

Cognitive Interviewing as Tool for Enhancing the Accuracy of the Interpretation of Quantitative Findings

PAMELA CAMPANELLI

Survey Coach, 33 Military Road, Colchester, Essex CO1 2AD UK. E-mail:

dr.pamela.campanelli@thesurveycoach.com. Phone: +44 7803 275317

MICHELLE GRAY and MARGARET BLAKE

NatCen Social Research, 35 Northampton Square, London EC1V 0AX, UK

STEVEN HOPE

Institute of Child Health, University College London, Gower Street, London, WC1E 6BT

Abstract. This paper contrasts findings from a quantitative survey with those from a cognitive interviewing follow-up investigation on a subset of the same respondents. The data were gathered as part of a larger study to explore measurement error across three modes of data collection, but this paper focuses on the question format experiments rather than mode effects part of the larger study. Three examples are presented which demonstrate how cognitive interviewing can cast new light on quantitative results by increasing the accuracy of the inferences made. These include instances where: (1) quantitative indicators of poor respondent behaviour (e.g., acquiescence bias on agree / disagree questions) are over-estimates, (2) similar quantitative response distributions across satisfaction and behavioural questions (from a fully-labelled versus end-labelled experiment) imply similar respondent satisficing behaviour, but cognitive interviews show that different response processes are at work and (3) unlikely quantitative findings (from an experiment comparing 3 versus 7 or 8 response options) could

easily be dismissed as due to chance but were instead the result of unforeseen respondent difficulties. The paper concludes with a discussion of the value of using a cognitive interviewing follow-up study as a tool in the interpretation of ambiguous quantitative findings.

Keywords: Cognitive interviewing, satisficing, acquiescence, end-labelled, polar point, number of response options

1. Introduction

Cognitive interviewing is traditionally thought of as a pretesting method. But this paper explores the use of cognitive interviewing as a pre-planned follow-up to a quantitative survey.

The original purpose for both the quantitative and cognitive interview phases of the research was to explore measurement error differences between face-to-face, telephone and web data collection and to explore the interaction of mode of data collection with seven question format experiments^{1,2}. Because of the need to provide a greater understanding of how mode effects

¹ The question formats experiments included fully-labelled versus end-labelled scales, scales with few versus many response options, agree/disagree statements versus balanced forced choice questions, rating versus ranking, ‘yes/no for each’ versus ‘mark all that apply’, unfolding versus one-step scales, and showcard versus no showcard on long lists in computer-assisted personal interviewing.

² For information on the quantitative mixed-mode results, see Nicolaas, Campanelli, Hope, Jäckle and Lynn (2011), Hope, Campanelli, Nicolaas, Lynn and Jäckle (2014) and Lynn, Hope, Jäckle, Campanelli and Nicolaas (2012). For information on the cognitive interviewing mode

happen even if they are not directly observed, the cognitive interviewer did not gather verbal reports about respondents' understanding of the survey questions. Instead the cognitive interviewing focused on how respondents arrived at their answers.

The aim of this paper is to illustrate how cognitive interviewing can be used after a quantitative study to inform inferences from quantitative findings. To meet this aim, we provide three examples that use data from question format experiments. The first example explores the issue of acquiescence in agree/disagree questions, the second looks at satisfaction and behavioural frequency questions in an end-labelled versus fully-labelled experiment, and the third looks satisfaction and factual questions in a 3 versus 7 or 8 response option experiment.

Before the methods and results sections, we discuss the concept of 'satisficing' and review the background literature associated with the three examples and with cognitive interviewing.

1.1. SATISFICING

Since Krosnick (1991) introduced Herbert Simon's term, 'satisficing', into 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 respondents' ability, respondents' motivation and the difficulty of the task, respondents may take a less than optimal shortcut to their answer.

results see Campanelli, Blake, Gray and Hope (2010). For information about the differences used in the cognitive interviewing methodology for a study of mode effects, see Gray, Blake and Campanelli (2014).

The satisficing concept was originally based on four cognitive stages of the response process: comprehension, recall, judgment and response (Tourangeau 1984). Weak satisficing was defined as cases where respondents were less thorough in their path through the four cognitive stages. *“They may be less thoughtful about a question’s meaning, they may search their memories less thoroughly, they may integrate retrieved information more carelessly, and they may select a response choice more haphazardly”* (Krosnick 1991, p. 214). Strong satisficing was defined as cases where respondents *“interpret each question only superficially and select what they believe will appear to be a reasonable answer to each question without referring to any internal psychological cues specifically relevant to the attitude, belief, or event of interest”* (Krosnick 1991, p. 215).

According to Krosnick (1991) and Krosnick and Fabrigar (1997), satisficing is the root cause of many response effects in surveys, such as selecting the first response option that constitutes a reasonable answer, or agreeing with items that make an assertion (weak satisficing) and selecting ‘don’t know’, when respondents do know; choosing the middle response option, when respondents have a different opinion; selecting the same response for every item or answering randomly (strong satisficing).

1.2. THE LITERATURE RELATED TO THE 3 EXAMPLES USED

1.2.1. *Example 1: Background on Acquiescence in Agree/Disagree Questions*

Although a very popular question format for attitude measurement (Converse and Presser 1986), the agree/disagree format is prone to acquiescence response bias (the propensity of individuals to agree with a statement regardless of its content).

Historically there have been many views on what causes the acquiescence. Some have believed that acquiescence arises from ambiguity in the agree/disagree statement and have attempted to differentiate ‘good’ statements from ‘poor’ ones (e.g., McBride and Moran 1967). A larger stream of research has viewed acquiescence as specifically due to limited cognitive ability / poor education (Schuman and Presser 1981), but results with respect to acquiescence and education are mixed (Ayidiya and McClendon 1990; Narayan and Krosnick 1996; Landsberger and Saavedra 1967). A slightly different perspective on low education is given by other authors (see, for example, Lenski and Leggett 1960; Warnecke, et al. 1997; Javeline 1999). Here it is believed that the link between low education and acquiescence is not cognitive ability but deference (e.g., ‘deference’ of a lower status respondent with a higher status interviewer). A further variation is ‘category fallacy’ which combines the idea of an ambiguous question with a socially desirable response; for example, ‘playing it safe’ *“by acquiescing rather than risk appearing foolish or ignorant”* (Warnecke, et al. 1997, p. 336). As discussed in Section 1.1, Krosnick (1991) labels acquiescing behaviour as ‘weak’ satisficing. According to Krosnick (2000), if the question appears to be offering only one alternative, then the respondent might assume that it must be ‘right’ as the researcher must know the ‘right’ answer. So, as long as the statement seems reasonable, the easiest response is just to agree.

Acquiescence can be detected in a split ballot experiment by comparing a single agree/disagree item to a question in balanced ‘forced-choice’ format (Schuman and Presser 1981; Ayidiya and McClendon 1990; Javeline 1999) or comparing conceptually opposite statements (Schuman and Presser 1981; and Javeline 1999). It is also possible to compare opposite statements in the same questionnaire (see Lenski and Leggett 1960; Landsberger and Saavedra

1967). This latter method has been used by the designers of multi-item scales (see DeVellis 2012).

1.2.2. *Example 2: Background on End-Labelled versus Fully-labelled Scales*

There are pros and cons to the use of fully-labelled versus end-labelled scales in a questionnaire. For example, after reviewing the reliability and validity literature on this topic, Krosnick and Fabrigar (1997) conclude that “*fully-labelled scales are more reliable and valid than partially-labelled scales*” (p. 152). However, they point out that this assumes researchers have selected labels with “*relatively precise meaning*” reflecting “*equal intervals along the continuum of interest*” (p. 152). They argue this is true because, even though numbers are more precise than verbal labels and easier to hold in memory, “*people rarely express complex conceptual meaning in everyday conversation via numbers*” (p.149) and the verbal labels help to clarify the meaning of the scale options. Toepoel and Dillman (2010) suggest that fully-labelled scales are more robust than end-labelled scales and less likely to be influenced by the spacing of the response options and the use of different end points labels. In contrast, there is the difficulty for the researcher of choosing appropriate adjectives for more than 5 response options (Fowler 1995).

The most important issue, however, is that the two formats do not produce the same results. Dillman and Christian (2005) suggest that respondents 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 appears to be greater use of the middle response option in end-labelled scales (with a difference of 21 percent between the two formats for question 1 and 22 percent for question 2).

1.2.3 *Example 3: 3 versus 7 or 8 response options for factual and non-factual questions*

For subjective, non-factual questions in the visual modes, there are recommendations for the use of 5 response options as a minimum and 7 as a maximum (Fowler 1995) and 4 as a minimum and 7 as a maximum (Krosnick and Fabrigar 1997). Thomas, Uldell, & Krosnick (2004), experimented with a number of aspects of response option formats and suggested the validity is optimised if the number of response options is increased from 2 to 5. Goggin and Stoker (2014) make a similar conclusion using simulations, with reliability increasing when response options increase from 2 to 5 or 7 points. On the other hand, the cognitive complexity of the task increases with the number of response options (Fowler 1995; Tourangeau, Rips and Raskinski 2000). For factual questions, there are no clear suggested limits to the number of response options, but the possibility of primacy effects urges designers to opt for shorter rather than longer lists of response options.

1.3. COGNITIVE INTERVIEWING

Since the inception of the ‘cognitive laboratory’ in 1984-1985, cognitive interviewing has been a common pretesting technique. The two main techniques are thinking aloud (‘protocol analysis’ from Ericsson and Simon 1984) and pre-determined and spontaneous verbal probing (see DeMaio and Rothgeb 1996; Willis 2005). Both techniques are used to reveal respondent difficulties with specific terms or phrases, to uncover whether the respondent is able to recall the necessary information from memory and how he or she goes about doing so, and to look at issues with the response options or question sensitivity. Beatty and Willis (2007, p. 287) offer a wider definition “*the administration of draft survey questions while collecting additional verbal*

information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends”.

Over the last 15 years, cognitive interviewing has expanded in several ways. One of the largest areas of growth is in the translation and pretesting of survey questions which are cross-national, cross-cultural and multilingual. For example, researchers have looked at the adaptations needed to use cognitive interviewing in other languages (e.g., Goerman 2006; Pan 2004). And cognitive interviewing has also been used to explore cultural and linguistic issues and be part of the actual translation process (e.g., Carrasco 2003; Padilla 2007; Levin et al. 2009).

Another direction is what some people call the ‘the ethnographic approach’ (see for example Gerber 1999 and Ainsworth 2000). Here the emphasis is on probing respondents’ circumstances because survey questions may be cognitively clear but not fit the situations in respondents’ lives.

It is still rare for cognitive interviewing, a pre-testing method, to be used after a quantitative survey. Some examples are Davis and her colleagues (2007), who used think-aloud to understand the discrepancies between the way parents and children answered a health-related quality of life questionnaire and Jakwerth and her colleagues (1999), who used standardised probes developed by their cognitive laboratory staff to explore why students were leaving certain questions blank in the an educational assessment questionnaire. Miller (2008) suggested that well-documented cognitive interview findings from the pretesting stage collected before the main survey fieldwork could then be used after the survey fieldwork to assist in the interpretation of quantitative results.

The cognitive interviewing described in this paper was from a pre-planned follow-up of a quantitative mixed modes experiment.

2. Methods

2.1. THE QUANTITATIVE MIXED MODES EXPERIMENT

For the quantitative mixed modes experiment which preceded the cognitive phase, initial data were collected using the 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 respondents and 1,600 face-to-face interviews were completed using computer-assisted personal interviewing³.

For the mixed modes experiment, NatCen Omnibus survey respondents from an earlier data collection who agreed to be re-contacted were randomly allocated to one of three modes: CAPI (computer-assisted personal interviewing), CATI (computer-assisted telephone interviewing) and CAWI (computer-assisted web interviewing) with data collection taking place during the fourth quarter of 2008 and second quarter of 2009. Response rates were CAPI: 78%, CATI: 68%, and CAWI: 47% (AAPOR 2011, RR5).

2.2. THE COGNITIVE INTERVIEWS

Thirty six respondents were recruited for the cognitive interviewing phase from respondents who had participated in the quantitative mixed modes experiment. The original purpose of the cognitive interviewing follow-up was two-fold: to gain a greater understanding of how mode effects happen even if they are not directly observed, and to serve as a follow-up to investigate

³ At the time this paper was written, NatCen Social Research was no longer running its Omnibus survey.

any unusual findings from the experiment. While mode differences are typically detected at an aggregate level, an analysis of the quantitative data showed that certain respondent ‘satisficing’ behaviours differed by mode. For example, acquiescence through agreeing to opposite agree/disagree statements was more common in the interview modes and non-differentiation in a ranking task was more common in web completion. The final quota was therefore based on satisficing behaviour: we included 18 respondents who had agreed to opposite statements, 9 who showed non-differentiation in a ranking task and 9 ‘non-satisficers’. The ‘non-satisficers’ were randomly selected from respondents showing the opposite demographic and socioeconomic profile to those who had exhibited satisfying behaviour, see Gray, Blake and Campanelli (2014) for more details.⁴

The cognitive interviews began with the administration of a subset of survey questions from the mixed modes experiment.⁵ These questions were carefully selected as a result of the quantitative analyses, based on findings that needed further investigation. Because there were more issues to investigate than could be included within a one hour cognitive interview (with 10 minutes as the limit for the survey questions), six versions of the survey questionnaire were

⁴ Note that although the quota was set up for 36 interviews, 37 respondents were interviewed.

We had required certain quota group respondents to go through certain versions of the questionnaire and one interviewer made an error and had to do an additional interview.

⁵ The survey questions were administered at the beginning of the cognitive interviews because the original quantitative survey had been conducted 5 months before.

created and each respondent received one of these versions.⁶ Question issues considered important by the research team were present in all six questionnaires (for example, acquiescence in the 12 agree / disagree questions). 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.⁷

Each of the six questionnaire versions consisted of questions in CAPI, CATI and CAWI. This meant that each respondent in the cognitive interviewing phase received questions in all three modes within the interview, in that order. These survey questions were administered in standard quantitative survey fashion and mimicked the modes as closely as possible. This involved the interviewer sitting with the respondent face-to-face (for the CAPI component), being in a different room in the respondent's home and talking over a phone (for the CATI component) and having the respondent use the interviewer's laptop completely on his/her own (for the CAWI component).

Although every respondent experienced the three modes, respondents 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

⁶ The quota distinctions were kept for each version of the questionnaire (i.e., one half of those who had agreed to opposite statements, one quarter who had showed non-differentiation on a ranking task and one quarter who were 'non-satisficers').

⁷ All 4 questions in end labelled format were in all six questionnaire versions. For the 3 versus 7 or 8 response option experiments, four of the six versions focused on the satisfaction questions and two of the six versions focused on 'type of dwelling' and 'years lived in area'.

one mode and the rest in the other modes within a particular version of the questionnaire. Which questions were asked in particular modes were randomly varied by version of the questionnaire, but the mode order remained constant (as explained earlier) to ease the task difficulty for the cognitive interviewers.

After this 10 minute administration of one of the six versions of the questionnaire with a respondent, there was a transition to the cognitive interviewing. Here the cognitive interviewer made use of retrospective think alouds. This was done by reminding the respondent of the survey question, data collection mode, the respondent's answer and any behaviour displayed whilst answering, for example hesitation. The retrospective approach was necessary so that respondents could experience the three modes of data collection as realistically as possible. The more widely used concurrent think aloud technique, which is carried out at the time a respondent is answering a survey question, would have distracted respondents from their experience of the different modes. Pre-written probes were then used on some questions to explore, for example, task difficulty. This part of the interview took 50 minutes. The cognitive interviewing was conducted by both researchers and survey interviewers who were trained and highly experienced in using cognitive interviewing methods. Together they interviewed respondents in their homes in six locations across England and Scotland.

It is important to note that these cognitive interviews differed from traditional cognitive interviews. Respondents were not asked about their understanding of the survey questions, but rather solely how they came up with their answers. This is because the original purpose of the

cognitive interviewing follow-up study was to gain a greater understanding of how mode effects happen even if they are not directly observed.⁸

All cognitive interviews were audio-recorded with respondent consent and then transcribed. Summarised transcript information was then entered into the qualitative data management program, 'Framework', which was used for analysis. Using this approach, summarised data are entered into a matrix which is set up to list themes explored as columns and respondents as rows. Qualitative data can then be read vertically, by theme and across all respondents/cases, or horizontally within a case. Here the four researchers involved in the analysis explored the themes behind respondents' verbal reports. The goal was to look for anything in the response process that was seen to differ by mode, including respondent satisficing.

With the cognitive interview data, it was feasible to make a distinction between 'clear' satisficing, 'possible' satisficing and no satisficing. Any cases which were not obvious (or 'clear' satisficing) were categorized as 'possible' satisficing. The distinctions can be seen more clearly with the examples given in Figure 1 and the categorisations were reviewed by the team leader to ensure reliability of categorisation among the researchers.

Clear satisficing

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

⁸ For a full description of the cognitive interviewing methodology used and some of its differences and innovations compared to standard cognitive interviewing, see Gray, Blake and Campanelli (2014).

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

Possible satisficing

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

Figure 1. Examples of ‘clear’ and ‘possible’ satisficing from the cognitive interview data

2.3. THE SURVEY QUESTIONS USED

The full wording and source of all the survey questions which are discussed in this paper are in the Appendix.

2.3.1. Example 1: Acquiescence

The quantitative experiment involved three short multi-item scales each with four items (with five point agree/disagree response options). These included a scale covering aspects of the respondent’s neighbourhood, a sensitive scale focusing on mental health patients or former prisoners living in the community, and a scale about the thoroughness of preparation one would do before making an important financial decision. Each scale contained both positive and

negative statements. The neighbourhood scale contained two statements which were meant to be exact opposites: ‘properties in poor state of repair’ versus ‘properties well kept’. For the sensitive scale, the opposites were less exact (‘would worry if mental health patients moved into neighbourhood’ versus ‘mental health patients have just as much right to live in neighbourhood’). The financial decision scale was more reflective of standard multi-item scales where statements represent opposites of the latent construct (e.g., the financial scale has ‘rarely read small print’ versus ‘do a lot of research’). Exact opposites in multi-item scales are often discouraged as this creates too much redundancy which can be seen as tedious by respondents.

The questions used in the cognitive interviews included all 3 scales from the quantitative experiment.

2.3.2. Example 2: End-Labelled versus Fully-Labelled Scales

The quantitative experiment included four questions in both fully-labelled and end-labelled format. Two of these were satisfaction questions and two were behavioural frequency questions.⁹

The cognitive interviews looked at all four questions but only in the end-labelled format. In addition, because the quantitative findings showed that ‘satisfaction with the economy’ behaved differently from other satisfaction questions, a new question ‘satisfaction with the environment’ was also included.

⁹ In survey practice it would be very rare for behavioural frequency questions to appear in end-labelled format, but a part of the research grant was to ensure that both behavioural and factual questions were tested as well as subjective, non-factual questions.

2.3.3. Example 3: 3 versus 7 or 8 Response Options

The quantitative experiment included six questions contrasting a 3 versus 7 or 8 response option version (two satisfaction questions and four factual questions).

The cognitive interviews focused on the two satisfaction questions and the two factual questions ‘type of dwelling’ and ‘years lived in area’ which had puzzling quantitative results. In the case of the two factual questions, respondents were asked the questions in the 3 response option format during the survey question administration, but the cognitive interviewers were instructed to present respondents with the 7 or 8 category version during the cognitive interview.

3. Results

In this section, we present results from the quantitative experiment and the cognitive interview results for each of the three examples.

3.1.1. Example 1: Quantitative Experiment Results

The quantitative data showed that respondents had agreed with opposite statements. As shown in Table 1, agreement with opposite statements becomes more prevalent as the exactness of the opposite statements was loosened, an expected finding. Using Cronbach’s alpha, the internal consistency (homogeneity) of the three scales was 0.75 (neighbourhood scale), 0.71 (sensitive scales) and 0.33 (financial decision scale). A principle components analysis of the scales showed the neighbourhood and sensitive scales to be uni-dimensional but not the financial decision scale. A possible explanation for the poorer performance of the financial decision scale was found by the cognitive interviewing. Two of the financial decision items used the problematic term,

‘rarely’ which respondents found hard to understand in combination with an agree/ disagree scale.

Table 1. Percentage of respondents agreeing to opposite statements

Scale	Percent who agreed	Percent who only agreed	Total	n
	to all items within a scale	to two opposite statements within a scale		
<i>Neighbourhood</i>				
<i>(most opposite)</i>	1.9	1.5	3.4	476
<i>Sensitive</i>	8.2	24.5	32.7	464
<i>Financial Decision</i>				
<i>(least opposite)</i>	11.9	32.9	44.8	478

3.1.2. *Example 1: Cognitive Interview Results*

Across all three agree / disagree scales administered at the beginning of the cognitive interviews there were 23 instances where respondents agreed to opposite statements.^{10 11} What was

¹⁰ Some respondents agreed to more than one set of opposite statements. These figures are based on the number of instances of agreement, not the number of respondents.

¹¹ 32 instances were originally found, but 9 instances were removed because the cognitive interviewing on the financial decision questions showed that use of the negative word ‘rarely’ in two of these questions caused 5 respondents to incorrectly agree to opposite statements. The cognitive interviewing also showed that lack of a ‘not applicable’ option caused another 4 respondents to incorrectly agree to opposite statements.

surprising was that 21 of these could not be attributed to acquiescence at all, as respondents gave very clear and justifiable reasons for why they chose the answer they did. For example, one respondent strongly agreed to N36 (more properties in poor repair) because some houses could do with some work and agreed to N38 (more properties well kept up) because *“in this village . . . it’s like half and half. There is a bit [that] . . . wants doing up and there’s other [parts which don’t]”* (SA04: Female, 50 to 59, no qualifications, employed, very low income, White British).

Of the remaining two cases, there was one clear and one possible case of acquiescence. The clear case shows a respondent unable to justify her answer to the neighbourhood question, ‘Compared to other neighbourhoods this neighbourhood has more properties that are in a poor state of repair’. In the cognitive retrospective think aloud she explained, *“I think I don’t understand that, I just say agree”* (JF03: Female, 30 to 39, no qualifications, low income, with poor English as rated by the interviewer). This respondent’s behaviour is in line with both the cultural norms theory (Javeline 1999) and category fallacy theory (see Warnecke et al. 1997) for acquiescence. The possible case of acquiescence was a respondent who had ambivalent feelings on the question about ‘family’s safety with former prisoners living in the area’ and commented that it was hard to choose agree or disagree. Interestingly, respondents with similar views to her chose the middle response option, thus her choice of ‘agree’ could be a type of acquiescence (SA03: Female, 40 to 49, higher education below degree level, employed, low income, White British).¹²

¹² Similar results were found with an exploration of non-differentiation on a battery of rating questions. The cognitive interviews found that most respondents displaying non-differentiation did so for valid and justifiable reasons. And in answer to a probe about difficulty, cognitive

3.2.1. *Example 2: Quantitative Experiment Results*

As shown in Table 2, the question, ‘satisfaction with democracy and personal freedom’, replicates the work of Dillman and Christian (2005) with the fully-labelled scale showing significantly more positive responses (as measured by a t-test of differences between the means of the two distributions) and the end-labelled scale showing significantly more middle response option choices (as measured by a t-test of the differences in percentages). The pattern is not replicated with a ‘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). In addition, there were no differences between formats in the selection of the middle response option for this question.

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 respondents in the fully-labelled version were significantly more likely to choose response options from the beginning, rather than the end, of the scale and respondents in the end-labelled version were significantly more likely to choose the middle response option.

Because of the similarity of the results of the behavioural frequency questions with the first satisfaction question and with Dillman and Christian (2005), it would be easy to conclude that

respondents found it difficult rather than easy to choose which of the non-differentiated items was most important.

similar satisficing processes were at work across all of these questions. More details on this are given in Section 4.

Table 2: Differences in means and choice of middle response option on end-labelled versus fully-labelled scales

Satisfaction with Democracy and Personal Freedom (On the whole, how satisfied are you with the way democracy and personal freedom work in Great Britain?)						
End-labelled			Fully-labelled			T-test between % in middle response options
Labels		%	Labels		%	
Very satisfied	1	5.0	Very satisfied	1	9.9	p < .001
	2	14.0	Moderately satisfied	2	35.4	
	3	16.9	Slightly satisfied	3	15.6	
	4	26.7	Neither satisfied nor dissatisfied	4	14.3	
	5	23.2	Slightly dissatisfied	5	11.9	
	6	7.5	Moderately dissatisfied	6	7.8	
Very Dissatisfied	7	6.7	Very dissatisfied	7	5.2	

T-test: Means (End-labelled = 3.98, n=479; Fully-labelled = 3.27, n=463), p < .001

Satisfaction with State of the Economy (And on the whole, how satisfied are you with the present state of the economy in Great Britain?)

End-labelled		Fully-labelled	
--------------	--	----------------	--

Labels		%	Labels		%	T-test between % in middle response options
Very Satisfied	1	0.8	Very satisfied	1	0.2	
	2	0.6	Moderately satisfied	2	3.2	
	3	4.8	Slightly satisfied	3	6.4	
	4	7.9	Neither satisfied nor dissatisfied	4	7.1	Not significant
	5	22.5	Slightly dissatisfied	5	15.5	
	6	26.7	Moderately dissatisfied	6	21.9	
Very Dissatisfied	7	36.5	Very dissatisfied	7	45.7	

T-test: Means (End-labelled = 5.77, n= 479; Fully-labelled = 5.83, n=466), Not significant

Frequency of grocery shopping (The next question is about grocery shopping which includes food, drinks, cleaning products, toiletries and household goods. How often do you personally do grocery shopping?)

End-labelled			Fully-labelled			T-test between % in middle response options
Labels		%	Labels		%	
Everyday	1	7.3	Every day	7	2.8	
	2	18.8	On most days	7	22.5	
	3	33.7	Once a month	7	59.0	

	4	21.8	Two or three times a month	7	7.7	p < .001
	5	9.8	Once a month	7	2.6	
	6	5.6	Less than once a month	7	2.4	
Never	7	2.9	Never	7	3.0	

T-test: Means (End-labelled = 3.37, n=478; Fully-labelled = 3.04, n=466), p < .001

Amount of hot beverages (In the last two weeks, how many teas, coffees and other hot beverages have you purchased outside the home?)

End-labelled			Fully-labelled			T-test between % in middle response options
Labels	%		Labels	%		
None	0	36.1	None	0	33.5	p < .001
	1	14.8	Less than 5 but more than one	1	41.1	
	2	17.7	Less than 10 but more than 5	2	15.5	
	3	13.4	Less than 15 but more than 10	3	4.7	
	4	7.3	Less than 20 but more than 15	4	1.9	
	5	5.4	Less than 25 but more than 20	5	0.6	
More than 25	6	5.2	More than 25	6	2.6	

T-test: Means (End-labelled = 1.78, n=479; Fully-labelled = 1.13, n=465), p < .001

3.2.2. Example 2: Cognitive Interview Results

The cognitive interviews shed light on the reasons for some of the response patterns which in the quantitative analysis suggested satisficing behaviour on the end-labelled scales. For the satisfaction questions (satisfaction with ‘democracy and personal freedom’, ‘the economy’, and ‘the environment’), it was clear that respondents understood the 7-point scales and were using these in the intended fashion. But using the scales proved to be a challenge for the behavioural frequency questions.

On ‘frequency of grocery shopping’ question, respondents were told that ‘1 is every day and 7 is never’. This question is relatively straightforward for respondents who do grocery shopping every day or never. The remaining respondents either tried to use the scale or created a different interpretation of it from that intended. Among those who tried to answer with the scale, ‘Once a week’ was a common shopping frequency, but their answers for ‘once a week’ varied between code ‘3’ to code ‘6’. Some respondents interpreted the numbers to mean the number of shopping trips: an answer of 5 means “*five times in a month*” (JF03: Female, 30 to 39, no qualifications, low income, with poor English as rated by the interviewer) or the interval between shopping: “*Four, I thought, I, you know, I just interpreted it as every fourth day*”. (SA01: female, 30 to 39, secondary school, employed, low income, White British).

For the ‘amount of hot beverage purchased’ question, respondents were told that ‘0 is none and 6 is more than 25’. Again the question is relatively straightforward for respondents who never buy hot drinks outside the house or who buy 25 or more. The common theme among respondents was giving the actual number of drinks they had had rather than using the scale. Other interpretations were that the scale numbers were multiples of 2 or 5.

In terms of middle response options, all respondents, except for one, believed they were choosing a valid number on the behavioural frequency scales rather than satisficing or meaning

‘don’t know’. This differed for the satisfaction questions. Here some cognitive respondents chose the middle response option as a valid answer and some chose the middle response option to ‘satisfice’ (see Figure 2) or indicate ‘don’t know’.

Clear satisficing

Interviewer: “*You said number four. Tell me why.*” Respondent: “*A cop-out one I suppose.*” (JF05: Female, 30 to 39, postgraduate degree, employed, high income, White British)

“*I just thought if I went for middle for diddle [both laughing] it, it would be okay.*” (SA04: Female, 50 to 59, no qualifications, employed, very low income, White British)

Possible satisficing

“*Yes, just the middle of the road, yeah.*” (JF01: Male, 60 plus, high school equivalent, high income, White British)

“*It’s not something I’ve got a very strong opinion on.*” (AR04: Female, 30 to 39, some higher education, employed, very low income, White – other)

“*Selfishly, it don’t affect me, so I’m in between with that one.*” (AR01: Female, 20-29, no qualifications, employed, no info on income, White British)

Figure 2. Examples of clear and possible satisficing on the end-labelled satisfaction questions

3.3.1. *Example 3: Quantitative Experiment Results*

All six questions used in the quantitative experiment showed significant differences between answers given on the 3 versus the 7 or 8 response option versions collapsed into 3 options. In retrospect, these significant differences by format made sense for the two satisfaction questions.

With response contraction bias (Tourangeau, Rips and Rasinki 2000) and fewer options to choose from, one would expect more middle response options on the 3 option rather the 7 option version and indeed this was found. But the significant differences in the two factual questions which were basic facts ('type of dwelling' and 'years lived in area') were not expected (see Table 3). These differences were not taken too seriously. Although respondents had been randomly assigned to question version, the two 'sub-samples' created could still differ with respect to 'type of dwelling' and 'years lived in area'. In addition, it is known that the significance of the chi-square statistic is highly influenced by sample size.

Table 3. Comparing two factual questions in 7 or 8 response options format with the same questions collapsed into 3 response option format

Type of Dwelling					
8 Response Options (n = 466)			3 Response Options (n = 479)		
Labels	%	Collapsed %	Labels	%	Significance Test
Detached house	26.0	87.3	House	83.5	$\chi^2 = 13.264,$ $p < .01$
Semi-detached house	31.5				
Terraced house	23.4				
Bungalow	6.4				
Flat in a block of flats	7.7	11.8	Flat or maisonette	11.7	
Flat in a house	2.8				
Maisonette	1.3	0.9	Other	4.8	
Other	0.9				

n=466			n=479		
Years lived in area					
7 Response Options (n = 479)			3 Response Options (n = 466)		
Labels	%	Collapsed %	Labels	%	Significance Test
Less than 12 months			$\chi^2 = 6.270,$ $p < .05$		
12 months or more but less than 2 years	3.4				
2 years or more but less than 3 years	6.9	15.9	Less than 3 years	11.8	
3 years or more but less than 5 years	5.6				
5 years or more but less than 10 years	7.7	26.1	3 years or more but less than 10 years	22.5	
10 years or more but less than 20 years	18.4				
20 years or longer	21.0	58.0	10 years or longer	65.7	
	37.0				
n=479			n=466		

3.3.2. Example 3: Cognitive Interview Results

The cognitive interviews studied the ‘type of dwelling’ and ‘years lived in area’ questions further. In the survey question part of the cognitive interviews, the respondents were given the 3

response option versions of these questions. During the cognitive interviewing component, interviewers then presented respondents with a showcard with the more detailed response options and asked the survey question again. The cognitive interviewers were given standardised probes for exploring any inconsistencies.

Although no inconsistencies were found, a further scrutiny of the transcripts suggested that there was confusion and difficulty with these basic facts. On ‘type of dwelling’, as shown in Figure 3, the themes apparent in respondent comments were lack of clarity of some of the answer choices in the 8 response option version and regional and cultural differences.

Theme of unclarity of response options

“Erm [hesitates] a semi-detached house I think” (OM03: Female, 20 to 29, high school equivalent, employed, low income, White British)

“I’m not sure what a flat in a house is” (JC06: Male, 49 to 49, first degree, employed, high income, White British)

“What the hell difference is there between a maisonette and a flat and a block of flats, a flat and a house?” (OM06: Male, 50 to 59, postgraduate degree, employed, high income, White British) part of a very lengthy debate between the respondent and another household member about what the various categories of the 8 response option version meant.

Theme of regional and cultural differences

Other household member: *“I’ll call it a duplex, yeah.”* Respondent: *“Well, it’s what they call it in the South, isn’t it, where my family’s from. Oh, you live in a maisonette.”* (OM06: Male, 50 to 59, postgraduate degree, employed, high income, White British)

The respondent answered 'flat,' but the interviewer observed the dwelling to be a semi-detached house. Here the respondent said a dwelling has to be very large to be called a house (PC01: female, 39 to 39, higher education below degree level, employed, medium income, other ethnicity).

Figure 3. Respondents' difficulties with 8 response option type of dwelling question

On 'years lived in area', the themes apparent in respondent comments were difficulty in remembering the number of years, feeling stuck between two options, and feeling the 3 option short version was much simpler (the 7 option version was phrased with 'less than 12 months, 12 months or more but less than 2 years, 2 years or more but less than 3 years', etc.)

4. Summary and Discussion

The cognitive interviewing discussed in the paper was developed to look at mode effects in a mixed modes survey experiment. Although the research project was originally designed to look at the interactions of question format and mode, this paper has focused on results from the format experiments, regardless of mode. More specifically it investigated the intersection between the quantitative results and what was found in subsequent the cognitive interviewing.

The results suggest that cognitive interviewing can be useful after a quantitative survey as well as before. Different methods and a different focus were used in this follow-up study. Cognitive interviewing at the pretesting stage typically focuses on the cognitive stages of comprehension, recall, judgment and response. In this follow-up study, no investigation of the cognitive stages was done, but rather the focus was on how and why respondents answered as they had. Cases of clear satisficing, possible satisficing and no satisficing could be

distinguished. And interestingly, despite the different cognitive interview methods and focus in the follow-up study, ambiguous terms were still discovered.

The cognitive interviewing suggests that all three examples of ‘typical’ quantitative conclusions drawn from the data may be incorrect. With respect to Example 1, the quantitative analysis predicted and found that the instances of agreement with opposite statements become more prevalent when the statements were less clearly exactly opposite. In multi-item scale construction, however, statements are meant to represent the opposites of the latent concept and ideally are not exact opposite wordings so as to minimise respondent boredom. The cognitive interview results suggest that true acquiescence may only account for a fraction of the cases of respondents who agree to opposite statements. Most importantly this casts doubt on the practice in multi-item scale research of handling acquiescence cases differently (through deleting or segregating them for use as a covariate) as many of these cases may be valid responses. Although only touched on in footnote 12, cognitive interviews suggested a similar conclusion with respect to non-differentiation in a battery of rating questions. Most instances on such non-differentiation reflected legitimate responses.

The quantitative findings for Example 2 suggest that respondents were displaying the same answer patterns for two behavioural frequency questions as for a satisfaction question (i.e., more choices at the top / positive end of fully-labelled scales and more choice in the middle response option on end-labelled scales). The cause of this pattern could be conjectured to arise from satisficing: primacy on the fully-labelled scales and middle response option short-cutting on the end-labelled scales. This explanation would hold for both the satisfaction and behavioural frequency questions. However, the cognitive interviews showed how differently the original satisfaction question on democracy and personal freedom (and an additional one on the

environment) were answered in comparison to the behavioural frequency questions. For the satisfaction questions, respondents understood the 7 response options and some who chose the middle category were labelled in the cognitive interviewing as ‘clearly’ or ‘possibly’ satisficing. In contrast, many of the respondents on the behavioural frequency questions were not using the scale as intended, creating their own interpretations while still trying to give a valid answer. The cognitive interviewing suggests that the behavioural frequency questions contain a lot of measurement error, but almost no satisficing. This was not noticed in the implementation of the quantitative mixed modes experiment. When questions are difficult, respondents do not always ask questions (DeMaio and Rothgeb 1996) and depend more on heuristics (Schwarz, Hippler, and Noelle-Neumann 1992). Also the quantitative survey environment can suppress the appearance of problems even with difficult questions (Beatty 2004).

In Example 3, the significant quantitative differences between the 3 and the 7 or 8 response option versions of ‘type of dwelling’ and ‘how long lived in area’ could have easily been dismissed as disparities in the two experimental groups due to random chance or the tendency of the significance test on the chi-square statistic to be influenced by sample size. The cognitive interviewing paints a different picture. Although no inconsistencies were found with the cognitive interviewing respondents between the 3 response option and 7 or 8 response option versions, it was clear that these commonly used questions are more difficult than expected. Not all respondents are knowledgeable and able to interpret the different types of housing response options in the same way. Similarly, answering a more detailed version of ‘years lived area’ means you have to be more precise in the information which you retrieve from memory and opens up the possibility for respondents desire to choose more than one response option.

In summary, these examples suggest that some typical quantitative indicators of satisficing can be an over-estimate of the true extent of the problem (e.g., with acquiescence and non-differentiation), that similar quantitative distributions which imply satisficing can be an incorrect inference (e.g., with satisfaction versus behavioural frequency questions) and that dismissing quantitative differences on simple questions can be problematic (e.g., with long and short versions of ‘type of dwelling’ and ‘how long lived in area’).

As with any research, both the mixed modes survey experiment and the cognitive interviewing phase have limitations. 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 respondents were dissatisfied with the economy in 2009. The financial decision scale used the confusing term ‘rarely’. Fortunately, in this case the cognitive interviewing could separate those respondents who were confused by this term from those who agreed to opposite statements for other reasons. We are also aware that end-labelled behavioural frequency questions would rarely, if ever, be used in practice and were only included to complete the larger mixed modes project’s hypothesis of comparing subjective and factual questions in the same format.

Another potential limitation involves the implementation of the cognitive interviewing: 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 overall respondents definitely appeared to have thought about the issues initially and remembered these during the retrospective work. There could be a concern on the 3 versus 7 or 8 response option experiment where the cognitive respondents purposely were given the 7 or 8 response option version in the

cognitive interviews. In principle this could lead to respondents trying to be helpful and suggest problems with the questions when given the longer version. Although this “helpfulness” does happen in standard cognitive interviewing, we believe this was not the case in this study as respondents were always focused on their answers and respondents’ understandings of the survey questions were never probed.

Other potential limitations relate to the respondents. All of the cognitive interview respondents had been interviewed twice previously and still agreed to taking part in the cognitive interviews thus making it their third interview in the project. We cannot know for certain whether these cooperative cognitive interview respondents were different from the respondents who had not agreed to participate. (1) However, the realistic nature of the cognitive interviews can be seen from that fact some respondents did satisfice (see Section 2.2), while others talked with pride about telling their true feelings and others confessed to the interviewers about how bad their opinions must seem (see Nicolaas et al. 2011, for specific examples). (2) In contrast, it must be remembered that the sample of respondents used for the cognitive interviews purposively over-sampled respondents who showed less than optimal behaviour on the mixed modes survey experiment.

Overall, we feel that the use of cognitive interviewing after a quantitative survey has been a useful exercise. This is not to exclude the usefulness of cognitive interviewing as a pretesting method, but rather to suggest its additional usefulness as a follow-up study to explore quantitative findings, particularly in research projects with a methodological focus. In particular, we were able to explore examples of response behaviours and see the differences in interpretation between the cognitive interviewing and quantitative views. Using cognitive interviewing after a survey to enhance the accuracy of the interpretation of quantitative findings,

is relatively novel. We hope this paper inspires more use of cognitive interviewing as a follow-up study to a quantitative survey, particularly when the quantitative results are unexpected.

Acknowledgments

The support of the UK Economic and Social Research Council (ESRC) is gratefully acknowledged. This work was funded by grant number RES-175-25-0007. The authors also thank core project team members: Gerry Nicolaas (Ipsos MORI), Peter Lynn and Annette Jäckle, (Institute for Social and Economic Research).

References

- AAPOR. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. AAPOR, Ann Arbor, Michigan.
http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156 (2011). Accessed 06 January 2012
- Ayidiya, S., McClendon, M.: Response effects in mail surveys. *Public Opinion Quarterly*. 54, 229-247 (1990)
- Beatty, P.: The dynamics of cognitive interviewing. In Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., Singer, E. (eds.) *Methods for Testing and Evaluating Survey Questionnaires* (eds.), pp. 45-66. Wiley, Hoboken, NJ (2004)
- Beatty, P. Willis, G.: Research synthesis: The practice of cognitive interviewing. *Public Opinions Quarterly*. 71, 287-311 (2007)

- Campanelli, P., Blake, M., Gray, M., Hope, S.: Mixed modes and measurement error: Using cognitive interviewing to explore the results of a mixed modes experiment. Paper presented at the 65th Annual Conference of the American Association for Public Opinion Research (2010)
- Carrasco, L.: 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. U.S. Census Bureau, Washington, DC (2003)
- Converse, J., Presser, S.: Survey questions: Handcrafting the standardized questionnaire. Sage, Thousand Oaks, CA (1986)
- Davis, E. Nicolas, C., Waters, E., Cook, K., Gibbs, L Gosch, A., Ravens-Sieberer, U.: 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 (2007)
- DeMaio, T., Rothgeb, J.: Cognitive interviewing techniques in the lab and in the field. In Schwarz, N., Sudman, S., (eds.) *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, pp. 177-196. Jossey-Bass: San Francisco, CA (1996)
- DeVellis, R.: *Scale development: Theory and applications*, 3rd Edition. Sage, Thousand Oaks, CA. (2012)
- Dillman, D. Christian, L.M.: Survey mode as a source of instability in responses across surveys. *Field Methods*. 17(1), 30-52 (2005)
- Fowler, F.J., Jr.: *Improving survey questions: Design and evaluation*. Sage, Thousand Oaks, CA (1995)

- Gerber, E.: The view from anthropology: Ethnography and the cognitive interview. In Sirken, M., Herrmann, D., Schechter, S., Schwarz, N., Tanur, J., Tourangeau, R. (eds.) *Cognition and Survey Research*. Wiley, New York (1999)
- Gray, M., Blake, M., Campanelli, P.: The use of cognitive interviewing methods to evaluate mode effects in survey questions. *Field Methods*. 26(2), 156-171 (2014)
- Goerman, P.: Adapting cognitive interview techniques for use in pretesting Spanish language survey instruments. Statistical Research Division Research Report Series, Survey Methodology #2006-3. U.S. Census Bureau, Washington, DC (2006)
- Goggin, S. and Stoker, L.: Optimal scale length and single-item attitude measures: Evidence from simulations and a two-wave experiment, APSA 2014 Annual Meeting Paper. <http://ssrn.com/abstract=2455794> (2014). Accessed 10 July 2014
- Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., Jäckle, A.: 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. 20141-20. ISER, University of Essex (2014)
- Jakwerth, P., Frances Stancavage, F. Reed, E.: 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 (1999)
- Javeline, D.: Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly*. 63, 1-28 (1999)
- Krosnick, J.: Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*. 5, 213-236 (1991)

- Krosnick, J.: The threat of satisficing in surveys: The shortcuts respondents take in answering questions, *Survey Methods Newsletter* 20(1). National Centre for Social Research, London, UK (2000)
- Krosnick, J., Fabrigar, L.: Designing rating scales for effective measurement in surveys. In Lyberg, L., Biemer, P., Collins, M., Leeuw, E., Dippo, C., Schwarz, N., Trewin, D. (eds.) *Survey Measurement and Process Quality*, pp. 141-64. Wiley, Hoboken, NJ (1997)
- Landsberger, H., Saavedra, A.: Response set in developing countries. *Public Opinion Quarterly*. 31, 214-229 (1967)
- Lenski, G., Leggett, J.: Caste, class, and deference in the research interview. *American Journal of Sociology*. 65, 463-467 (1960)
- Levin, K., Willis, G., Forsyth, B., Norberg, A., Kudela, M., Strack, D., Thompson, F.: Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods*. 3(1), 13-25 (2009)
- Lynn, P., Hope, S., Jäckle, A., Campanelli, P., Nicolaas, G.: 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 (2012)
- McBride, L., Moran, G.: Double agreement as a function of ambiguity and susceptibility to demand implications of the psychological situation. *Journal of Personality and Social Psychology*. 6, 115-118 (1967)
- Narayan, S., Krosnick, J.: Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*. 60, 58-88 (1996)

- Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A., Lynn, P.: Is it a good idea to optimise question format for mode of data collection? Results from a mixed modes experiment. ISER Working paper no. 2011-31. ISER, University of Essex (2011)
- Padilla, J.: The use of cognitive interviewing to assess “construct overlap” in cross-cultural surveys. Paper presented at second Conference of European Survey Research Association, Prague (2007)
- Schuman, H., Presser S.: Questions and Answers in Attitude Surveys. Academic Press, New York (1981)
- Schwarz, N., Hippler, H., Elisabeth Noelle-Neumann, E.: A cognitive model of respondent-order effects in survey measurement. In Schwarz, N., Sudman, S. (eds.) Context Effects in Social and Psychological Research. Springer-Verlag, New York (1992)
- Toepoel, V. Dillman, D.: Words, numbers, and visual heuristics in web surveys: Is there a hierarchy of importance? *Social Science Computer Review*. 29(2), 193-207 (2011)
- Thomas, R., Uldall, B., Krosnick, J.: How many are too many?: Number of response categories and validity. Paper presented at the 59th Annual Conference of the American Association of Public Opinion Research (2004)
- Tourangeau, R.: Cognitive science and survey methods: A cognitive perspective. In Jabine, T., Straf, M., Tanur, J., Tourangeau, R. (eds.) *Cognitive Aspects of Survey Design: Building a Bridge Between the Disciplines*, pp. 73–100. National Academy Press, Washington, DC (1984)
- Tourangeau, R., Rips, L.J., Rasinski, K.: *The Psychology of Survey Response*. Cambridge University Press, Cambridge (2000)

Warnecke, R., Johnson, T., Chavez, N., Sudman, S., O'Rourke, D., Lacey, L., Horm, J.:

Improving question wording in a survey of culturally diverse populations. *Annals of Epidemiology*. 7, 334-342 (1997)

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

Appendix: The Full Question Wording and Source for each Question

Agree/Disagree Scales

Each of the items in each of the three scales used the following five categories: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree.

Neighbourhood scale

The next four questions are about the extent to which you agree or disagree with statements about your neighbourhood. Here is the first statement.

N35. This neighbourhood is not a bad place to live.

N36. Compared to other neighbourhoods, this neighbourhood has more properties that are in a poor state of repair.

N37. Compared to other neighbourhoods, this neighbourhood does not suffer from things like litter, dog mess and graffiti.

N37alt. Compared to other neighbourhoods, there are many things I do not like about this neighbourhood.

N38. Compared to other neighbourhoods, this neighbourhood has more properties that are well kept.

Questions adapted from the UK Southern Housing Association questionnaire.

Note N37 was used in the quantitative survey, N37alt was used in cognitive interviewing to create another pair of opposite statements.

Sensitive scale

How strongly do you agree or disagree with the following 4 statements.

N52. I would worry if housing were provided near my home for people with mental health problems leaving hospital.

N53. People who have serious mental health problems have just as much right to live in my neighbourhood as any other people.

N54. I would be concerned for my family's safety if housing were provided near my home for people who were leaving prison.

N55. People who have been in prison have just as much right to live in my neighbourhood as any other people.

Questions adapted and extended from two questions from the 2006 British Social Attitudes survey.

Financial scale

To what extent do you agree or disagree with the following 4 statements about making important financial decisions such as taking out a mortgage, loan or pension.

FM64. I would rarely read all the small print before making important financial decisions.

FM65. I would do a lot of research before making an important financial decision.

FM66. I would rarely talk to a financial advisor before making an important financial decision.

FM67. I definitely would talk to family and friends before making an important financial decision.

Questions extended from the UK Attitudes to Pensions Survey.

End-labelled versus Fully-labelled Experiment

Questions displayed in end-labelled format

GB16. On the whole, how satisfied are you with the way democracy and personal freedom work in Great Britain, where 1 is very satisfied and 7 is very dissatisfied?

Question from European Social Survey, 2006 with addition of 'and personal freedom' to make the question more difficult.

GB17. And on the whole, how satisfied are you with the present state of the economy in Great Britain, where 1 is very satisfied and 7 is very dissatisfied?

Question from European Social Survey, 2006.

GB17alt And on the whole, how satisfied are you with the present state of the environment in Great Britain, where 1 is very satisfied and 7 is very dissatisfied?

Question newly created for cognitive interviewing as GB17 in the quantitative data did not behave as expected.

FM68. The next question is about grocery shopping which includes food, drinks, cleaning products, toiletries and household goods. How often do you personally do grocery shopping, where 1 is every day and 7 is never?

Question newly created for this research to fit hypotheses.

FM74. In the last two weeks, how many teas, coffees and other hot beverages have you purchased outside the home, where 0 is none and more than 25 is 6?

Question newly created for this research to fit hypotheses.

3 versus 7 or 8 Categories Experiment

Questions displayed in face-to-face version ^a

FM75. Which of these best describes your home? Would you say a . . . (READ OUT) . .

8 Category Version	3 Category Version
<p>Detached house</p> <p>Semi-detached house</p> <p>Terraced house</p> <p>Bungalow</p> <p>Flat in a block of flats</p> <p>Flat in a house</p> <p>Maisonette</p> <p>Or other?</p> <p><i>Question taken from the Survey of Public Attitudes and Behaviours Towards the Environment, 2007.</i></p>	<p>House</p> <p>Flat or maisonette</p> <p>Or other?</p>
FM82. How long have you lived in this area?	Would you say . . . READ OUT . . .
7 Category Version	3 Category Version
<p>Less than 12 months</p> <p>12 months or more but less than 2 years</p> <p>2 years or more but less than 3 years</p> <p>3 years or more but less than 5 years</p> <p>5 years or more but less than 10 years</p> <p>10 years or more but less than 20 years</p> <p>20 years or longer</p>	<p>Less than 3 years</p> <p>3 years or more but less than 10 years</p> <p>10 years or longer</p>

<p><i>Question taken from the British Crime Survey, 2006.</i></p>	
<p>N43. I would like you to tell me how satisfied or dissatisfied you are with the local household waste collection, recycling collection and other recycling collection points. Would you say you are . . . (READ OUT). . .</p> <p><i>Question taken from Citizenship survey, 2007 and adapted to make it more difficult.</i></p> <p>N44. And how satisfied or dissatisfied are you with street cleaning are . . . (READ OUT). . .</p> <p><i>Question taken from Citizenship survey, 2007.</i></p>	
<p>7 Category Version</p>	<p>3 Category Version</p>
<p>Very satisfied</p> <p>Moderately satisfied</p> <p>Slightly satisfied</p> <p>Neither satisfied nor dissatisfied</p> <p>Slightly dissatisfied</p> <p>Moderately dissatisfied</p> <p>Very dissatisfied</p>	<p>Satisfied</p> <p>Neither satisfied nor dissatisfied</p> <p>Dissatisfied</p>

^a Note: In the quantitative experiment there was a further sub-experiment where the 7 or 8 answer categories were either read out or placed on a show card in CAPI. The results did not differ between the show card and no show card conditions for either the 'type of dwelling' or the 'years lived in area' questions.