ISER Working Paper Series

Is it a good idea to optimise question format for mode of data collection? Results from a mixed modes experiment

Gerry Nicolaas
National Centre for Social Research

Pamela Campanelli
The Survey Coach

Steven Hope
University College London

Annette Jäckle
Institute for Social and Economic Research
University of Essex

Peter Lynn
Institute for Social and Economic Research
University of Essex

No. 2011-31
December 2011

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

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

www.iser.essex.ac.uk

Non-technical summary

It is common practice for survey designers to change how questions are asked depending on whether the questions are asked face-to-face, over the telephone or included on a self-completion form. For example, it is possible to show long lists of response options on self-completion forms and on cards for respondents in face-to-face interviews. However, such long lists are not feasible in telephone interviews and this often leads to the use of radically different question formats in telephone interviews compared to face-to-face interviews and self-completion forms.

There has been limited research into the impact that changes in question formats will have on how people will answer in telephone interviews, face-to-face interviews and self-completion forms. This paper analyses the results from an experiment testing the effect of changing two commonly used question formats, contrasting the answers in these three different settings. We also carry out further in-depth interviews to explore possible causes for answering questions differently depending on the question format and the setting in which the questions were asked.

Our results show that changing question formats can change how people answer questions. However, we also find differences when we use the same question format in the telephone interviews, face-to-face interviews and web format. These results suggest that differences in answers between telephone interviews, face-to-face interviews and web forms are not only caused by changes in question format. We discuss other possible causes for these differences.

Is it a good idea to optimise question format for mode of data collection? Results from a mixed modes experiment

Gerry Nicolaas, National Center for Social Research
Pamela Campanelli, The Survey Coach
Steven Hope, University College London
Annette Jäckle, University of Essex
Peter Lynn, University of Essex

Abstract

It is common practice to adapt the format of a question to the mode of data collection. Multi-coded questions in self-completion and face-to-face modes tend to be transformed for telephone into a series of 'yes/no' questions. Questions with response scales are often branched in telephone interviews, that is, converted into two or more questions, each with shorter response lists. There has been limited research into the impact of these format differences on measurement, particularly across modes. We analyse data from an experiment that contrasted these question formats in face-to-face, telephone and web surveys. The study also included a cognitive interviewing follow-up to further explore the quantitative findings.

Keywords: mixed mode survey, measurement error, question format, code all that apply, branching, NatCen Omnibus Survey, British Household Panel Survey

JEL classification: C81, C83

Acknowledgements: This research was funded by the UK Economic and Social Research Council (ESRC) Survey Design and Measurement Initiative for the project "Mixed Modes and Measurement Error" (award RES-175-25-0007). We are grateful to Rebecca Taylor (NatCen) and Alita Nandi (ISER) for their contributions to the initial study design, Chloë Robinson (NatCen) for managing the experimental data collection, Margaret Blake and Michelle Gray for their contributions to the design, management and analysis of the cognitive interviews, NatCen interviewers and operations staff for collecting and processing the data, and David Hussey (NatCen) for contributing to the initial analysis of the data. Finally, we would like to thank all those members of the public who gave their time and co-operation in responding to the surveys.

Contact: Gerry Nicolaas, National Center for Social Research, 35 Northampton Square, London EC1V 0AX, gerry.nicolaas@natcen.ac.uk

1 Introduction

Face-to-face interviewing has been the dominant data collection mode in the UK for national surveys of the general population since the Second World War. A move towards the cheaper mode of telephone interviewing which was witnessed in the USA and elsewhere in the 1970s failed to materialise in the UK because of the lack of a suitable sampling method (for an overview see Nicolaas and Lynn, 2002), and later attempts were hindered by low telephone response rates and non-coverage of mobile-only households (e.g. Nicolaas et al, 2000; Hope, 2005). The even cheaper mode of postal questionnaires has been limited in UK national surveys to specific groups for which lists of named individuals are available, such as benefit claimants. Finally, the more recent allure of web surveys, a mode which is very cheap and potentially fast, is undermined by severe under-coverage of the general population and low response rates (Lozar Manfreda et al, 2008; Smyth and Pearson, 2011).

Nonetheless, the increasing cost of fieldwork for face-to-face surveys coupled with declining survey budgets is pushing survey clients and survey practitioners to look for cheaper ways of collecting survey data. Given the specific limitations of each of the main modes, it is therefore not surprising that attention is drawn towards the use of mixed modes. As noted by de Leeuw (2005, p. 235), "Survey designers choose a mixed-mode approach because mixing modes gives an opportunity to compensate for the weaknesses of each individual mode at affordable cost. The most cost-effective method may not be optimal for a specific study. By combining this method with a second more expensive method the researcher has the best of both worlds: less costs and less error than in a unimode approach." Nonetheless, there is a trade-off that certain mixed mode designs can have with measurement error.

Mixing modes of data collection can reduce data comparability because people may answer questions differently depending on the mode (for further details on why different modes can produce different responses, see Jäckle et al, 2011). Moreover, many standard questions in the UK have been designed to be 'optimal' for the face-to-face mode and use formats which have to be adapted considerably when used in other modes, thus increasing the risk of differences in measurement by mode.

In this paper we examine the effects on measurement of adapting the question format for use in different modes, using data from a mixed mode survey that experimentally contrasted two

different question formats in computer assisted face-to-face interviews (CAPI), computer assisted telephone interviews (CATI), and computer aided web interviews (CAWI).

1.1 'Mark all that apply' versus 'yes/no' formats

A common question format is 'code all that apply' with a list of items displayed on a show card. This format is equivalent to 'mark all that apply' in self-completion modes such as postal questionnaires and web questionnaires. However, these questions are difficult to administer in telephone interviews that rely solely on aural communication and they are therefore often converted into a series of 'yes/no' questions for each item on the list. However, there is evidence to suggest that the 'yes/no' format and the 'mark all that apply' format are not functionally equivalent. Sudman and Bradburn (1982) were the first to recommend that the 'mark all that apply' format should be avoided because of the difficulty in interpreting what the absence of a check mark means (e.g., the item did not apply to the respondent, the respondent did not notice the item or the respondent did not know how to answer the item) and recommended the 'yes/no' format as a more suitable alternative. Several experimental studies have shown that for the same item the percentage of 'yes' responses in the 'yes/no' format is higher than the percentage choosing the item in the 'mark all that apply' format (Rasinski et al 1994, Smyth et al 2006, Thomas and Klein 2006). This finding has been replicated across various behavioural topics, languages and countries of residence (Thomas and Klein 2006). Smyth et al (2006) demonstrated that the 'yes/no' format takes longer to complete and seems to encourage deeper processing of the response options which results in a higher number of options being selected and less weak satisficing behaviour (e.g. primacy) and therefore in a larger number of options being selected. Smyth et al (2008) were the first to compare the two formats across modes (web and telephone) and found that the 'yes/no' format performed similarly across telephone and web modes, suggesting that this format is not prone to mode effects.

In this paper we replicate the research by Smyth et al (2006, 2008) and extend it by including a comparison with face-to-face interviewing in addition to telephone and web, by using probability samples of the general adult population rather than university students (thus increasing its generalisability). We also contrasted easy and difficult question series and employed cognitive interviewing techniques after the quantitative methods to enhance our understanding of the causes of differences in measurement.

1.2 Branching versus non-branching formats

Another common adaptation for telephone interviews that tends not to be used in other modes is branching[1] which involves splitting a question into two or more steps. For example, a five-point attitude scale can be divided into two steps: first respondents are asked to report the direction of their attitudes (positive, negative or neutral), then they are asked for the strength of their attitudes (e.g. very positive or just positive). Branching is also often used in telephone interviews when respondents are required to make difficult calculations or to provide information about income or expenditure; i.e. the task is decomposed into two or more steps.

However, there is some evidence to suggest that branching and non-branching formats are not functionally equivalent. Several studies have experimentally compared both formats and concluded that branching increases the accuracy and reliability of judgements (Armstrong et al 1975, Groves and Kahn 1979, Krosnick and Berent 1993, Yu et al 2003, Malhotra et al 2009), with the exception of Miller (1984) who concluded that non-branched formats were preferable. The studies examined different indicators of data quality. Branching produced fewer extreme responses in one study (Groves and Kahn 1979), and more extreme responses in another (Yu et al 2003). The authors however disagreed about the desirability of extreme responses; while Groves and Kahn (1979) argued that extreme responses were an indicator of bias, Yu et al (2003) were of the view that a reluctance of respondents to give extreme responses was an indicator of error. Branching in addition produced higher inter-item correlations (Groves and Kahn 1979), data with better predictive power (Yu et al 2003), and higher criterion validity (Malhotra 2009). In the study by Miller (1984), however, branching produced more item non-response and lower inter-item correlations. These mixed results are possibly due to differences in study design such as data collection modes and types of items, and in some cases format comparisons were confounded with other differences between questions. Groves and Kahn (1979) examined seven-point ratings scales in a telephone survey. Their comparison was confounded with differences in the scales between the two formats. Miller (1984) tested the two formats in a telephone experiment, using the same seven-point satisfaction scale in both formats and including a follow-up probe for those who selected the middle category in the branched format. The labelling however differed between the two formats, with only end labels and a middle label for the non-branched format and full verbal scales for the branched format. Krosnick and Berent (1993) conducted a meta-analysis of eight experimental studies contrasting branching bipolar attitude scales with fully verbally labelled

[1] Also referred to as 'unfolding' in the literature (Groves, 1979; Groves & Kahn, 1979; Miller, 1984; Sykes & Collins, 1988).

response options with partially labelled seven-point scales administered in a single reporting step. Yu et al (2003) carried out two experiments in the USA and Hong Kong using semantic differential scales in self-completion questionnaires. Malhotra et al (2009) used three separate studies, each using a different mode (face-to-face interview, telephone interview and web).

Although these studies used different modes, all of them compared the two question formats within modes rather than across modes. We are aware of one experiment which compared the formats across modes: the 1999 Welsh Assembly Election Study (Nicolaas et al, 2000). The questionnaire included 64 attitude questions with four-point or five-point response scales ranging from an extreme positive to an extreme negative response. The non-branched format was used in a face-to-face interview (show cards for 52 questions and a mini self-completion document for 12 questions) and a two-step branched format was used in a telephone interview for all but six questions, thus confounding question format and mode. The results showed that telephone respondents were more likely than face-to-face respondents to choose an extreme response for all 64 questions. Although question format was confounded with data collection mode, the authors concluded that it was very likely that the branching of responses in the telephone interview was the main cause for this strong tendency towards extremeness.

In this paper, we analyse differences in measurement between branching and non-branching across three modes: face-to-face interview, telephone interview and web questionnaire. To avoid confounding of question format and mode, we include the branched format in all three modes but the non-branched format is only used in the face-to-face interview and web questionnaire because it is not considered a feasible format for a telephone interview. In addition, we contrasted attitudinal and factual questions[2] and employed cognitive interviewing techniques to explore the underlying causes of differences in measurement.

2 Hypotheses

For both comparisons - that is (A) 'mark all that apply' compared to a series of 'yes/no' questions and (B) branching versus non-branching - our overarching hypothesis is that any observed

[2] It is generally assumed that factual, non-sensitive questions are less prone to mode effects than subjective questions. There is some evidence to support this (Lozar Manfreda and Vehovar, 2002; Schonlau et al, 2003). Van Soest, and Kapteyn (2009) found no differences between CAPI / CATI and web with respect to questions on checking and saving accounts and stocks and stock mutual funds.

differences in measurement across modes are due to changes in question format rather than an effect due to mode.

Our specific hypotheses for the comparison of 'mark all that apply' and a series of 'yes/no' questions are:

A1: Based on the findings of several studies (Rasinski et al 1994, Smyth et al 2006, Thomas and Klein 2006), when using a 'yes/no' series in CATI and 'mark all that apply' in CAPI and CAWI, we expect a higher percentage of items chosen in CATI than in CAPI and CAWI.

A2: When the 'mark all that apply' format is used in CAPI and CAWI, we expect no differences between CAPI and CAWI.

A3: Based on the findings from Smyth et al (2008), when a 'yes/no' series is used in all modes, we expect no differences between CATI, CAPI and CAWI.

A4: Using information about respondent question completion time as an indicator for respondent effort and based on the findings from Smyth et al (2006),

(A4a) We expect longer completion times with a 'yes/no' series than 'mark all that apply',

(A4b) We expect respondents who answer the 'mark all that apply' question under the mean response time to show evidence of primacy effects, and

(A4c) We expect respondents who spend at least the mean response time to complete the 'mark all that apply' question, to select as many items as those who complete a 'yes/no' series.

A5: We expect the observed response differences caused by format to be greater for the difficult than the easy question.

Our specific hypotheses for the comparison between branching and non-branching are:

B1: Based on the findings of Nicolaas et al 2000 and Yu et al 2003, when branching questions in CATI but not in CAPI and CAWI,

(B1a) We expect more extreme responses in CATI compared to CAPI and CAWI,

(B1b) We expect this effect to be more prevalent (in the expected direction) for attitudinal than factual questions.

B2: Within each mode,

(B2a) We expect more extreme responses when branching is used compared to no branching,

(B2b) We expect this effect to be more prevalent (in the expected direction) for attitudinal than factual questions.

B3: We expect no difference in extreme responses between modes when branching is used across all modes.

3. Methodology

3.1 The experimental data

The collection of experimental data took place in two surveys: the NatCen Omnibus survey and the British Household Panel Study (which has become part of the UK Household Longitudinal Survey). The NatCen Omnibus survey is a probability sample of adults aged 16 and over in Great Britain whereby clients are able to buy questionnaire space. The survey is administered quarterly to a fresh sample of respondents and 1,600 interviews are administered face-to-face using CAPI. The British Household Panel Study (BHPS) is based on an original probability sample of 5,000 households in Great Britain in 1991 and is also interviewed using CAPI. Individuals from these BHPS chosen households have continued to be followed annually ever since.

Prior to the mixed modes experiment, 15 questions from the BHPS were selected to address some of the wider project's hypotheses. These were administered in the NatCen Omnibus survey and as part of the main BHPS survey. Six months later, all NatCen Omnibus survey respondents who agreed to be recontacted were randomly allocated to one of three modes (CAPI, CATI, and CAWI) for the mixed modes experiment. For the BHPS, a sub-sample of respondents was selected, ensuring just one respondent per household to match the NatCen Omnibus design, and randomly allocated to CATI or CAWI follow-up samples.[3] The remainder of the BHPS sample would later be interviewed by CAPI as part of the standard BHPS survey. At the time of writing this paper all modes of NatCen Omnibus data collection had been completed, but only the CATI and CAWI components of the BHPS were available.[4]

[3] The CAWI sample for both studies was restricted to respondents who had access to and used the internet. Although this restriction did not hold for CAPI and CATI respondents. For analyses comparing CAWI with other modes, only respondents who had access to and used the internet were included in comparisons.

[4] The CAPI data will be from what would have been Wave 19 of the BHPS, but is now part of Wave 2 of the UK Household Longitudinal Survey.

The mixed modes questionnaire repeated the original module of 15 questions and included an additional 67 questions designed to test a set of hypotheses about the causes and consequences of mode effects, including those set out in Section 2 above. These 67 additional questions were classified according to type of question (satisfaction, other attitudinal, behavioural, other factual), task difficulty and sensitivity of the question. In addition, seven different question format comparisons were experimentally contrasted: (1) short versus long scales, (2) rating versus ranking, (3) agree/disagree statements versus balanced questions addressing more than one side of the issue, (4) 'yes/no for each' versus 'mark all that apply', (5) branching versus non-branching, (6) fully-labelled versus end-labelled scales and (7) show card versus no show card on long lists in CAPI. This paper uses the question comparisons (4) 'yes/no for each' versus 'mark all that apply' and (5) branching versus non-branching.

The response rates for the mixed mode experiment are listed in Table 1. A key concern was the possibility of differential nonresponse bias which would confound the substantive question comparisons between modes, in particular since the response rates for CAWI were so much lower than for CAPI and CATI. After considering several adjustment options including standard weighting, propensity score weights, and modelling with an optimal set of control variables, we opted for modelling and the final set of control variables comprised sex, age, ethnicity, marital status and economic activity status.

Table 1: Mixed mode experiment response rates

	NatCen Omnibus	BHPS
CAPI	73%	Not available
CATI	69%	70%
CAWI	47%	37%

Table 2: Mixed mode experiment sample sizes after exclusion of non-internet access or use cases from CAPI and CATI samples

	NatCen Omnibus	BHPS
CAPI	282	Not available
eCATI	314	421
CAWI	349	334
TOTAL	945	755

3.2 Analysis methods for the comparison of 'yes/no for each' versus 'mark all that apply' formats

The questionnaire contained two split ballot experiments contrasting 'yes/no for each' versus 'mark all that apply' (see Figure 1). One was based on 8 different suggestions to reduce poverty. The other was based on 8 attributes you could like about your neighbourhood. The poverty questions, as opposed to the neighbourhood questions, were considered the difficult series as we hypothesised most people would not have pre-formed attitudes about this topic.

We first examined cross-tabulations for each 'yes/no' question versus its respective 'mark all that apply' question overall and by the 3 modes of data collection. We then summed up the total number of endorsements within each of the two question series for use as the dependent variables in ordinary least squares regression. The control variables described above were included in all models. We complemented the analyses by examining data on respondent completion times for the two question series to compare completion times between modes and formats and to classify respondents are faster or slower.

Figure 1: The 'yes/no for each' versus 'mark all that apply' questions presented in 'yes/no' format for CAPI

The 'poverty' questions *	The 'neighbourhood' questions ◊
GB21. 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? Would increasing pensions reduce poverty? Yes 1 No 2	**N56.** What are the things that you like about your neighbourhood? Do you like your neighbourhood because of its community spirit? Yes 1 No 2
GB22. Would investing in education for children reduce poverty? Yes 1 No 2	**N57.** Do you like your neighbourhood because it feels safe? Yes 1 No 2
GB23. Would improving access to childcare reduce poverty? Yes 1 No 2	**N58.** Do you like your neighbourhood because of the neighbours? Yes 1 No 2
GB24. Would the redistribution of wealth reduce poverty? Yes 1 No 2	**N59.** Do you like your neighbourhood because of the character of its buildings? Yes 1 No 2
GB25. Would increasing trade union rights reduce poverty? Yes 1 No 2	**N60.** Do you like your neighbourhood because of Its cleanliness? Yes 1 No 2
GB26. Would reducing discrimination reduce poverty? Yes 1 No 2	**N61.** Do you like your neighbourhood because of Its location? Yes 1 No 2
GB27. Would increasing income support reduce poverty? Yes 1 No 2	**N62.** Do you like your neighbourhood because it is quiet? Yes 1 No 2
GB28. Would investing in job creation reduce poverty? Yes 1 No 2	**N63.** Do you like your neighbourhood because of its transport facilities? Yes 1 No 2
* Taken from the Poverty and Social Exclusion Survey of Britain, 1999, with the addition of an item on increasing income support and one on investing in job creation.	◊ Adapted from a London Housing Association questionnaire [5]

[5] Although labelled as easy, it is clear to a survey researcher that the neighbourhood questions are problematic. This is because each question assumes that respondents like their neighbourhood; that the given characteristic applies to their neighbourhood (i.e., the facilities exist); and if it applies, that it is something they like about their neighbourhood. The cognitive interviewing explored these issues. From cognitive respondents' think alouds and answers to probes, it was clear that they had no problems with the questions. Their views suggested that the questions were easy, straightforward and about, as one respondent described them, "*everyday stuff*".

3.3 Analysis methods for the comparison of branching versus non-branching formats

The questionnaires contained four split ballot experiments contrasting branching versus non-branching for attitudinal and factual questions, further classified as easy or difficult (see Figure 2)[6]. The attitudinal questions rated local shopping facilities and perceived change in standard of living. 'Rating of local shopping facilities' was classified as the easier question because it was expected that respondents would be more likely to have a well-formed attitude about this than 'changes in standard of living'. The two factual questions were the respondent's household's 'monthly rent or mortgage' and 'monthly grocery shopping costs'. Of these, the 'grocery shopping expenditure' questions were considered the more difficult series. This was because the immediately preceding question defines grocery shopping as including food, drinks, cleaning products, toiletries and household goods, thus increasing the difficulty of the task to come up with an overall estimate.

We first examined cross-tabulations for each branching question versus its respective non-branching question overall and by the 3 modes of data collection. We then compared the proportion of extreme and non-extreme answers, using two versions of the dependent variables: one indicating whether the respondent had selected either the highest or lowest categories versus the other categories (referred to as 'one high/low') and the other indicating whether the respondent had selected one of the two highest or two lowest categories versus the other categories (referred to as 'two high/low'). To test for differences between formats and modes we used logistic regression, adding the control variables for nonresponse.

[6] Note that only the 'change in standard of living' question uses a middle category.

Figure 2: Branching versus non-branching, presented in branched format for CAPI

The 'rating shopping facilities' questions ◊	The 'change in standard of living' questions *
N40. Please tell me whether you consider your local shopping facilities to be poor or good? Poor 1 GO TO N41 Good 2 GO TO N42 **N41.** Would this be poor, very poor or extremely poor? Poor 1 GO TO N43 Very poor 2 GO TO N43 Extremely poor 3 GO TO N43 **N42.** Would this be good, very good or extremely good? Good 1 Very good 2 Extremely good. 3 ◊ Newly developed to test hypotheses	**GB18.** Thinking back to the last general election, would you say that the standard of living has increased or decreased, or has it stayed the same? Increased 1 GO TO GB19 Decreased 2 GO TO GB20 Stayed the same 3 GO TO GB21 **GB19.** Would you say it has increased by a small amount, a medium amount or a large amount? Increased by a small amount 1 GO TO GB21 Increased by a medium amount 2 GO TO GB21 Increased by a large amount 3 GO TO GB21 **GB20.** Would you say it has decreased by a small amount, a medium amount or a large amount? Decreased by a small amount 1 Decreased by a medium amount 2 Decreased by a large amount 3 * Shortened from Welsh Assembly Election Study, 1999
The 'rent or mortgage expenditure' questions *	**The 'grocery shopping expenditure' questions ◊**
FM76. How much did your household spend last month in rent or mortgage for the accommodation you live in? Was it more or less than £300? Less than £300 1 → GO TO FM77 More than £300 2 → GO TO FM79 **FM77**. Was it more or less than £200? Less than £200 1 → GO TO FM78 More than £200 2 → GO TO FM81 **FM78.** Was it more or less than £100? Less than £100 1 → GO TO FM81 More than £100 2 → GO TO FM81 **FM79.** Was it more or less than £400? Less than £400 1 → GO TO FM81 More than £400 2 → GO TO FM80 **FM80.** Was it more or less than £500? Less than £500 1 More than £500 2 * Newly developed to test hypotheses	**FM69.** How much did your household spend last month on grocery shopping? Was it more or less than £300? Less than £300 1 → GO TO FM70 More than £300 2 → GO TO FM72 **FM70**. Was it more or less than £200? Less than £200 1 → GO TO FM71 More than £200 2 → GO TO FM74 **FM71.** Was it more or less than £100? Less than £100 1 → GO TO FM74 More than £100 2 → GO TO FM74 **FM72.** Was it more or less than £400? Less than £400 1 → GO TO FM74 More than £400 2 → GO TO FM73 **FM73.** Was it more or less than £500? Less than £500 1 More than £500 2 ◊ Newly developed to test hypotheses

3.4 The cognitive interviewing methodology

In the survey context, cognitive interviewing is traditionally used as a pretesting method (Presser et al, 2004). In contrast, we pre-planned a cognitive interviewing follow-up study, designed to gain a greater understanding of how mode effects happen-even if they were not directly observed-and to seek explanations for any unusual quantitative findings.

Thirty seven respondents were recruited for the cognitive interviewing phase. These respondents had participated in the NatCen Omnibus mixed modes experiment and were selected using specific quotas, contrasting respondents who had displayed satisficing behaviour connected with mode effects[7] versus those who had not.

The cognitive interviews began with a carefully selected subset of survey questions from the mixed modes experiment. These questions were administered in standard quantitative fashion and across 3 modes (CAPI, CATI, and CAWI). This involved the interviewer sitting with the respondent face-to-face (for the CAPI component), being in a different room in the respondent's home and talking over a landline/mobile phone (for the CATI component) and having the respondent use the interviewer's laptop completely on his/her own (for the CAWI component). Although being exposed to all 3 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 one mode and the rest in a different mode.[8] 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.

After the administration of all survey questions, there was the transition to the actual cognitive interviewing. Here the cognitive interviewer made use of retrospective think alouds and many pre-written probes. For most questions, this was done by reminding the respondent of the survey question, data collection mode, his or her answer and any behaviour displayed whilst answering e.g., hesitation. The respondent then talked through how he or she had gone about answering the question and how he or she had decided on the answer. Then, where appropriate, the interviewer

[7] Although mode differences are typically detected at an aggregate level, we found that certain respondent 'satisficing' behaviours differed by mode (i.e., acquiescence through agreeing to opposite agree/disagree statements and non-differentiation in a ranking task)

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

asked the respondent a number of structured open probes, such as, *"Tell me your thinking behind choosing that category"*, to explore further anything about the response process that had not been covered by the think aloud account.

All cognitive interviews were transcribed and the data introduced into the qualitative charting programme, "Framework", for analysis. 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 (2011).

4. Results

4.1 'Yes/no for each' versus 'mark all that apply'

Using a two sample t-test for each pair of items, it can be seen that the 'yes/no' format produced significantly higher endorsements compared to the 'mark all that apply' format, for each of the 16 items from both the poverty and the neighbourhood questions (see Table 3). Using the Bonferroni method to adjust for multiple tests (alpha/16=.003), all results remain significant.[9] Although the easy neighbourhood questions showed more endorsements in both formats, the range of differences (between question formats for each item) were similar to those for the more difficult poverty questions, with both varying from 14.2 to 36.9 percentage points. It is also important to note that the results in Table 3 show no trace of primacy effects as large differences between items were found for the first few, middle few, and last few items.

[9] Format differences within CAPI and CAWI modes were also explored. (Note that comparisons were not made within CATI as 'mark all that apply' questions cannot be asked in that format.) Within CAPI, all are significant and in the expected direction at $p < .05$. Nine remain significant after the Bonferroni adjustment. Within CAWI, all are significant and in the expected direction at $p < .05$. Ten remain significant after the Bonferroni adjustment.

Table 3: Percent of CAPI and CAWI NatCen Omnibus respondents endorsing 'yes/no' versus 'mark all that apply' ♦

Response format	Yes/No for each % Endorsed	Mark all that apply % Endorsed	Difference between percentages T-test for each pair shows (p<.001 for each pair)
Poverty questions	(*n* varies from 452 to 474 due to missing data.)	(*n* = 313)	
Increasing pensions	71.3	52.4	18.9
Education for children	69.8	55.6	14.2
Improving access to childcare	67.3	33.2	34.1
Redistribution of wealth	66.8	33.2	33.6
Increasing trade union rights	26.3	5.1	21.2
Reducing discrimination	46.9	18.8	28.1
Increasing income support	46.1	11.5	34.6
Investing in job creation	92.8	77.6	15.2
Average	**60.9**	**35.9**	**25.0**
Neighbourhood questions	(*n* varies from 458 to 466 due to missing data.)	(*n* = 318)	
Community spirit	56.1	30.8	25.3
Feels safe	84.7	62.3	22.4
Neighbours	75.3	60.7	14.6
Character of buildings	54.3	27.4	26.9
Cleanliness	75.6	38.7	36.9
Location	93.1	78.3	14.8
Quiet	82.5	61.0	21.5
Transport facilities	57.9	34.9	23.0
Average	**72.4**	**49.3**	**23.2**

♦ Analyses restricted to respondents with internet access. These are the raw percentages unadjusted by the regression control variables.

A1: Higher endorsement in CATI 'yes/no' than CAPI/CAWI 'mark all that apply'

Table 4 shows that, as expected, the mean number of endorsements with the 'yes/no' format in CATI was higher than with the 'mark all that apply' format in CAPI and CAWI. As described in Section 3.2, significance was tested through OLS regression models with controls for nonresponse. The regressions show that Hypothesis A1 is clearly supported. The regression

coefficients (*b*) and standardised regression coefficients (β)[10] for question format for the poverty questions were *b*=2.106, β=.525 (p<.001) for CATI > CAPI and *b*=2.064, β=.529 (p<.001) for CATI > CAWI and for the neighbourhood questions were *b* =-1.977, β=.441 (p<.001) for CATI > CAPI and *b* =2.318, β=.556 (p<.001) for CATI > CAWI. This result was replicated in the CATI/CAWI comparison from the BHPS data *b* =3.414, β=.493 (p<.001) for the poverty questions and b=2.343, β=.299 (p<.001) for the neighbourhood questions.

A2: No differences in endorsement between CAPI and CAWI using 'mark all that apply'

Table 4 shows that the mean number of endorsements in CAPI and CAWI were similar when 'mark all that apply' formats were used for the poverty question and that the mean for CAPI was slightly larger than for CAWI for the neighbourhood questions. But, the OLS regression analyses with controls for the NatCen Omnibus data[11] showed that there were no mode differences between CAPI and CAWI for either question series, thus supporting Hypothesis A2.

A3: No differences in endorsement between modes when 'yes/no' format is used

In contrast to the other hypotheses, Table 4 suggests that Hypothesis A3 is not supported. There were differences between CATI, CAPI and CAWI when the 'yes/no' format is used in all modes. This is confirmed by the OLS regressions with controls. For the poverty questions CAPI respondents endorsed more items than CAWI respondents (*b* =.768, β=.192, p<.001) and CATI respondents endorsed more items than CAWI respondents (*b* =.517, β=.137, p<.05). CAPI and CATI respondents did not differ significantly from each other. For the neighbourhood questions, CATI respondents endorsed more items than CAWI respondents (*b* =.483, β=.125, p<.05) and contrary to the poverty questions, CATI respondents almost endorsed significantly more items than CAPI respondents (*b*=.376, β=.097, p=.075). CAPI and CAWI respondents were not significantly different. For the BHPS poverty questions, CATI respondents had a higher mean than CAWI respondents, but this didn't reach significance (*b*=.351, β=.091, p=.111). For the neighbourhood questions, CATI respondents had a significantly higher mean than CAWI respondents (*b*=.876, β=.256, p<.001).

[10] The standardised coefficients are reported to give the reader an idea of the magnitude of the relationship between the characteristic (format and/or mode) and the dependent variable. Standardised coefficients in this case can be treated like partial correlation coefficients.

[11] This could not be tested in the BHPS data as the CAPI results were not available.

Table 4: Unadjusted mean number of items endorsed by question set and mode for NatCen Omnibus respondents♦

Response format	Yes/no for each		Mark all that apply	
	Mean	n	Mean	n
Poverty questions				
CAPI	5.13	110	2.84	147
CATI	5.02	113	NA	NA
CAWI	4.41	178	2.90	166
Neighbourhood questions				
CAPI	5.74	141	4.09	135
CATI	6.01	142	NA	NA
CAWI	5.60	166	3.94	183

Data on completion times[12]

A4a: Longer completion times with 'yes/no' than 'mark all that apply'

Table 5 shows that question completion times. Much longer time was taken with the 'yes/no' format than 'mark all that apply'. Using factorial ANOVA models with the nonresponse control variables confirms that the differences are significant ($p < .001$ in all 4 comparisons: two modes by two question series), thus supporting Hypothesis A4a. In addition, there is also an interaction with mode. Although CAPI and CAWI respondents spent a roughly similar amount of time on the 'yes/no' format, CAPI respondents took longer on the 'mark all that apply' format than did CAWI respondents ($p < .001$ for both question series). This suggests more thorough answers in the 'yes/no' format and more thorough answers in CAPI than CAWI when the 'mark all that apply' format is used. Similarly, the variability between respondents, as indicated by the relative standard deviations in Table 5, is roughly similar for CAPI and CAWI with the 'yes/no' format. But with the 'mark all that apply' format, CAWI respondents were more variable in the amount of time they took to answer questions than CAPI respondents. This could indicate that CAWI respondents vary more in their motivation to fully engage with the question (some evidence of this from the cognitive interviewing phase).

[12] The time completion data were considered both with and without the exclusion of outliers defined as cases greater than 2 standard deviations from the mean. The conclusions were the same, but the results in this section exclude outliers. Also excluded, just for the time completion data analyses, are cases where respondents did not endorse any items. These cases were excluded from both formats.

Table 5: Question completion time by question set, mode and format

Questions	Mode	Format	Unadjusted mean completion time in seconds♦	Std Dev	Relative Std Dev: $\left\lvert \frac{std\ dev}{mean} \right\rvert$	n
Poverty	CAPI	Yes/No	82.19	27.00	3.29	129
	CAWI	Yes/No	84.56	28.40	3.36	163
	CAPI	Mark all	52.02	27.78	5.34	143
	CAWI	Mark all	18.78	12.74	6.78	157
Neighbourhood	CAPI	Yes/No	53.62	15.94	2.97	138
	CAWI	Yes/No	57.87	18.05	3.12	153
	CAPI	Mark all	37.63	15.69	4.17	130
	CAWI	Mark all	14.88	9.75	6.55	183

♦ Table contains unadjusted mean completion times, but statistical comparisons were made after control variables were applied. Outliers greater than 2 standard deviations removed.

A4b: Respondents answering 'mark all that apply' questions in less than mean response time show evidence of primacy effects

We examined the relationship between respondents who completed the 'mark all that apply' format in less than the mean response time and the tendency to choose items from the top of the list. For the poverty questions, neither cross-tabulations nor logistic regressions with controls found any evidence for primacy effects[13] being associated with faster responding times. For the neighbourhood questions, cross-tabulations suggested that both primacy and recency choices were associated with faster responding. But in the logistic regressions only the recency effect approaches significance ($p = .054$). It is unclear what this possible effect means, but the key point is that there is no support for Hypothesis A4b.

A4c: Respondents who answer 'mark all that apply' questions in at least mean response time select as many items as those completing the 'yes/no' format

Table 6 shows that Hypothesis A4c was not supported. Combining the data from all three modes, the mean number of items endorsed was still higher in the 'yes/no' format than in the 'mark all that apply' format for respondents who took an average amount of time or longer on the 'mark all that apply' questions. This held true for both the poverty and neighbourhood questions ($p < .001$ in both cases).

[13] For the time completion data, these are defined as covering the first 4 items.

Table 6: Mean number of items endorsed by question series and format

Question Series	Format	Analysis Base	Unadjusted mean number of items endorsed	Std Dev	n
Poverty	Yes/No	All respondents	4.68	.096	288
	Mark all	Just respondents expending more than the mean amount of time	3.09	.159	105
Neighbourhood	Yes/No	All respondents	5.66	.105	307
	Mark all	Just respondents expending more than the mean amount of time	4.20	.170	117

A5: Effects are greater for difficult than easy questions

Drawing on evidence discussed so far in Section 4, we can conclude that Hypothesis A5 was not supported. This can be clearly seen in Table 3. Although the easy neighbourhood questions showed a higher average percentage of endorsements in both formats (72.4% for 'yes/no'; 49.3% for 'mark all that apply') than the poverty questions (60.9% for 'yes/no'; 35.9% for 'mark all that apply'), the average of differences between question formats were very similar 23.2% and 25.0%, respectively. The range of differences was also analogous with 14.2 to 34.6 for the poverty questions and 14.6 to 36.9 for the neighbourhood questions. A comparable pattern can be seen in Table 6 among the subset of 'mark all that apply' respondents who had spent the average amount of time or more on their task compared to 'yes/no' respondents. Although respondents are likely to endorse more neighbourhood items than poverty items, the difference in the number selected between the two formats is not that different (i.e., 4.68 – 3.09 = 1.59 and 5.66 – 4.20 = 1.46). This evidence suggests that the way respondents attend to the two formats is similar regardless of whether the questions are easy or difficult.

4.1.2 Cognitive interview findings

The focus of the cognitive interviewing project came from the quantitative results. It was a surprise that Hypothesis A3 was not supported, i.e., that the 'yes/no' format was not comparable across modes. So this particular hypothesis was investigated using the poverty questions, comparing just CAPI respondents to CAWI respondents. The cognitive interviewing results suggest that there were many subtleties that could have affected aggregate mode comparisons in different directions.

Firstly, there were instances of possible and clear satisficing; more in CAWI than CAPI and almost all of these were in the 'yes' category. This could indicate that in the absence of a middle category that 'yes' is an easy answer. One respondent specifically said she "*erred on the side of 'yes'*" (Female, 30 to 39, postgraduate degree, employed, high income, White British).

Secondly, if respondents were in the middle ground (e.g., qualified their answer or said it depends), they were much more likely to choose 'yes' than 'no'. Of these respondents in the middle ground, more were in CAPI than were in CAWI. It is also interesting that the two respondents who changed their answers during the cognitive interview debriefing both changed their answers from 'no' to 'yes'.

Thirdly, there is usually a difference between modes with respect to giving a socially undesirable answer (i.e., more likely in CAWI than CAPI). According to respondent comments, two of the six questions[14] were unexpectedly a bit sensitive: GB27 (increasing income support) and GB24 (redistribution of wealth). In both cases there were slightly more 'no' answers, the socially undesirable answer, in CAWI than in CAPI. These are the only two questions which show this pattern. On GB27 (increasing income support), one CAPI respondent commented, "*it's a hard one to say 'no' . . . what's somebody going to think me saying no*" (Female, 40 to 49, high school level equivalent, employed, low income, White British). On GB24 (redistribution of wealth) a CAWI respondent 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*" (Female, 30 to 39, first degree, employed, high income, White British).

So overall, these findings raise questions about the validity of the 'yes' answers in the 'yes/no' format as they are more prone than the 'no' answers to contain satisficing answers, and clarified and dependent answers. Moreover, this could affect mode comparisons. Although caution is warranted given this small unrepresentative sample, findings suggest that 'yes' answers due to satisficing may be more likely to occur in CAWI (a pattern repeated across analyses on other topics in the mixed modes experiment) while 'yes' answers due to 'clarified' and 'dependent' answers may be more likely to occur in CAPI. In addition, a 'yes' answer due to giving a socially desirable answer may be more likely to occur in CAPI.

[14] Only 6 of the 8 questions were included in the cognitive interviews (3 which showed the most mode differences and 3 which showed the least mode differences.

Interestingly the latter two findings are in line with the quantitative mode results for the 'yes/no' questions, whereas the first more general finding about satisficing is at odds with the quantitative mode results. The effect of these different types of respondent behaviour on the quantitative data would depend on their prevalence.

4.2 Branching versus non-branching

B1a: More extreme answers with branching in CATI than non-branching in CAPI/CAWI

B1b: Stronger effect (in the expected direction) for attitudinal than factual questions

Using the 'two high/low' dependent variable, which combines the top two and bottom two categories as an indicator of extremeness, we estimated a logistic regression model of the probability that the respondent selected an extreme response using the question format indicator and nonresponse controls as explanatory variables. The results from the NatCen Ominbus data show the expected pattern for the 'change in standard of living' question, with CATI responses significantly more extreme than CAWI responses (Odd Ratio (OR)=1.848, $p < .01$) and almost significantly more extreme than CAPI responses (OR=1.592, $p = .053$). As can be seen from the actual percentages in Table 7, the expected pattern is also present in the 'rating of shopping facilities' question, but did not reach significance in the logistic regression. The factual questions paint a different picture. For the 'rent or mortgage expenditure' questions, the percentages do not appear to differ in Table 7 and this is confirmed by the logistic regression. For the 'grocery shopping expenditure' questions, Table 7 shows a lower rather than higher percentage of extreme answers with branching in CATI. The logistic regression confirm that CATI branched responses are significantly less extreme than CAWI non-branched responses (OR=.601, $p < .05$).[15]

The attitude questions appear to support Hypothesis B1a with a clear and significant case of more extremeness with branching in CATI for the 'change in standard of living' questions and a clear, but not significant, case for the 'rating of shopping facilities' questions. The results for the factual questions were either non-significant or in the opposite direction, implying support for Hypothesis B1b.

[15] The BHPS data show mostly similar, but non-significant, results. CATI responses are slightly more extreme than CAWI responses on the two attitudinal variables and also on the 'rent or mortgage expenditure' questions, but only reach significance on the 'two high/low' version of 'rating of shopping facilities', OR=1.873, $p<.01$. Although not significant, the opposite pattern is seen for the 'grocery shopping expenditure' questions as was seen for the NatCen Omnibus data.

But why is the hypothesised effect found clearly on the 'change in standard of living' questions and much less so on the 'rating of shopping facilities' questions? [16] There are three confounding factors. The former used a showcard and the latter did not, the former had a middle category and the latter did not, and finally the two questions are on very different topics with different extremeness in wording. First with respect to the showcard / no showcard difference in implementation, there is some evidence to suggest that the branching effect is robust to this. On the 'change in standard of living' questions, both within CAPI with a showcard and CATI without a showcard there are significant branching effects. Second, with respect to having or not having a middle category, there is evidence to suggest this is not a large issue. For both questions large percentages of respondents are choosing the middle on the 'change of standard of living' (26 percent on average across formats and modes) and a central category on the 'rating of shopping facilities' ('good' as opposed to very good or extremely good with 44 percent on average across formats and modes). Although the percentage is higher for the 'rating of shopping facilities question' than the 'change in standard of living' questions, we will see below that it is what happens in the extremities which is probably causing the differences. Third, with respect to topic and extremity of wording, respondents were more likely to say that their standard of living had 'decreased' rather than 'increased' and that their shopping facilities were 'good' rather than 'poor'. But there were also important differences. On the 'change in standard of living' questions (for all three modes), respondents who had received the branching format were more likely to choose the last two categories ('decreased by a medium amount' and 'decreased by a large amount') compared to those who received the non-branching format. For the 'rating of shopping facilities' (for CAPI and CAWI), respondents who had received the branching format were more likely to choose the 5th rather than the last category (i.e., very good rather than extremely good) compared to those who received the non-branching format. But it is unclear whether this is due to respondents avoiding the extreme labels in the 'rating of shopping facilities' questions (i.e., response contraction bias – see Tourangeau et al, 2000) or that respondents felt more strongly about the topic of 'change in the standard of living' questions.

[16] Interestingly, only the 'change in standard of living' question comes from the Welsh Assembly Election Study where extremeness with branching had been detected before.

Table 7: Unadjusted percent of respondents choosing the 'two high/low' pattern by format and mode

Dependent variable	Format	CAPI	CATI	CAWI	Total n
Change in standard of living	Branching	57.8	54.7	65.4	476
	Non-branching	41.8	37.4	39.2	459
Rating of shopping facilities	Branching	44.8	37.3	33.9	463
	Non-branching	34.8	36.0	31.7	479
Grocery shopping expenditure	Branching	43.8	43.2	48.8	458
	Non-branching	48.9	53.5	52.2	476
Rent or mortgage expenditure	Branching	83.1	81.4	74.0	463
	Non-branching	80.1	77.1	80.4	448

B2a: Within mode, more extreme responses with branching than non-branching

B2b: Stronger effects (in the expected direction) for attitudinal than factual questions

Table 8 shows that the extent of extreme reporting within modes was indeed greater with the branched than non-branched questions for the attitude items. Within each mode, the 'change in standard of living' question shows significantly more extreme responses with the branched format. For the 'rating of shopping facilities' significance is only approached within CAPI ($p = .073$).

Again a different pattern emerges for the factual questions. Using the 'two high/low' dependent variable, the 'grocery shopping expenditure' questions only shows a significant difference within CATI and the 'rent or mortgage expenditure' questions almost show a significant difference within CAWI ($p = .063$). The latter case, is however significant with the 'one high/low' dependent variable). Similarly for the 'grocery shopping expenditure' questions and using the 'one high/low' dependent variable, significant differences are found within CAPI. Most

importantly, however, is that the effects for the factual questions are in the opposite direction of the expected pattern, with higher extreme percentages in the non-branched format.[17]

Table 8: Branching versus non-branching within mode, significance of logistic regression coefficients

Dependent variable	Extremeness measured with	CAPI	CATI	CAWI
Change in standard of living	Two high / low	Branching more extreme, OR=2.072, p<.01	Branching more extreme, OR=1.867, p<.05	Branching more extreme, OR=2.903, p<.001
Rating of shopping facilities	Two high / low	Branching almost more extreme, OR=1.579, p=.073	-	-
Grocery shopping expenditure	One high / low ♦	Non-branching more extreme, OR=5.096, p<.01	-	-
Grocery shopping expenditure	Two high / low	-	Non-branching more extreme, OR= 1.652, p<.05 ◊	-
Rent or mortgage expenditure	One high / low ♦	-	-	Non-branching more extreme, OR=1.732, p<.05
Rent or mortgage expenditure	Two high / low	-	-	Non-branching more extreme, OR=1.679, p=.063

♦ 'One high/low' results only reported if results were significant.

◊ This finding is due to a large value in the second category for CATI.

- No significant differences found.

How seriously should we take the findings for the factual questions? There is more inconsistency. On the 'rent or mortgage expenditure' questions for NatCen Omnibus CAPI and CATI respondents, the figures for branching are higher than for non-branching, but the differences are not significant. It is only within CAWI, that differences in the opposite direction were found. On the 'grocery shopping' expenditure questions, NatCen Omnibus CAPI and CATI respondents and BHPS CAWI respondents all show the pattern of non-branching being more extreme. But these are not large differences. This can be illustrated with the BHPS CATI findings for the 'grocery shopping expenditure' questions. There are only slightly more cases in the first category (10.6 percent non-branching compared to 5.9 percent branching), second category (23.9 percent non-

[17] With respect to the BHPS data, there is a good replication of the NatCen Omnibus findings. Significantly more extreme answers are found with branching on the 'change in living standards' question within CATI when using the 'two high/low' dependent variable ($p < .05$) and this almost holds within CAWI ($p = .087$). It does hold within CAWI when the 'one high/low' dependent variables is used ($p < .05$). For the 'rating of shopping facilities' questions, more extreme answers are found only found within CAWI and for the 'two high/low' dependent variable ($p < .01$). On the factual questions, 'grocery shopping expenditure' is significantly more extreme within CAWI for the non-branching format as was found with the NatCen Omnibus data. This was true when using the 'two high/low' variable ($p < .05$) and almost true when using the 'one high/low' variable ($p = .094$). There were no significant differences on the 'rent or mortgage expenditure' questions.

branching compared to 21.1 percent branching), second to last category (15.6 percent non-branching compared to 11.8 percent branching) and none in the last category (3.9 percent non-branching compared to 3.9 percent branching). But when combined these produced a significant difference. Thus although there are several examples of more extreme answers in the non-branched format for the factual questions, the circumstances under which these occur would suggest that these findings should be viewed with some caution. Other replications on factual data are needed.

So overall, the attitude questions lend support to Hypothesis B2a and suggest support for Hypothesis B2b.

B3: No differences in extreme responses when branching used in all modes

Surprisingly this hypothesis was not supported. When branching was used, there were many differences in extreme responses between modes. There were however no clear patterns (see Table 9). In the BHPS data there is only one significant mode difference, but that is not too surprising given that 60 percent of the NatCen Omnibus results involve CAPI and the BHPS CAPI data were not yet available.

Table 9: Looking at branching across modes, significance of logistic regression coefficients

Dependent variable	Extremeness measured with	NatCen Omnibus results	BHPS CATI / CAWI results
Change in standard of living	Two high / low	CAWI > CATI, OR=1.666, p<.05	-
Rating of shopping facilities	Two high / low	CAWI < CAPI, OR=.575, p<.05	-
Grocery shopping expenditure	One high / low ♦	CAWI > CAPI, OR=4.080, p<.01	-
	Two high / low	-	-
Rent or mortgage expenditure	One high / low ♦	-	CAWI < CATI, OR=.621, p<.05
	Two high / low	CAWI < CAPI, OR=.524, p<.05 CAWI < CATI, OR=.501, p<.05	-

♦ 'One high/low' results only reported if results were significant.

- No significant differences found.

4.2.2 Cognitive interview findings

Branching effects are typically discovered at the aggregate level through an experimental comparison. For the cognitive interviewing, we could not rely on this technique. We needed to identify individuals prone to branching effects in order to be able to talk to them and understand their experiences. For 12 respondents, we decided to ask the question in branching and non-

branching formats in the same interview using both the 'change in standard of living' questions and the 'rating of shopping facilities' questions. This was done as follows. The question was first asked in branched format as part of the survey questions. Later during the actual cognitive interviews the interviewer presented the respondent with a showcard in 'non-branched' format (without reminding the respondent of his/her answer to the survey question). If the respondent chose a different answer with the showcard, the interviewer was then instructed to probe, "*I noticed that you have come up with a slightly different answer, can you tell me what you were thinking?*", reminding the respondent of his/her answer to the branched version, if needed.

A few of the 12 respondents were inconsistent between the two formats and one of these respondents was inconsistent on both questions. The reasons for inconsistency were due both to the vagueness of the answer categories and confusion about what to include or exclude from the question.

For example, with respect to 'rating of shopping facilities, one respondent (Male, 60 or older, high school equivalent, employed, high income, White British) answered 'very good' on the branched survey questions and later during the cognitive interviewing, chose 'extremely good' off the showcard. When asked about the discrepancy, he first said, *"To me, extremely good and very good are the same thing"*. Later on in the 'change in standard of living' questions he chose 'decreased by a large amount' on the branched survey questions and 'decreased by a medium amount' on the showcard. When asked about this he answered: *"Strange, I don't really know, to tell you the truth, to be honest."*" And continued, *"Just medium, I don't know how you'd describe a medium amount, you know what I mean? What's a large amount?"*

With respect to the 'rating of shopping facilities' questions, one respondent's answers were actually contradictory (choice of good through branching and choice of poor on the showcard). The respondent justified this by saying that the local shops in the village were poor but there was a Morrisons 5 miles away (Female, 50-59, high school equivalent, employed, low income, White British). The respondent also commented that the questions were confusing and obtuse. On the same questions, a different respondent chose 'very good' in the branched version and 'good' on the showcard non-branched version. The respondent indicated confusion about whether 'local shopping facilities' are for food shopping or clothes shopping (Female, 20 to 29, employed, low income, White British).

Although not explaining the branching effect, these findings illustrate the instability of respondents' answers.

With the remaining 25 respondents the cognitive interviewing explored the possibility for different processing of the branching questions in the different modes. The 'change in standard of living' questions, a new equivalent 'change in crime' questions, the 'rent or mortgage expenditure' questions and a new equivalent 'electricity expenditure' questions were asked in different modes. Although many things were found about how respondents understood and answered the questions[18] (see Campanelli et al, 2011), there was nothing to indicate any systematic differences by mode.

5 Discussion

In this paper we have explored the impact of adapting two question formats that are commonly used in face-to-face interviews and self-completion questionnaires so that they can be administered more easily in a telephone interview:

1. 'mark all that apply' versus a series of 'yes/no' questions for each item;
2. presentation of all response options in one step versus branching the question into two or more steps.

Looking first at 'mark all that apply' versus a series of 'yes/no' questions, we found that the number of response options selected was higher in the 'yes/no' format compared to the 'mark all that apply' format which is consistent with previous research. This effect was apparent in each of the modes using the 'mark all that apply' format, i.e., CAPI and CAWI. We also found that respondents took more time to answer the 'yes/no' questions compared to the 'mark all that apply' questions, which Smyth et al (2006) suggested could indicate deeper processing of the question in the 'yes/no' format. However contrary to the findings of Smyth et al (2006), we did not find that those who took longer than average to complete the 'mark all that apply' questions selected the same number of items as those who completed the 'yes/no' questions, which is what we were expecting if time to complete was associated with deeper processing. Furthermore, we found no clear evidence of a primacy effect among those who completed the 'mark all that apply' format in

[18] Both the 'rent and mortgage expenditure' questions and the 'electricity expenditure' questions suffer from a basic flaw; there is no category for respondents who fall on the cusp. For example, the 'rent or mortgage expenditure' questions are asked in terms of 'more or less than £300', more or less than £200', and so on. Thus there is no category for the respondent who pays exactly £300 or exactly £200 and so on. In addition, although the category 'less than £100' includes zero, respondents felt they needed to make the distinction that their true answer was zero.

less than the mean response time, which is what we were expecting if respondents were getting through these questions quickly because they were not processing the full list of response options.

In contrast to Smyth et al (2008), the 'yes/no' format did not perform similarly in CAWI and CATI modes. With the poverty questions, both CATI and CAPI respondents were more likely than CAWI respondents to answer 'yes'. With the neighbourhood questions, only CATI respondents were more likely than CAWI respondents to answer 'yes'.

Results from the cognitive interviews (which compared CAPI and CAWI for the poverty questions) raised questions about the validity of the 'yes' answers in the 'yes/no' format. There was some suggestion of slightly more 'yes' answers due to 'clarified' and 'dependent' answers occurring in the CAPI mode, slightly more 'yes' answers due to satisficing in the CAWI mode, and social desirability resulting in more 'yes' answers in the CAPI mode and more 'no' answers in the CAWI mode.

All in all, these results suggest that there could be several processes involved in the selection of more items in the 'yes/no' format compared to the 'mark all that apply' format. But differences between our study and that of Smyth et al (2006, 2008) also need to be considered.

First, there are differences in the number of items. Smyth et al (2006) used 10 to 15 items where as our study used 8. Perhaps our scales weren't long enough to trigger primacy effects. But Thomas and Klein (2006) found significant primacy effects in their comparison of 'yes/no for each' versus 'mark all that apply' in as few as 5 categories (see Experiment 2) but equally showed no evidence of primacy effects on a list of 20 categories (see Experiment 3). This would suggest that the essential prerequisite for a primacy effect may be more than just a long list.

Second, there are differences in question wording. Smyth et al (2006, 2008) included both the positive and negative categories as part of the question stem (e.g., 'Do you think that each description does or does not describe this campus?') to avoid prose that would encourage respondents to mark a 'yes' answer" (Smyth et al, 2006, p. 75). But this was not done in our experiment. Could the larger number of 'yes' answers in the interview modes come from 'yea-saying'? The work of Schuman and Presser (1981) suggests that a token alternative (e.g., Do you favour or oppose X?) provides the same results as an unbalanced question (e.g., Do you favour X?) and that the only way to change the pattern of response is to use a full alternative (e.g. 'Do

you favour X or Y?'). Also if 'yea-saying' is the cause, it is not clear why we did not find the effect among CAPI respondents on the neighbourhood questions and both CAPI and CATI respondents on the poverty questions?

With respect to the poverty questions, one explanation for the higher proportion of 'yes' answers in both CAPI and CATI could be that some of the questions are potentially sensitive as noted by cognitive interview respondents and thus interview respondents were giving polite answers. But not all of these questions are necessarily sensitive. An alternative explanation comes from the fact that respondents took longer on the poverty questions than the neighbourhood questions. There is inherent difficulty in the poverty questions. These questions addressed deep issues which required thoughtful responses compared to the neighbourhood questions. The poverty questions were the type of questions which would encourage 'qualified' and 'it depends' answers. And as seen in the results of the cognitive interviewing, such respondents would veritably choose 'yes' and were more likely to do so in CAPI as opposed to CAWI. For the neighbourhood questions, other processes could be a work. In related research we have noted a CATI positivity bias (Hope et al, 2011). These results for the neighbourhood questions were in line with the CATI positivity bias described by Christian, Dillman and Smyth (2008) and Ye, Fulton, and Tourangeau (2011). This would suggest there are no special differences by mode for the Yes/No format for CAPI and CAWI respondents for easy questions, but that CATI positivity could also be present in this format.

Third, there are differences in the populations. The Smyth et al (2006, 2008) studies were conducted among the student population at Washington State University whereas this study was based on a sample of the general population of Great Britain who used the internet. Experiment 4 from the work of Thomas and Klein (2006) included an array of countries including the United Kingdom and the United States in response to a web survey. They found no appreciable differences between how respondents answered a yes/no grid versus 'mark all that apply'. More interestingly, an analysis of the 'yes/no for each' data suggests that the characteristics of individuals answering 'yes' differed between the two series of questions in ways that could not have been detected in the studies of Smyth et al. For the poverty questions, young people aged 25 to 34 (as compared to all older respondents) may have been more likely to choose 'yes' ($p=.059$) and those with higher education or a degree (as compared to those with lower qualifications) may have been more likely to choose 'yes' ($p=.086$). An interaction term did not reach significance. In contrast, on the neighbourhood questions, it was respondents 45 and older (as compared to

younger respondents) who or were more likely to say 'yes' ($p < .001$) and although not reaching significance, it was those respondents without qualifications who were more likely to say 'yes'.[19] This further suggests that the poverty questions were considered differently to the neighbourhood questions.

The second adaptation of a question format tested in this study was the branching of four questions into two or three steps compared to non-branching. Our hypotheses that the branched format would produce more extreme responses, and that this would be more likely for attitudinal than factual questions, were clearly supported. One of the attitude variables displayed a clear significant case of the expected pattern. The same pattern was present for the other attitudinal variable, but it did not reach significance. Although a lot more inconsistent and less clear, there was a trend on the two factual questions for a contrary finding, in that the non-branched questions were more extreme.

Of great concern for the design of mixed mode surveys is our finding that the non- branched format did not produce equivalent results across all modes; many differences were observed across modes but no clear pattern was discerned. Results from the cognitive interviews revealed respondents' variability in dealing with perceived vagueness of answer categories and definitions, but there was nothing to indicate any systematic differences by mode. Further research is required to explore the causes of these inconsistent findings.

In conclusion, our initial analysis of these question formats more or less confirms previous research; we also found that the 'yes/no' format in CATI mode produces more affirmative responses than the 'mark all that apply' format in CAPI and CAWI modes, and branching of attitude questions in CATI mode produces more extreme responses than non-branching in CAPI and CAWI modes. Furthermore, these effects were also apparent within each mode, suggesting that the observed differences could be due to format effects rather than mode effects. If we follow the arguments made by previous researchers[20] we could therefore infer that the 'yes/no' format and the branching format produce better quality data and should be used across all modes. However, further analysis casts doubt on this inference and the quantitative and qualitative results

[19] After deletion of the internet uses, this later group is very small. This could account for the non-significant result.
[20] For example, more time spent on answering 'yes/no' format suggests deeper processing of response options (Smyth et al, 2006) and the branching format involves decomposing the task into smaller and easier steps (Armstrong, 1975).

from this study suggest that the 'yes/no' format and the branching format are not functionally equivalent across all three modes.

So to answer the question in the title, our results show that it is not a good idea to 'optimise' these particular question formats for the data collection mode if comparable data are to be collected using different modes. This could imply that a uni-mode approach would be preferable; i.e, the same question format should be used across modes (Dillman, 2000). However, our results also show that using the same format across all modes will not necessarily produce comparable data. The largest differences in this study were found when using 'yes/no' for CATI and 'mark all that apply' for CAPI and CAWI ($p<.000$ in all comparisons). Nonetheless, we acknowledge past research which has shown that the 'yes/no' format produces better quality data than the 'mark all that apply' format. For this reason we would still recommend using the 'yes/no' format in all modes until our results have been replicated elsewhere. However, we hope that we have shown that survey designers should be cautious about using the 'yes/no' format if questions are potentially sensitive and the socially desirable response is 'yes'. And researchers should also be aware of the risk of CATI positivity bias.

With respect to the branching and non-branching of questions, more research is needed on the effects of branching on responses to factual questions. All of the articles in our literature review studied subjective phenomenon. For subjective questions, we would suggest based on the majority of prior research to assume that branching is a better format than non-branching, but our inconsistent findings suggest that more research is still needed to fully understand how branching affects responses. And most importantly, more research is needed on the effects of branching across modes.

Since we included the web as one of our data collection modes in the experiment, we had to restrict the mode experiment to respondents with web access so that any observed differences between the groups could be attributed to mode rather than differences in the responding samples. This could cast some doubt on the ability to generalise the results from this experiment to the general population.[21] Nonetheless, we have a stronger basis for extrapolation to the general population than many other mixed mode experiments that have had to rely on samples of specific

[21] At the time of writing this paper, key analyses from the 'yes/no' versus 'mark all that apply' part of the paper were re-run using all cases from CAPI and CATI, not just those restricted to internet access, and no differences in results were found.

groups such as students. Furthermore, the question experiments from the NatCen Omnibus survey were replicated on the BHPS, with the BHPS findings closely coinciding with those of the Omnibus. Nonetheless, given the discrepancies found between this study and other studies, we strongly encourage researchers to design mixed mode experiments that use samples that are drawn from a broader spectrum of the general population.

Table 10: Summary of results

Hypothesis	**Supported or Rejected**
A1: When using a 'yes/no' series in CATI and 'mark all that apply' in CAPI and CAWI, expect a higher percentage of items chosen in CATI than in CAPI and CAWI.	Supported
A2: When the 'mark all that apply' format is used in CAPI and CAWI, expect no differences between CAPI and CAWI.	Supported
A3: When a 'yes/no' series in used in all modes, expect no differences between CATI, CAPI and CAWI.	Rejected
A4a: Expect longer completion times with a 'yes/no' series than 'mark all that apply'.	Supported
A4b: Expect respondents who answer the 'mark all that apply' question under the mean response time to show evidence of primacy effects	Rejected
A4c: Expect respondents who spend at least the mean response time to complete the 'mark all that apply' question, to select as many items as those who complete a 'yes/no' series.	Rejected
A5: Expect the observed response differences between 'yes/no' and 'mark all that apply' caused by format to be greater for the difficult rather than the easy questions.	Rejected
B1: When using question branching in CATI and not in CAPI and CAWI, (B1a) expect more extreme responses in CATI compared to CAPI & CAWI. (B1b) Expect this effect to be more prevalent (in the expected direction) for attitudinal than factual questions.	 Supported Supported
B2: Within each mode, (B2a) Expect more extreme responses when branching is used compared to non-branching. (B2b) Expect this effect to be more prevalent (in the expected direction) for attitudinal than factual questions.	Supported for attitude questions but not for factual questions. Supported
B3: Expect no difference in extreme responses between modes when branching is used across all modes.	Rejected

References

Armstrong, J.S., Denniston, W.B., and Gordon, M.M. (1975). "The Use of the Decomposition Principle in Making Judgments." Organizational Behavior and Human Performance, 14(3): 257-263.

Campanelli, P., Gray, M., Blake, M. and Hope, S. (2011). Mixed Modes and Measurement Error: Using Cognitive Interviewing to Explore the Results of a Mixed Modes Experiment, Technical Report, National Centre for Social Research, UK.

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

Dillman, D. (2000). Mail and Internet Surveys -The Tailored Design Method (2nd Edition). Hoboken, New Jersey: Wiley.

Gray, M., Blake, M. and Campanelli, P. (2011). "The Use of Cognitive Interviewing Methods to Evaluate Mode Effects in Survey Questions", paper to be presented at the fourth Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland.

Groves, R.M. (1979). "Answers and questions in telephone and personal interview surveys." Public Opinion Quarterly, 95:199-204.

Groves, R.M. and Kahn, R.L. (1979). Surveys by Telephone: A National Comparison with Personal Interview. New York: Academic Press.

Hope, S. (2005). Scottish Crime and Victimisation Survey Calibration exercise: a comparison of survey methodologies, Research report for The Scottish Executive, MORI Scotland. (Web only publication) Accessed on 12 June 2011 at http://www.scotland.gov.uk/Publications/2005/12/22132936/29366 .

Hope, S., Campanelli, P., Nicolaas, G., Jäckle, A., Lynn, P., and Nandi, A. (2011). "The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration", paper presented at the fourth Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland, 18-22 July 2011.

Jäckle, A., Lynn, P., Campanelli, P., Nicolaas, G., Hope, S. and Nandi, A. (2011). "Causes of Mode Effects on Survey Measurement", paper presented at the fourth Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland, 18-22 July 2011.

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

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

Lozar Manfreda, K. and Vehovar, V. (2002). "Mode effect in web surveys." In the proceedings from The American Association for Public Opinion Research (AAPOR) 57th Annual Conference, 2002. Accessed on 30 November 2011 at http://www.amstat.org/sections/srms/Proceedings/y2002/files/JSM2002-000972.pdf

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I. and Vehovar, V. (2008). "Web surveys versus other survey modes: a meta-analysis comparing response rates." International Journal of Market Research; 50: 79-104.

Malhotra, N., Krosnick, J.A. and Thomas, R.K. (2009). “Optimal design of branching questions to measure bipolar constructs.” Public Opinion Quarterly, 73(2): 304-324.

Miller, P. (1984). “Alternative Question Forms for Attitude Scale Questions in Telephone Interviews.” Public Opinion Quarterly, 48(4): 766-778.

Nicolaas, G., Thomson, K. and Lynn, P. (2000). “Feasibility of Conducting Electoral Surveys in the UK by Telephone.” National Centre for Social Research.

Nicolaas, G. and Lynn, P. (2002). “Random-digit dialling in the UK: viability revisited.” Journal of the Royal Statistical Society, Series A; 165(2):297-316.

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., and Singer, E. (eds) (2004). Methods for Testing and Evaluating Survey Questionnaires, Hoboken, NJ: Wiley.

Rasinski, K.A., Mingay, D. and Bradburn, N.M. (1994). “Do Respondents Really ‘Mark All That Apply’ on Self-Administered Questions?” Public Opinion Quarterly, 58:400–408.

Schonlau, M., Zapert, K., Simon, L., Sanstad, K., Marcus, S., Adams, J., Spranca, M., et al. (2003). “A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey.” Social Science Computer Review, 22(1), 128-138. Accessed on 30 November 2011 at http://www.schonlau.net/publication/03socialsciencecomputerreview_propensity_galley.pdf

Schuman, H. and Presser, F. (1981) Questions And Answers in Attitude Surveys, New York, Academic Press.

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

Smyth, J.D. and Pearson, J.E. (2011). “Internet Survey Methods: A Review of Strengths, Weaknesses, and Innovations.” In M. Das, P. Ester, and L. Kaczmirek (eds), Social and Behavioural Research and the Internet: Advances in Applied Methods and Research Strategies. London: Routledge.

Sudman, S. and Bradburn, N.M. (1982). Asking Questions. San Francisco: Jossey-Bass.

Sykes, W. and Collins, M. (1988). “Effects of mode of interview: experiments in the UK”. In R. Groves et al. (eds), Telephone Survey Methodology. New York: Academic Press.

Thomas, R. and Klein, J. (2006). “Merely incidental?: Effects of response format on self-reported behaviour.” Journal of Official Statistics, 22(2), 221-244.

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

van Soest, A. & Kapteyn, A. (2009). “Mode and Context Effects in Measuring Household Assets.” Chapter 12 in M. Das, P. Ester and L. Kaczmirek (eds.), Social and Behavioural Research and the Internet: Advances in Applied Methods and New Research Strategies. London: Routledge.

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

Yu, J.H., Albaum, G. and Swenson, M. (2003). “Is a central tendency error inherent in the use of semantic differential scales in different cultures?” International Journal of Market Research, 45(2), 213-228.