?		
		GB27.	**INCREASING INCOME SUPPORT:** Would increasing income support reduce poverty?		
		GB28.	**INVESTING IN JOB CREATION:** Would investing in job creation reduce poverty?		
'Yes/no' list (from a 'mark all that apply' versus 'yes/no' experiment)	No showcards		**THINGS YOU LIKE ABOUT YOUR NEIGHBOURHOOD** What are the things that you like about your neighbourhood?	Two categories: Yes and No	Adapted from a London Housing Association questionnaire
		N56.	**COMMUNITY SPIRIT:** Do you like your neighbourhood because of its community spirit?		
		N57.	**SAFETY:** Do you like your neighbourhood because it feels safe?		
		N58.	**NEIGHBOURS:** Do you like your neighbourhood because of the neighbours?		
		N59.	**CHARACTER OF BUILDINGS:** Do you like your neighbourhood because of the character of its buildings?		
		N60.	**CLEANLINESS:** Do you like your neighbourhood because of its cleanliness?		
		N61.	**LOCATION:** Do you like your neighbourhood because of its location?		
		N62.	**QUIETNESS:** Do you like your neighbourhood because it is quiet?		
		N63.	**TRANSPORT:** Do you like your neighbourhood because of its transport		

For Example 2 and 3					
Item Formats	Showcard in CAPI?	Question Number	Topic of question and actual wording Note that only questions highlighted in grey were used in the cognitive interviewing	Response Options	Source
End labelled scales (from an end-labelled versus fully-labelled experiment)	Showcards used	FM5	**IMPORTANCE OF MONEY:** On a scale from 1 to 10 how important is having a lot of money is to you, where '1' equals 'Not important at all' and '10' equals 'Very important.	Eleven categories with end labels of Not important at all and Very important	New
		GB16	**SATISFACTION WITH DEMOCRACY AND PERSONAL FREEDOM:** On the whole, how satisfied are you with the way democracy and personal freedom work in Great Britain?	Seven categories with end labels Very Satisfied and Very dissatisfied	
		GB17	**SATISFACTION WITH THE ECONOMY:** And on the whole, how satisfied are you with the present state of the economy in Great Britain?	Seven categories with end labels Very Satisfied and Very dissatisfied	European Social Survey, 2006
		FM68	**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?	Seven categories with end labels Every day and Never	New
		FM74	**AMOUNT OF HOT BEVERAGES PURCHASED OUTSIDE THE HOME:** 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.	Seven categories with end labels None and More than 25	New

For correlations in Example 2, Table 2					
Item Formats	Showcard in CAPI?	Question Number	Topic of question and actual wording	Response Options	Source
From initial 15 questions with no format experiment	No	FM1	**HOW WELL MANAGING FINANCIALLY:** How well would you say you yourself are managing financially these days? Would you say you are	Living comfortably, doing alright, just about getting by, finding it quite difficult or finding it very difficult	British Household Panel Study, Wave 17
	No	FM3	**FINANCIAL EXPECTATION A YEAR FROM NOW:** Looking ahead, how do you think you will be financially a year from now? Will you be	Better off, Worse off than you are now, or about the same	British Household Panel Study, Wave 17
	Yes	FM6	**MONTHLY SPEND ON EATING OUT:** Please look at this card and tell me about how much you personally spend in an average month on eating out at, or buying takeaway food from, a restaurant, pub or café, including school meals or meals at work?	Under £10, £10-£19, £20-£29, £30-£39, £40-£49, £50-£59, £60-£79, £80-£99, £100-£119, £120-£139, £140-£159, £160 or over.	British Household Panel Study, Wave 17
Set of Four Agree / disagree statements	No showcards	FM64-FM67	**THOROUGHNESS OF PREPARATION BEFORE MAKING A LARGE FINANCIAL DECISION:** To what extent do you agree or disagree with the following statements about making important financial decisions such as taking out a mortgage, loan or pension. • I would rarely read all the small print before making important financial decisions. • I would do a lot of research before making an important financial decision. • I would rarely talk to a financial advisor before making an important financial decision. • I definitely would talk to family and friends before making an important financial decision.	Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree	Modified and extended from two statements from 2006 British Social Attitudes survey
Long scales (from a long versus short scale experiment) crossed with showcard/no showcard in CAPI	A random half of CAPI participants received a showcard and others did not	FM75	**HOUSING TYPE:** 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
		FM81	**MONTHLY SPEND 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

For correlations in Example 2, Table 3					
Item Formats	Showcard in CAPI?	Question Number	Topic of question and actual wording	Response Options	Source
From initial 15 questions with no format experiment	No	FM1	**HOW WELL MANAGING FINANCIALLY:** How well would you say you yourself are managing financially these days? Would you say you are	Living comfortably, doing alright, just about getting by, finding it quite difficult or finding it very difficult	British Household Panel Study, Wave 17
	Yes	FM7	**HOW FREQUENTLY GO TO PUB:** Tell me how frequently you go for a drink at a pub or club?	At least once a week, at least once a month, several times a year, once a year or less, never/almost never	British Household Panel Study, Wave 17
	Yes	N15	**PEOPLE ARE TOO READING TO CRITICISE:** Tell me how strongly you agree or disagree with this statement. People in Britain are too ready to criticise their country.	Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree	British Household Panel Study, Wave 17

Example 4, Table 5					
Item Formats	Showcard in CAPI?	Question Number	Topic of question and actual wording Note that only questions highlighted in grey were used in the cognitive interviewing	Response Options	Source
Long scales (from a long versus short scale experiment) crossed with showcard/no showcard in CAPI	A random half of CAPI participants received a showcard and others did not	N43	**SATISFACTION WITH WASTE AND RECYCLING COLLECTION:** I would like you to tell me how satisfied or dissatisfied you are with local household waste collection, recycling collection and other recycling collection points. Would you say you are	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)
		N44	**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

Example 3, Table 4 (continued)

Item Formats	Showcard in CAPI?	Question Number	Topic of question and actual wording Note that only questions highlighted in grey were used in the cognitive interviewing	Response Options	Source
Set of Four Agree / disagree statements	No showcards	N35-N38	**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 questionnaire
Set of Four Agree / disagree statements	No showcards	FM64-FM67	**THOROUGHNESS OF PREPARATION BEFORE MAKING A LARGE FINANCIAL DECISION:** To what extent do you agree or disagree with the following statements about making important financial decisions such as taking out a mortgage, loan or pension. • I would rarely read all the small print before making important financial decisions. • I would do a lot of research before making an important financial decision. • I would rarely talk to a financial advisor before making an important financial decision. • I definitely would talk to family and friends before making an important financial decision.	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	N52-N55	**MENTAL HEALTH PATIENTS AND FORMER PRISONERS IN R'S NEIGHBOURHOOD:** How strongly do you agree or disagree with the following four 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