Revisiting "yes/no" versus "check all that apply": Results from a mixed modes experiment

Gerry Nicolaas
Ipsos MORI
London, UK

Pamela Campanelli
The Survey Coach, UK

Steven Hope
University College
London, UK

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

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

The work of Smyth, Dillman, Christian, and Stern (2006) and Smyth, Christian, and Dillman (2008) compares "yes/no" questions to "check all that apply" questions. They conclude that the "yes/no" format is preferable as it reflects deeper processing of survey questions. Smyth et al. (2008) found that the "yes/no" format performed similarly across telephone and web modes. In this paper we replicate their research and extend it by including a comparison with face-to-face in addition to telephone and web and by using probability samples of the general adult population. A cognitive interviewing follow-up was used to explore the quantitative findings. Our results suggest there are times when the "yes/no" format may not perform similarly across modes and that there may be factors which limit the quality of answers.

Keywords: survey methods, mode of data collection, questionnaire design, check all that apply, forced choice

1 Introduction

A common question format in self-administered modes of data collection, such as mail and web, is "check all that apply". This format is equivalent to "code all that apply" in face-to-face surveys where a list of items is displayed on a showcard. However, these questions are difficult to administer in telephone interviews that rely solely on aural communication. Therefore, they are 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"[1] 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). They recommended the

"yes/no" format, along with "applies/does not apply", "true for me/not true for me" and similar formats, as more suitable alternatives. These alternatives require the respondent to look at each item individually and Sudman and Bradburn (1982) believed these alternatives could remove or reduce list order effects. 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, Mingay, & Bradburn, 1994; Smyth, Dillman, Christian, & Stern, 2006; Thomas & Klein, 2006). This is true for comparisons of individual items (Thomas & Klein, 2006) as well as for comparisons of the mean number of endorsements (Rasinski et al., 1994; Smyth et al., 2006). This finding has been replicated across various behavioral topics, languages, countries of residence (Thomas & Klein, 2006) and opinion-based items (Smyth et al., 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. Smyth, Christian, and Dillman (2008) were the first to compare the

---

Contact information: Gerry Nicolaas, Ipsos MORI, 3 Thomas More Square, London E1W 1YW, UK; e-mail: Gerry.Nicolaas@ipsos.com

---

[1]For the remainder of the paper we use the term "mark all that apply" rather than "check all that apply" or "code all that apply" as this phrase can be used for all modes.

two formats across modes (web and telephone) and found that the "yes/no" format performed similarly across both, suggesting that this format is not prone to mode effects. They also found no clear differences in item nonresponse rates between web and telephone.

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 and by using probability samples of the general adult population rather than university students (thus increasing its generalizability). We also employed cognitive interviewing techniques after the quantitative methods to enhance our understanding of the causes of differences in measurement. The research was conducted as part of a larger research grant funded to explore "mixed modes and measurement error" across a variety of question formats and levels of question sensitivity. This paper focuses specifically on question format and mode of data collection effects for "yes/no" versus "mark all that apply".

1.1   Theories and Findings in the Literature

In an ideal scenario, according to the model proposed by Tourangeau, Rips, and Rasinski (2000), a respondent undergoes four stages when answering survey questions: comprehension of the question, retrieval of the information needed to answer the question, use of this information to make a judgment, and then finally reporting of the answer. The extent to which respondents actually complete all four stages varies. There may be overlap in the stages, backtracking and skipping or truncating a stage. There may also be errors at each stage (e.g., respondents misunderstanding the survey questions or key concepts, not knowing or being unable to recall the needed information from memory, using an inappropriate heuristic for making a judgment or preferring to hide certain information or provide a socially desirable answer). The theory of satisficing (Krosnick, 1991) suggests that the quality of respondents' answers also depends on their ability and motivation to apply sufficient cognitive effort and on the difficulty of the task itself.

The way a respondent processes a "mark all that apply" question is different compared to a series of "yes/no" questions. A "mark all that apply" question requires considerable cognitive work from respondents as they are asked to retrieve information on a full range of possible answers, but they are also required to map their multiple answers to a list of given response alternatives. This contrasts with a series of "yes/no" questions for each item which focus on one specific item at a time and requires a simple "yes" or "no" answer.

The more difficult a question task is, the more likely that respondents will take shortcuts in coming up with answers (Krosnick, 1991; Tourangeau et al., 2000). The response alternatives for "mark all that apply" questions tend to be displayed visually. The visual display can make the task of answering these questions easier (Schwarz, Strack, Hippler, & Bishop, 1991). But the "mark all that apply" format can still remain a difficult task and satisficing in this format can take the form of a "primacy effect". It is well documented that there is a tendency for respondents to select response alternatives at the beginning of a list when these are presented visually (Krosnick & Alwin, 1987; Schuman & Presser, 1981; Schwarz et al., 1991). According to the theory of survey satisficing (Krosnick, 1991), such "primacy effects" occur because respondents tend to select the first satisfactory or acceptable response alternative(s) rather than exerting the required cognitive effort to process thoroughly all items on the list. Similarly, the cognitive elaboration model (Schwarz, Hippler, & Noelle-Neumann, 1992) predicts "primacy effects" in visual modes, because the respondent has more opportunity to consider and therefore select the response alternatives at the top of the list. However, if the response alternatives at the top of the list are not plausible, then the opposite "recency effect" is predicted.

It is generally assumed that the "yes/no" format will lead to more complete information and reduced response order effects than the "mark all that apply" format (Dillman, 2000; Sudman & Bradburn, 1982). The "yes/no" format encourages the respondent to look at every item because an answer is required for each. This should reduce primacy effects. For example, Smyth et al. (2006) found higher levels of item reporting and longer response times with "yes/no" questions compared to the equivalent "mark all that apply" questions. They inferred that the longer response times were due to deeper cognitive processing for each "yes/no" item because the clear evidence of primacy effects in the "mark all that apply" format was significantly related to the time spent. Because primacy effects appear to be reduced with the "yes/no" format (Dillman, 2000), one would also expect the number of items chosen in the "yes/no" format to be similar in all the modes. There is as yet limited evidence to support this but Smyth et al. (2008) found that the "yes/no" format performed similarly across telephone and web modes suggesting that this format may be less prone to mode effects.

However, some research on mode effects suggests that the likelihood for satisficing in general is greater in self-administration modes than interviewer-administered modes. For example, the internet allows respondents to multitask and quickly skip from one topic to the next. "This in turn may lead to more superficial cognitive processing, more top of the head answers, and more satisficing in responding to survey questions" (de Leeuw, 2005, p. 244). In contrast, interviewers control the survey process and more channels of communication are available in interviewer modes such as verbal communication (e.g., words, text) and paralinguistic communication (e.g., tone, timing, emphasis). Nonverbal communication is also available in face-to-face interviews (e.g., gestures, expressions, body posture). de Leeuw

(2005, p. 244) suggests, "the effective use of these channels facilitates the communication with the respondent and improves the total question-answer process." A well-trained interviewer can also motivate respondents to produce complete and accurate answers (e.g., through probing) and reduce the difficulty of the task by offering support and providing explanations of what is needed (Hope, Campanelli, Nicolaas, Lynn, & Jäckle, 2014; Skjak & Harkness, 2003). Furthermore, there is evidence that satisficing is more likely to occur in telephone interviews than face-to-face interviews (Holbrook, Green, & Krosnick, 2003; Jäckle, Roberts, & Lynn, 2006; Jordan, Marcus, & Reeder, 1980). It is unclear whether the findings from these general studies will affect the "mark all that apply" and "yes/no" formats.[2]

The percentage of item nonresponse may also differ between the self-administered and interviewer-administered modes. Based on the reasoning of de Leeuw (2005), Smyth et al. (2008) expected higher item nonresponse in web than telephone because it would be more difficult for telephone respondents to leave an item blank as they would have to tell the interviewer they did not want to answer. Web respondents could simply leave the item blank. But from their findings, Smyth et al. (2008, p. 111) concluded that "item nonresponse rates did not differ substantially by mode" because only one comparison showed significantly higher nonresponse for web than telephone and only two comparisons showed significantly higher nonresponse for telephone than web.

1.2 Hypotheses

Our hypotheses are based on a replication of Smyth et al. (2006) and Smyth et al. (2008) with respect to the "yes/no" format versus "mark all that apply". We therefore first hypothesize that each item would show more endorsements in the "yes/no" than in the "mark all that apply" format (Hypothesis 1). This is based on combined Computer Assisted Personal Interviewing (CAPI) and Computer Aided Web Interviewing (CAWI) data for the two question formats. The testing of this hypothesis did not include Computer Assisted Telephone Interviewing (CATI) data because as described in Section 2.2, the response lists for the "mark all that apply" questions were too long for CATI administration.

Our second set of hypotheses is based on a replication of the analyses and results of Smyth et al. (2006). Smyth et al. (2006) had concluded that the greater number of endorsements with the "yes/no" compared to the "mark all that apply" format was due to deeper cognitive processing. They reached this conclusion because of (a) the longer completion times for the "yes/no" compared to the "mark all that apply" format. They also found that respondents who spent more, rather than less, time on the task selected more items. More importantly (b), respondents who spent at least the mean completion time or more to answer the "mark all that apply"

questions selected as many items as those who completed the "yes/no" series of questions and (c) respondents who answered the "mark all that apply" question under the mean completion time showed evidence of primacy effects. Thus our second set of hypotheses is that these three findings will be replicated. This is based on combined CAPI and CAWI data for the two question formats.

Replicating Smyth et al. (2008), we would expect no differences in mean endorsements (i. e., mean number of "yes" answers) between CATI and CAWI for the "yes/no" format (Hypothesis 3a). Hypothesis 3b extends this to the expectation of no differences between CAPI, CATI and CAWI.

Replicating the results of Smyth et al. (2008), we expected no differences in mean item nonresponse rates in the "yes/no" format across CATI and CAWI (Hypothesis 4a) and Hypothesis 4b extends this to the expectation of no differences between CAPI, CATI and CAWI.

2 Methods

2.1 The Data from the Larger Research Project

This paper uses data from a larger research grant-funded project which included a longitudinal study and a quantitative mixed mode experiment to explore measurement error across modes (NatCen Social Research, 2014). The first wave of the longitudinal study included 15 questions. The data were collected by the NatCen Social Research Omnibus survey. This survey used a probability sample of adults aged 16 and over in Great Britain whereby 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 CAPI.[3] The first wave data was collected over two implementations of the Omnibus survey. The response rate for the first wave, averaged over both implementations, was 54% (response rate RR5 American Association for Public Opinion Research, 2011).

For the mixed mode experiment, NatCen Omnibus survey respondents from the first wave of data collection who agreed to be re-contacted (82.5% of respondents) were randomly allocated to one of three modes: CAPI, CATI and CAWI. Separate surveys for each of these three modes were collected by NatCen Social Research and conducted between the January and June of 2009. The questionnaire contained 67 questions which were classified according to type of question content, task difficulty, and sensitivity of the question[4]

---

[2]Although not a focus of this paper, it is also well-known that interviewers reduce the privacy of the reporting situation, which can have an impact on respondents' willingness to answer truthfully to sensitive questions (Kreuter, Presser, & Tourangeau, 2008; Tourangeau & Smith, 1996; Tourangeau & Yan, 2007).

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

[4]In advance of the survey, the 67 questions were categorized as

and contained seven different question format experiments.[5] Using the "uni-mode" approach of Dillman (2000), all aspects of the question order, wording and display were identical in all three modes. However, there was a necessary operational difference between the interviewer modes and CAWI. If CAPI and CATI respondents spontaneously said that they did not know the answer or refused to give an answer, then the interviewer was able to record this. In CAWI, all respondents who hit the "submit" button without marking an answer were prompted with the additional question "Did you intend to skip this question?".

The response rates for the mixed mode experiment were 73% for CAPI, 69% for CATI and 47% for CAWI (response rate RR5 American Association for Public Opinion Research, 2011). These response rates are based on respondents issued to the mixed mode experiment. There are, of course, the two preceding levels of nonresponse: nonrespondents to the first wave of NatCen Omnibus survey data collection and those who participated in that data collection, but refused to be re-contacted. These first two levels of nonresponse do not differ by mode and therefore do not confound our mode comparisons. However, the nonresponse to the mixed modes experiment does. We avoided standard weighting because CAWI respondents only included those who had internet access and use. Propensity score weights would have been cumbersome given more than two modes. We opted for modelling and the best set of variables to control for differential nonresponse by mode proved to be sex, age, ethnicity, marital status and labor force status.[6]

For analyses comparing CAWI with other modes, only respondents who had access to and used the internet were included in comparisons. The resulting sample sizes for analysis were 282 in CAPI, 314 in CATI and 349 in CAWI.

2.2   Quantitative Analysis Methods

This paper focuses on the "mark all that apply" versus "yes/no" experiment. The questionnaire contained two split ballot experiments contrasting the "yes/no" format with "mark all that apply". One was based on eight different suggestions to reduce poverty. The other was based on eight attributes you could like about your neighbourhood (see Figures 1 and 2, respectively). Showcards were used in CAPI for the "mark all that apply" format. The "yes/no" format was used in all three modes, but the "mark all that apply" questions were only used in CAPI and CAWI because administering eight category "mark all that apply" questions in CATI would not be feasible.

Analysis of variance with the control variables was used when the dependent variable was a mean and estimated marginal means were of interest. In the case of a simple two-way model this would be

$$y_{ijk} = \mu + \tau_j + \lambda_k + \epsilon_{ijk} \qquad (1)$$

where μ is the grand mean and τ and λ are explanatory variables. This included Hypotheses 2a-c which explored mean completion time "paradata" in seconds. Consistent with Smyth et al. (2006), completion time was measured in CAWI from when the page loaded to when the respondent clicked on the "submit" button and in CAPI from when the interviewer entered the page to when the interviewer clicked on the "next" button. Primacy effects were defined as any pattern of endorsement where only items from the top half of the list were chosen when using the "mark all that apply" format. Analysis of variance was also used for Hypothesis 3a-b, where a summary measure across items was created so that mean number of "yes" answers could be investigated. It was also used for Hypotheses 4a-b, where item nonresponse was measured by adding together the number of missing answers from don't knows and refusals for each respondent.

Logistic regression with the control variables was used when the dependent variable was a dichotomy.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \quad , \quad (2)$$

where the log of the odds (probability of the outcome happening over it not happening) is equal to a standard linear regression model, with x_m being the explanatory variables used. This included Hypothesis 1, where differences in the percentage of endorsements in each format were compared. It also included Hypotheses 4a-b where item nonresponse in the "yes/no" format was considered for each of the individual items. (This was in addition to the analysis of variance of the mean number of items with missing data.)

2.3   Cognitive Interviewing Methods

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.

Respondents were recruited for the cognitive interviewing phase from respondents who had participated in the Nat-

---

an easy/difficult recall task as part of an easy/difficult comprehension task as part of being a non-sensitive/sensitive question. This was done with each researcher determining his or her on views and differences worked out through group discussions. Neither of the two question series used in this paper were categorized as sensitive.

[5](1) short versus long scales, (2) rating versus ranking, (3) agree/disagree statements versus forced choice, (4) "yes/no" versus "mark all that apply", (5) branching versus non-branching, (6) fully-labelled versus end-labelled scales and (7) showcard versus no showcard on long lists in CAPI.

[6]Note that although respondent education levels were among the variables considered, education was not related to differential nonresponse by mode.

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

GB22. Would investing in education for children reduce poverty?
 Yes   1
 No   2

GB23. Would improving access to childcare reduce poverty?
 Yes   1
 No   2

GB24. Would the redistribution of wealth reduce poverty?
 Yes   1
 No   2

GB25. Would increasing trade union rights reduce poverty?
 Yes   1
 No   2

GB26. Would reducing discrimination reduce poverty?
 Yes   1
 No   2

GB27. Would increasing income support reduce poverty?
 Yes   1
 No   2

GB28. Would investing in job creation reduce poverty?
 Yes   1
 No   2

Figure 1. The individual poverty questions used for the "yes/no" versus "mark all that apply" experiment.[a,b]

[a] Questions are presented as they would appear in "yes/no" format for CAPI and CATI where each of the items from the "mark all that apply" version becomes a separate question.

[b] Questions are 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.

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

N57. Do you like your neighbourhood because it feels safe?
 Yes   1
 No   2

N58. Do you like your neighbourhood because of the neighbours?
 Yes   1
 No   2

N59. Do you like your neighbourhood because of the character of its buildings?
 Yes   1
 No   2

N60. Do you like your neighbourhood because of its cleanliness?
 Yes   1
 No   2

N61. Do you like your neighbourhood because of its location?
 Yes   1
 No   2

N62. Do you like your neighbourhood because it is quiet?
 Yes   1
 No   2

N63. Do you like your neighbourhood because of its transport?
 Yes   1
 No   2

Figure 2. The individual neighbourhood questions used for the "yes/no" versus "mark all that apply" experiment.[a,b]

[a] Questions are presented as they would appear in "yes/no" format for CAPI and CATI where each of the items from the "mark all that apply" version becomes a separate question.

[b] Questions have been adapted from a London Housing Association questionnaire.

Cen Omnibus mixed modes experiment. Rather than selecting a "representative" sample of respondents from the mixed modes experiment, we ideally wanted respondents who had displayed mode effects in the mixed modes experiment. But mode effects are typically detected at the aggregate rather than individual level. Fortunately, quantitative analysis of the mixed modes data showed that there were two aspects of respondent behavior that could be identified at an individual level that when aggregated, varied across modes. These behaviors were (1) agreeing to opposite statements (a typical indicator of acquiescence behavior) and (2) misunderstanding a ranking task and giving the same ranking to all items or to all but one of the items (i. e., non-differentiation). The respondents from these two groups were more likely to be

those with lower levels of education and income, not working (or if working, in lower level occupations), a social (i. e., government supported) renter, and non-white. A third group of contrasting respondents was also included. They had higher levels of education and income, were in higher level occupations, of white ethnicity and house owners. Thirty seven respondents from these three groups were recruited and interviewed.[7]

Cognitive respondents were first interviewed with a subset of survey questions (10 minutes worth) from the mixed modes experiment questionnaire because the original quantitative survey had been conducted five months before. The survey questions were carefully selected based on the quantitative analyses. All of the surprising findings, as well as a selection of other findings that deserved further investigation, were selected and incorporated into a questionnaire with six versions. Six versions were used because there were more issues to investigate than would fit into the 10 minutes survey question part of the cognitive interview. Each of the six questionnaire versions consisted of questions in CAPI, CATI and CAWI as needed to follow up on particular findings. Each respondent was exposed to questions in each of the modes in that order.[8] In all versions, the plan was to administer the survey question in standard quantitative fashion and mimic the modes as closely as possible. This involved the interviewer sitting with the 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).[9] Sets of questions with a particular format (e.g., "yes/no"), were divided into two parts based on the results from the quantitative analysis with the goal of creating two equivalent groups of questions which could be used in different modes. This ensured that no respondent was asked the same question more than once.

The administration of the survey questions was followed by the actual cognitive interview component in the form of retrospective think-alouds which lasted approximately 50 minutes. The retrospective think-alouds proceeded by reminding the respondent of the survey question, data collection mode, his or her answer and any behavior displayed whilst answering e.g., hesitation. The respondent then talked through how he or she had gone about answering the question and how he or she had decided on the answer.

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

All cognitive interviews were audio-recorded with respondent consent and then transcribed. The transcribed information was summarized to capture the key points from re-

spondents in a more efficient way. The summarized information was then entered into the qualitative data management program, "Framework", which was used for analysis. The Framework program also allowed analysis of different sub-groups of respondents based on further variables, such as mode, the respondent's answer to the survey question at the start of the cognitive interview, etc. The goal was to look for anything in the response process that was seen to differ by mode. The themes that were revealed were then written down and compared across modes. This also included the detection of respondent "satisficing". With the cognitive interview data, it was possible to make a distinction between "clear" satisficing, "possible" satisficing and no satisficing. An example of clear satisficing is "I'll be truthful, I just answered that, with no thought in my head" (male, 60 or older, no qualifications, low income, White British). Any cases which were not obvious or "clear" satisficing were categorized as "possible" satisficing.

Many researchers consider cognitive interviewing to be a qualitative method. As such the use of quantification would be precluded. But to try to understand mode differences, which are usually manifested at the aggregate level, it was difficult to avoid looking at the magnitude of the differences across modes. A compromise was the use of occasional vague quantifiers like "a few" or "most".

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).

For this paper, the cognitive interviewing protocol involved an exploration of the "yes/no" format in the poverty questions as these showed the clearest mode effects. Six of the eight poverty questions were chosen and the comparison took place between CAPI and CAWI administered questions.

3   Results

3.1   Hypothesis 1: Investigating Endorsement Levels

The results in Table 1 confirm Hypothesis 1. Combining CAPI and CAWI responses, the "yes/no" format produced significantly higher endorsements compared to the "mark all

---

[7]Thirty six interviews had been planned (18 who showed acquiescence behavior; 9 who had misunderstood the ranking task and 9 contrasting respondents). One interviewer had conducted six interviews but still needed a particular type of person to fulfill her quota, so an additional interview was conducted.

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

[9]The cognitive interviews ended with a discussion of what respondents thought of the three modes they had experienced. As outlined in Campanelli, Blake, Mackie, and Hope (2015), an analysis of the themes from this discussion showed that the mode mimicking had been effective. Respondents brought up comments that were clearly mode related.

that apply" format for each of the eight items in both the poverty and the neighbourhood question series. Even when the Bonferroni method was used to adjust alpha for multiple tests, all results remained significant at the $p < .001$ level.

In line with Smyth et al. (2008), we also looked to see if these results replicated within mode. Within CAPI, all of the sixteen items across both question series showed significant differences in the expected direction ($p < .05$ in all cases). Nine of these remained significant after the Bonferroni adjustment. Similarly within CAWI, all of the sixteen items across both question series showed significant differences in the expected direction ($p < .05$ in all cases). Ten of these remained significant after the Bonferroni adjustment.

3.2 Hypothesis 2: Investigating Completion Times

Hypothesis 2a. Table 2 shows that significantly longer completion times were taken with the "yes/no" format than "mark all that apply" for CAPI and CAWI respondents across both the poverty and neighbourhood questions. This supports Hypothesis 2a. In addition, there was 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. The F test statistics for the interaction terms were $F = 78.743$ ($p < .001$) for the poverty questions and $F = 135.82$ ($p < .001$) for the neighbourhood questions. 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 was used. The relative standard deviations were roughly similar for CAPI and CAWI within the "yes/no" format, but within 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 came from the cognitive interviewing phase).

Hypothesis 2b. Table 3 shows that Hypothesis 2b was not supported. Combining the data from CAPI and CAWI, the mean number of items endorsed was still higher in the "yes/no" 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.

Hypothesis 2c. Hypothesis 2c was not supported. The analyses of variance with the control variables showed no differences in selecting the top four categories (a primacy effect) between respondents who had completed the "mark all that apply" format in less than the mean response time and those who took the mean response time or more. This was true for both the poverty and neighbourhood questions.

3.3 Hypothesis 3: Investigating Endorsement Levels for the "yes/no" Format by Mode

Hypothesis 3a. Table 4 shows no support for Hypothesis 3a that there would be no differences in mean endorsements (i. e., the mean number of "yes" answers) between CATI and CAWI for the "yes/no" format. For both the poverty questions and the neighbourhood questions, pairwise comparisons showed that CATI means were significantly higher than CAWI means.

Hypothesis 3b. The expectation of no differences also does not hold when extended to include CAPI. For the poverty questions, pairwise comparison showed that CAPI means were significantly higher than CAWI means but not statistically different from CATI. For the neighbourhood questions, however, pairwise comparisons showed that CAPI means were not significantly different from CAWI means and that CAPI means were almost lower than CATI means with $p < .057$.

3.4 Hypothesis 4: Investigating Item Nonresponse Levels for the "yes/no" Format by Mode

Overall there was only a small amount of item nonresponse in the data. Across the eight poverty questions, item nonresponse ranged from 0.6 to 4.6% and for the eight neighbourhood questions, it ranged from 0 to 1.7%.

As described in the Section 2.2, we also created a summary measure for each question series to capture the number of non-answered items by respondents. Overall, 88.8% of respondents answered all eight poverty questions and 96.4% answered all eight neighbourhood questions. For the poverty questions, 8.0% of the respondents had left only one item unanswered, 2.5% had left two items unanswered and 0.7% had left three or four items unanswered. For the neighbourhood questions, 2.8% of the respondents had left only one item unanswered, 0.8% had left 2-4 items unanswered.

Hypothesis 4a. Hypothesis 4a (that there would be no differences in mean item nonresponse levels across CATI and CAWI) was not supported. The estimated marginal means were 0.274 for CATI and 0.097 for CAWI. This difference was significant (difference = 0.177, s.e. = 0.052, $p < .01$). Smaller, but still significant differences were also found with the neighbourhood questions. The estimated marginal means were 0.068 for CATI and -0.018[10] for CAWI. This difference was significant (difference = 0.087, s.e. = 0.033, $p < .05$).

Hypothesis 4b. Hypothesis 4b was also not supported. The estimated marginal means were 0.272 for CAPI and 0.097 for CAWI. This difference was significant (difference=0.175, s.e. = 0.054, p<.01). Although smaller, differences were also found with the neighbourhood questions. The estimated marginal means were 0.054 for CAPI and

---

[10]In reality, there were no missing cases in CAWI.

Table 1

Differences in the percentage of respondents endorsing "yes/no" versus "mark all that apply" among the general population respondents in Great Britain who used the Internet

Questions	Response format		Difference between percentages	Wald test statistic (df=1)[a,b]
	Yes/No format (% endorsed)	Mark all that apply (% endorsed)		
Poverty questions[c]	(n = 452 to 474)[d]	(n = 313)		
Increaing pensions	71.3	52.4	18.9	22.616***
Education for children	69.8	55.6	14.2	13.332***
Improving childcare access	67.3	33.2	34.1	73.201***
Redistribution of wealth	66.8	33.2	33.6	58.266***
Increasing trade union rights	26.3	5.1	21.2	37.669***
Reducing discrimination	46.9	18.8	28.1	48.044***
Increasing income support	46.1	11.5	34.6	76.431***
Investing in job creation	92.8	77.6	15.2	25.530***
Average	60.9	35.9	25.0	
Neighbourhood questions[c]	(n = 458 to 466)[d]	(n = 318)		
Community spirit	56.1	30.8	25.3	34.832***
Feels safe	84.7	62.3	22.4	34.671***
Neighbours	75.3	60.7	14.6	13.389***
Character of buildings	54.3	27.4	26.9	42.601***
Cleanliness	75.6	38.7	36.9	76.731***
Location	93.1	78.3	14.8	29.708***
Quiet	82.5	61.0	21.5	25.918***
Transport facilities	57.9	34.9	23.0	18.200***
Average	72.4	49.3	23.2	

[a] Note that these logistic regressions (with the control variables) are restricted to CAPI and CAWI respondents as "mark all that apply" with eight categories was not feasible in CATI. Analyses are also restricted to respondents with internet access and use.

[b] α was set to 0.05. With the Bonferroni adjustment for multiple tests (0.05/8 tests), the new alpha level was 0.00625.

[c] Exact wordings of questions are found in Figures 1 and 2. [d] n varies due to item nonresponse data.

*** $p < 0.001$.

-0.018 for CAWI. This difference was significant (difference=.072, s.e. = 0.034, $p < .05$).

Overall. There were two patterns present: (1) there was a slightly higher percent of item nonresponse for the poverty than for the neighbourhood questions (this is explored in the Discussion Section) and (2) there was a significantly higher mean item nonresponse rate for the interviewer modes than for CAWI. This latter finding could be due to the differences in recording item nonresponse in CAPI and CATI compared to CAWI. As discussed in Section 2.1, interviewers in CAPI and CATI recorded item nonresponse when the respondent spontaneously said that they either did not know the answer or did not want to say, whereas all CAWI respondents who tried to move to the next screen without marking an answer were prompted with the additional question "Did you intend to skip this question?"

3.5 Cognitive Interview Findings

The cognitive interviewing results were based on six of the eight poverty questions in CAPI and CAWI. As described in Section 2.3 on the cognitive interviewing methods, the cognitive data consisted of retrospective think-alouds of respondents talking about why they chose the answer they did to the survey questions. The data for each of the poverty questions were explored for themes.

For each of the six questions, themes were placed under one of four headings based on how the respondent had answered the survey question. These consisted of CAPI-YES, CAPI-NO, CAWI-YES and CAWI-NO. Themes represent summarized data, each theme containing comments from one or many respondents. Table 5 shows an example of this using the "reducing discrimination" item.

Using the example of Table 5, firstly it can be seen that a lot of similar themes appeared in both modes. One exception was the instance of "possible" and "clear" satisficing. These only appeared in the CAWI mode. This was part of a larger

Table 2

Mean completion time by question series, mode of data collection and format among general population respondents in Great Britain who used the internet[a]

Mode	Poverty questions		Neighbourhood questions	
	Yes/No	Mark all	Yes/No	Mark all
CAPI				
n	129	143	138	130
Est. marginal mean completion time in sec.[b]	91.69	61.00	58.94	43.72
Mean difference (Std. Err.)	30.69 (2.96)***		15.22 (1.76)***	
Std. Dev.	35.94	35.40	20.55	21.03
Relative Std. Dev.[c]	39.19	58.02	34.86	48.09
CAWI				
n	163	157	153	183
Est. marginal mean completion time in sec.[b]	94.74	28.53	63.88	21.26
Mean difference (Std. Err.)	66.21 (2.73)***		42.62 (1.58)***	
Std. Dev.	39.03	37.55	21.93	23.38
Relative Std. Dev.[c]	41.20	131.63	34.33	109.96

[a] Note that these analyses of variance are restricted to CAPI and CAWI respondents as "mark all that apply" with eight categories was not feasible in CATI.   [b] While holding constant the control variables. Outliers greater than two standard deviations removed because similar results were found with and without their exclusion. Also excluded from both formats were cases where respondents did not endorse any items.   [c] $\left| \frac{\text{std. dev.}}{\text{mean}} \right| \cdot 100$
*** $p < 0.001$.

pattern that was noted on the other five poverty questions and on other questions in the larger mixed modes project, i. e., that satisficing was more common in CAWI than CAPI. Across the six poverty questions, almost all of the satisficing responses were in the "yes" category. This could indicate that in the absence of a middle category that "yes" was an easy answer. As shown in Table 5, one respondent specifically said she "erred on the side of "yes"".

Secondly, the think-aloud data themes in Table 5 show that although respondents chose a "yes" or "no" to the survey question, some were really in a middle ground (saying "possibly yes", "it depends", "hopefully yes"). Looking across all six of the poverty questions, there was a clear pattern that these respondents in the "middle ground" were much more likely to have chosen a "yes" to the survey question than a "no". Of these respondents in the "middle ground", more were in CAPI than were in CAWI.

Thirdly (not shown in Table 5), respondent comments suggested that unexpectedly, two of the six poverty questions could be a bit sensitive. These were "income support" (a UK government benefit) and "redistribution of wealth". On the "increasing income support" question, 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, secondary school equivalent education, low income, white British). On the "redistribution of wealth" question, 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, high income, white British).[11] There is usually a difference between modes with respect to giving a socially desirable answer (i. e., more likely in CAPI than CAWI in quantitative data). For both questions, the cognitive data showed slightly more "no" answers, the socially undesirable answer, in CAWI than in CAPI.

So overall, these findings raise questions about what a "yes" answer means in the "yes/no" format. The "yes" answers, as opposed to the "no" answers, included more respondents who gave a satisficing answer and in contrast, more respondents who had thought about the question and found themselves in the middle ground, wanting an option between "yes" and "no". The "yes" answers for some of the questions may have also resulted from respondents giving a socially desirable answer. The cognitive interview findings suggested that the "yes" answers due to satisficing may be more likely to occur in CAWI, whereas the "yes" answers from thoughtful respondents desiring a middle option may

---

[11] As discussed in footnote 4, in advance of the survey the research team had given the poverty questions an overall categorization of non-sensitive and this is true for most of the individual items. But in retrospect we would agree with the cognitive interview findings that these two items may be sensitive for some people. It is not possible to know how many respondents would have been affected by this (more on this in the Discussion Section under the heading of Exploration of Mode Differences).

Table 3

Mean number of items endorsed by "mark all" respondents expending at least mean amount of time or more compared to "yes/no" respondents[a]

	n	Est. mean number of items endorsed[b]	Mean difference	F-Test Statistic
Poverty questions				
Yes/No (All respondents)	288	5.34	1.64	74.803[***]
Mark all (Just respondents expending more than the mean amount of time)	105	3.70		
Neighbourhood questions				
Yes/No (All respondents)	307	5.77	1.66	50.455[***]
Mark all (Just respondents expending more than the mean amount of time)	117	4.11		

[a] Note that these analyses of variance are restricted to CAPI and CAWI respondents as "mark all that apply" with eight categories was not feasible in CATI.   [b] While holding constant the control variables. Outliers greater than 2 standard deviations removed. Also excluded from both formats were cases where respondents did not endorse any items.
[***] $p < 0.001$.

Table 4

Mean number of items endorsed by "yes/no" respondents, by question series and mode of data collection

	n	Estimated mean number of items endorsed[a]	Mean Difference of Comparisons to Reference Group	t-test statistic
Poverty questions				
CAPI	100	5.86	0.78	3.924[***]
CATI	113	5.61	0.53	2.798[**]
CAWI	178	5.08	Ref.	
Neighbourhood questions				
CAPI	141	5.74	−0.40	−1.905
CATI	142	6.14	Ref.	
CAWI	166	5.63	0.51	2.532[*]

[a] While holding constant the control variables.
[*] $p < 0.05$,   [**] $p < 0.01$,   [***] $p < 0.001$.

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.

Interestingly the latter two findings were in line with the quantitative mode results of more "yes" answers in CAPI than CAWI, whereas the first more general finding about satisficing was at odds with the quantitative mode results as it suggests more "yes" answers in CAWI. The effect of these different types of respondent patterns on the quantitative data would depend on the magnitude of the occurrence of each behavior pattern relative to each other in the population under study.

4 Discussion

4.1 Review of our Findings

In this paper we have explored differences between the "mark all that apply" format (which is commonly used for face-to-face interviews and self-administered modes) with a series of "yes/no" questions for each item (which is commonly used for telephone). In doing so, we have replicated and extended the work of Smyth et al. (2006, 2008) to include an additional mode, CAPI, a broader general adult population rather than university students and the use of cognitive interviewing after the quantitative data collection.

Table 5
Themes from respondents' think-alouds during the retrospective cognitive debriefing: Example of the answers to the poverty item on reducing discrimination

CAPI Respondents	CAWI Respondents
Clear reasons for choosing "yes"	
With respect to workplace:	With respect to workplace:
• General about getting a job	• If job goes to someone else "person who's discriminated against might be left in poverty"
• Women still suffer inequality	
General comments about discrimination:	Personal experience:
• Discrimination due to race and colour in general – they could lose out	• Yes because self-experience of being an agency worker
Qualified or dependent reasons for choosing "yes"	
Possibly:	
• Possibly, but there are other ways which are more effective	
It depends:	
• Depends on what's being done and how	
• Would have liked a category between "yes" and "no"	
Hope it will:	
• Not sure but hopes it will	
No discrimination exists:	
• Yes, but people are not "allowed to discriminate as lots of laws to prevent"	
"Yes" answer given, but problems with response process	
	Resp. didn't understand what was meant by or what type of discrimination
	Possible satisficing:
	• Less easy to answer but erred on the side of "yes"
	Clear satisficing:
	• "I'm not very sure really how, how I've come to that answer"
	• "To tell you the truth, I just clicked it"
Clear reasons for choosing "no"	
Can't see any connection between discrimination and poverty:	Can't see any connection between discrimination and poverty:
• Can't see why it would	• "I don't think discrimination has got anything to do with poverty"
• No influence on poverty at all	• "If you're a different race or ethnic background, I don't think it automatically means that you are poverty-stricken"
• Have never connected discrimination with poverty	
No discrimination exists:	No discrimination exists:
• Doesn't see gender discrimination in jobs	• Don't honestly believe there is discrimination in the workplace nearly as much as there used to be
• No discrimination in benefit systems	
"No" answer given, but problems with response process	
	Resp. didn't understand what was meant by "discrimination"

Hypotheses 1 and 2: Question format effects. We found that overall 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 the research of Smyth et al. (2006) (as well as Rasinski et al. (1994) and Thomas and Klein (2006)). We found that respondents took more time to answer the "yes/no" questions compared to the "mark all that apply" questions. But using Smyth et al. (2006) indicators of better quality we did not see any evidence that completion times were linked to the thoroughness of respondents' processing. Those respondents who took longer than average to complete the "mark all that apply" questions did not select as many items as those who completed the "yes/no" questions. And those respondents who took less than the average completion time to complete the "mark all that apply" questions showed no clear evidence of primacy effects. This is dissimilar to the findings of Smyth et al. (2006).

Differences in design. There are some differences between our research and that of Smyth et al. (2006). (1) Finding no primacy effects could mean that our scales may not have been long enough to trigger primacy effects. Smyth et al. (2006) used 10 to 15 items where as our study used eight. But Thomas and Klein (2006) found significant primacy effects in their comparison of "yes/no" versus "mark all that apply" in as few as five categories (see their Experiment 2) but equally showed no evidence of primacy effects on a list of 20 categories (see their Experiment 3). This would suggest that the essential prerequisites for a primacy effect may be more than just a long list. (2) Smyth et al. (2006) used both self-reported behavior as well as opinion-based questions whereas this paper focused on opinion-based questions. Smyth et al. (2006) found evidence to suggest that respondents were more thoughtful with the opinion than the factual questions. But using only opinion questions in this paper does not impair our replication of their work. (3) There was a difference in how primacy was measured. Smyth et al. (2006) were able to include questions in reverse order. We measured the number of items selected from among the first four (top half of the list) by completion time. If there had been primacy effects, we would have expected more items selected among the first four for respondents who completed the task in less than the mean response time. This was not found.

Although these differences exist between our design and that of Smyth et al. (2006) we do not believe these differences are large enough to distort the comparison.

Hypotheses 3 and 4: Mode effects. In contrast to the findings of Smyth et al. (2008), the "yes/no" format did not perform similarly across modes. Firstly, both CATI and CAPI respondents were more likely than CAWI respondents to answer "yes" to the poverty questions. With the neighbourhood questions, only CATI respondents were more likely than CAWI respondents to answer "yes" whereas

CAPI and CAWI respondents did not differ. Secondly, the mean item nonresponse rate was not similar across modes for the poverty and neighbourhood questions. CAPI and CATI respondents had higher mean levels of item nonresponse than CAWI respondents for both question series.

Results from the cognitive interviews (which compared CAPI and CAWI for six of the poverty questions) raised questions about the content of the "yes" answers in the "yes/no" format. "Yes" answers could include more satisficing and socially desirable answers. But could also include respondents who had thought thoroughly about the topic and decided they were in a middle position between "yes" and "no". The use of the "yes" category differed by mode. "Yes" answers due to a "thoughtful" middle position and due to a possible socially desirable response were slightly more common in CAPI whereas "yes" answers due to satisficing were more common in CAWI. All in all, these results suggest that there could be several processes involved in the selection of items in the "yes/no" format and depending on the distribution of these types of behaviors in the population, mode comparisons could be affected in different ways.

Exploration of the Mode Differences. As discussed in the section on Hypotheses 3 and 4: Mode effects, there were mode differences between the interview modes and CAWI on both the mean number of "yes" answers and the mean number of items resulting in item nonresponse, with higher means in the interviewer modes for both the poverty and neighbourhood questions. The exception was CAPI respondents on the neighbourhood questions, as their mean number of "yes" answers did not differ from CAWI respondents. In terms of item nonresponse, as discussed in Section 2.1, part of the mode differences could be due to a questionnaire implementation difference, where CAWI respondents were prompted when they had failed to choose an answer. Despite this, the item nonresponse levels are slightly higher on the poverty questions than the neighbourhood questions for all three modes. Over the next four paragraphs we explore possible reasons for these results. The first paragraph explores the effects of question sensitivity and more thoughtful answers in relation to the mode effects found on the poverty questions. The second paragraph looks at the general issue of "yea-saying" for both the poverty and neighbourhood questions, detecting possible "yea-saying" for CATI respondents on the neighbourhood questions. The third paragraph explores further differences between the two question sets and what may be unique about CATI responses on the neighbourhood questions. The fourth paragraph summarizes the findings.

The cognitive interview data suggested that two of the poverty questions "redistribution of wealth" and "increasing income support" could be sensitive questions. Could the higher rates of "yes" answers and item nonresponse in the interviewer modes on the poverty questions be due to

question sensitivity? This issue can be better understood by looking at the individual items themselves. If sensitive questions are the cause of the mode differences for both the "yes" answers and item nonresponse, then one would expect the same questions to be affected. But there was no clear pattern. There were only three instances of significant item nonresponse differences by mode and one of these ("redistribution of wealth") shows no mode difference in the mean number of "yes" answers whereas the other two ("increasing trade union rights" and "reducing discrimination") do. Similarly there were three other items which showed significant mode differences in the mean number of "yes" answers, but showed no significant mean item nonresponse differences by mode. The remaining two questions showed no significant differences of any kind. There is also no evidence regarding the two questions specifically identified by the cognitive interviews as sensitive. These two questions were either not sensitive enough in general or not sensitive enough for most people in the sample for an aggregate mode effect to occur. "Redistribution of wealth" only showed differences in item nonresponse differences by mode and "increasing income support" only showed differences in the mean number of "yes" answers by mode. These findings suggest that question sensitivity is probably not the cause of the mode differences. The cognitive interviewing had suggested that CAPI rather than CAWI respondents were being more thoughtful in their answers to the poverty questions. The time data suggest no differences in completion time between CAPI and CAWI respondents, although the web design across the whole questionnaire of one question per page may have been a factor.

Could the higher rate of "yes" answers in the interview modes for the poverty questions and neighbourhood questions be due to "yea-saying" (acquiescence)?[12] There are differences in the question wording used by Smyth et al. (2006, 2008) compared to our questions. 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). It is not clear that these differences in wording would be the cause of "yea-saying". (1) There are different findings in the literature. Summarising over a number of split ballot experiments on question balance, Kalton and Schuman (1982), p.51, suggested it was not surprising that there was little difference in results between an unbalanced question (e.g., Do you favor X?) and one with a token alternative (e.g., Do you favor or oppose X?). "On the other hand, large differences have often – but not always – been found between the responses given to questions asked with and without a substantive alternative" (e.g., "Do you favor X or Y?"). In contrast, Schaeffer, Krosnick, Langer, and Merkle (2005) found that the token and full alternative questions (which they called "minimally" and "fully" bal-

anced) yielded similar distributions and levels of concurrent validity. But they did not compare their results to an unbalanced question version. (2) If "yea-saying" is the cause, it is not clear why we did not find the effect in the mean number of "yes" answers among CAPI respondents on the neighbourhood questions and both CAPI and CATI respondents on the poverty questions. We did an additional analysis looking at the percentage of respondents who had answered "yes" to all items (an extreme form of "yea-saying"). There were no significant differences for the poverty questions (CAPI = 10.0 %, CATI = 8.3 % and CAWI = 5.6%). For the neighbourhood questions, the respective percentages were CAPI = 17.0%, CATI = 23.2% and CAWI = 15.7%. The CATI/CAWI difference was statistically significant with the odds of answering "yes" to all items being increased by a factor of 1.91 for CATI respondents ($p < .05$). This suggests that there is something else going on than "yea-saying".

There is evidence that the CATI respondents on the neighbourhood questions were responding differently to CAPI and CAWI respondents. These results were found when comparing the meaning number of "yes" answers and whether the respondent had chosen all "yes" answers. These results don't hold for the poverty questions. So is the difference in results due to the different topics in the two question sets, CATI respondents or a combination of both? Differences between the two different question sets were highlighted through a further analysis by age and education.[13] Looking at those who had said "yes", some clear patterns emerged. For the poverty questions it was young people aged 25 to 34 (as compared to all older respondents, $t = 1.89$, $p < .10$) and those with higher education or a degree (as compared to those with lower qualifications, $t = 1.72$, $p < .10$) who were more likely to have said "yes". For the neighbourhood questions, it was respondents aged 45 and older (as compared to younger respondents, $t = 3.34$, $p = .001$) and those without qualifications (as compared to those with qualifications although this was not significant due to the very small sample size once non-internet users were excluded) who were more likely to have said "yes". Summarizing, on the poverty questions more "yes" answers occurred among the young people and those with high education and the opposite was true for the neighbourhood questions. This suggests the two questions sets were processed differently by respondents. Krosnick (2000) suggests that it is people with less education and who are older who are more prone to acquiesce. Summarizing over several studies, Weisberg (2005) suggests that age and education can be seen as proxies for the respondent's sophis-

---

[12] For example, Dillman and Christian (2005) and Bowling (2005) suggest the presence of an interviewer may result in respondents being more likely to select agreement-oriented response options because social encounters between people show that agreeing with one another is easier than disagreeing.

[13] Due to small sample sizes, α was set to .10 for this analysis.

tication of cognitive processing, knowledge base and memory capacity. He concluded that respondents with lower education and those who were older may be more prone to satisficing. Thus it is possible that in general the poverty questions were given more thought and neighbourhood questions given less thought. This is backed up by the completion time data. Respondents in both CAPI and CAWI[14] spent longer on the poverty questions than the neighbourhood questions. Evidence that there may be something unusual about the behaviour of CATI respondents on the neighbourhood questions is backed up by the literature and analysis of other questions that were part of the larger grant project. Dillman et al. (2009) found more extreme answering among telephone as opposed to web respondents. Hope et al. (2014) found CATI but not CAPI positivity effects on other question sets in the larger grant project (e.g., seven versus three category satisfaction questions and agree/disagree questions – separate from acquiescence bias). This suggests that part of the higher number of "yes" answers for CATI respondents on the neighbourhood questions may have been due to satisficing.

In summary, this exploration of mode differences suggests that the poverty and neighbourhood question series were processed differently by respondents. For the poverty questions, the higher mean number of "yes" answers in CAPI and CATI are probably not due to question sensitivity or "yea-saying", but may be due to more thoughtful answers because the topic of reducing poverty would engender this. For the neighbourhood questions, the higher mean number of "yes" answers in CATI but not CAPI is due to "yea-saying" and this may be reflected as satisficing or a general positivity bias.

4.2 Conclusions and Limitations

Our research is limited by having smaller sample sizes than used by Smyth et al. (2006, 2008) and having only two question sets compared to the Smyth et al. (2006) study which used three question sets in their first web experiment, six question sets in their second web experiment and one question set in their paper-based experiment. Our probability sample of the adult population of Great Britain has three levels of nonresponse with the resulting achieved sample being more cooperative than members of the general public as a whole. Nonetheless we can discern a lot from our two question sets. First, this is due to our broad population-based sample. The Smyth et al. (2006, 2008) studies were conducted among the student population at Washington State University. Having general population data also allowed for subgroup analysis by age and education which would not have been possible with student respondents. Second, having more cooperative respondents could mean that we are less likely to find measurement error (particularly measurement error due to lack of motivation). But we did find significant measurement error differences. The conclusions in the paper are not harmed by the more cooperative achieved sample that

we had.

A positive addition to our study was the use of cognitive interviewing after the quantitative survey. Results from the cognitive interviews shed further light on the unexpected mode differences between CAWI respondents and those in the interview modes for the poverty questions. For example, the research team would not have considered that the greater endorsement of "yes" answers in the interviewer modes could be due to more thoughtful answers if it had not been for the cognitive interviewing. The cognitive interview also provided a more complex understanding of how the "yes/no" format performs in different modes.

Our analyses both confirm and do not confirm Smyth et al. (2006, 2008). We clearly found respondents take more time and endorse more items in the "yes/no" format. However, our analyses cast some doubt on whether the "yes/no" format is always problem free and both the quantitative and cognitive interview results from this study suggest that the "yes/no" format is not always functionally equivalent across all three modes.

Our results show that it is not a good idea to "optimize" this particular question format for the data collection mode (i. e., "yes/no" for telephone and "mark all that apply" for other modes) 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. 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 difficult or sensitive and the socially desirable response is "yes".

Acknowledgements

References

American Association for Public Opinion Research. (2011). Standard definitions: final dispositions of case codes and outcome rates for surveys (7th ed.) Retrieved from

---

[14]CATI not included in time completion data as eight item "mark all that apply" questions could not be asked in CATI.

http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/StandardDefinitions2011_1.pdf

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27(3)*, 281–291.

Campanelli, P., Blake, M., Mackie, M., & Hope, S. (2015). Mixed modes and measurement error: using cognitive interviewing to explore the results of a mixed modes experiment. ISER Working paper, No. 2015-18. Wivenhoe, UK. Institute for Social and Economic Research, University of Essex.

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

Dillman, D. (2000). *Mail and internet surveys – the tailored design method (2nd ed.)* Hoboken, NJ: Wiley.

Dillman, D. & Christian, L. (2005). Survey mode as a source of instability in responses across surveys source. *Field Methods*, *17(1)*, 30–52.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, *38(1)*, 1–18.

Gray, M., Blake, M., & Campanelli, P. (2014). The use of cognitive interviewing methods to evaluate mode effects in survey questions. *Field Methods*, *26(2)*, 156–171.

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

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., & Jäckle, A. (2014). The role of the interviewer in producing mode effects. Results from a mixed modes experiment comparing face-to-face, telephone and web administration. ISER Working paper, No. 2014-20. Wivenhoe, UK: Institute for Social and Economic Research, University of Essex.

Jäckle, A., Roberts, C., & Lynn, P. (2006). Telephone versus face-to-face interviewing: mode effects on data quality and likely causes. Report on phase II of the ESSGallup mixed mode methodology project. ISER Working Paper, No. 2006-41. Wivenhoe, UK: Institute for Social and Economic Research, University of Essex.

Jordan, L., Marcus, A., & Reeder, L. (1980). Response styles in telephone and household interviewing. A field experiment. *Public Opinion Quarterly*, *44(2)*, 210–222.

Kalton, G. & Schuman, H. (1982). The effect of the question on survey responses: a review. *The Journal of the Royal Statistical Society, Series*, *145(1)*, 42–57.

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

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

Krosnick, J. (2000). The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Newsletter*, *20*(1). Published by the NatCen Social Research, London, UK.

Krosnick, J. & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51(2)*, 201–219.

NatCen Social Research. (2014). Mixed modes and measurement error, 2009 (computer file). Colchester, Essex: UK Data Archive (distributor), June 2014. SN: 7515. Retrieved from http://dx.doi.org/10.5255/UKDA-SN-7515-1

Presser, S., Couper, M., Lessler, J., Martin, E., Martin, J., Rothgeb, J., & Singer, E. (2004). Methods for testing and evaluating survey questions. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: Wiley.

Rasinski, K., Mingay, D., & Bradburn, N. (1994). Do respondents really mark all that apply on self-administered questions? *Public Opinion Quarterly*, *58(3)*, 400–408.

Schaeffer, E., Krosnick, J., Langer, G., & Merkle, D. (2005). Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. *Public Opinion Quarterly*, *69(3)*, 417–428.

Schuman, H. & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Schwarz, N., Hippler, H., & Noelle-Neumann, E. (1992). A cognitive model of response-order effects in survey measurement. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 187–199). New York: Springer.

Schwarz, N., Strack, F., Hippler, H., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, *5(3)*, 193–212.

Skjak, K. & Harkness, J. (2003). Data collection methods. In J. Harkness, F. vande Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods*. New York: Wiley.

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

Smyth, J., Dillman, D., Christian, L., & Stern, M. (2006). Comparing check-all and forced-choice question for-

mats in web surveys. *Public Opinion Quarterly*, *70(1)*, 66–77.

Sudman, S. & Bradburn, N. (1982). *Asking questions*. San Francisco, CA: Jossey-Bass.

Thomas, R. & Klein, J. (2006). Merely incidental? Effects of response format on self-reported behavior. *Journal of Official Statistics*, *22(2)*, 221–244.

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

Tourangeau, R. & Smith, T. (1996). Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60(2)*, 275–304.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133(5)*, 859–883.

Weisberg, H. (2005). *The total survey error approach: a guide to the new science of survey research*. Chicago: University of Chicago Press.