Yes–no answers versus check-all in self-administered modes

A systematic review and analyses

Mario Callegaro
Google UK Ltd
Michael H. Murakami
Latinum Network
Ziv Tepman
Staffing Industry Analysts
Vani Henderson
Google, Inc.

When writing questions with dichotomous response options, those administering surveys on the web or on paper can choose from a variety of formats, including a check-all-that-apply or a forced-choice format (e.g. yes–no) in self-administered questionnaires. These two formats have been compared and evaluated in many experimental studies. In this paper, we conduct a systematic review and a few meta-analyses of different aspects of the available research that compares these two formats. We find that endorsement levels increase by a factor of 1.42 when questions are posed in a forced-choice rather than check-all format. However, when comparing across a battery of questions, the *rank order* of endorsement rates remains the same for both formats. While most authors hypothesise that respondents endorse more alternatives presented in a forced-choice (versus check-all-that-apply) format because they process that format at a deeper cognitive level, we introduce the acquiescence bias hypothesis as an alternative and complementary explanation. Further research is required to identify which format elicits answers closer to the 'true level' of endorsement, since the few validation studies have proved inconclusive.

Introduction

When surveyors ask respondents to select statements that apply to them, actions they have ever taken, or items they recognise, two formats are typically used in self-administered questionnaires on the web or on paper:

Received (in revised form): 9 September 2014

DOI: 10.2501/IJMR-2015-014

a check-all-that-apply format or a forced-choice (e.g. yes–no) format. Employed in a wide variety of survey applications, these two formats have been used to investigate consumer preferences (Ares *et al.* 2010; Lado *et al.* 2010; Parente *et al.* 2011), employment status (Thomas *et al.* 2009), and type of health care coverage (Ericson & Nelson 2007), to name a few examples. Check-all or forced-choice answers are also used as screening questions to find eligible respondents for a complete study, or for particular sections of a questionnaire (Thomas 2011).

A check-all-that-apply format requires the respondent to check or mark the items of interest, with the implicit assumption that the checked items are a 'yes' and the non-checked items are a 'no'. Check-all-that-apply is also called 'mark all that apply' (Rasinski *et al.* 1994; Nicolaas *et al.* 2011), 'multiple response format' (Thomas & Klein 2006) or, simply, 'check all' (Smyth *et al.* 2006; Dykema *et al.* 2011). An alternative method of asking the same question is to have the respondent explicitly provide a 'yes' or 'no' answer (or similarly dichotomous options: applies to me, does not apply to me; describes me, does not describe me; etc.) for each item in the list. This format is also referred to as 'forced choice' (Smyth *et al.* 2006; Stern *et al.* 2007, 2012).

The forced-choice response format is generally straightforward for respondents to answer and for the researcher to code; a marked yes is a 'yes', a marked no is a 'no' and, if nothing is selected, that entry is considered item non-response. The situation is quite different for check-all-that-apply formats. As Bradburn *et al.* (2004, p. 173), Sudman and Bradburn (1982, p. 168) and Birnbaum (2001, p. 53) have pointed out, the difficulty lies in interpreting a non-checked box, since four explanations are possible:

1. the response option does not apply and therefore is a 'no'
2. the respondent might have missed that entry in the list
3. the respondent was not sure, or
4. the respondent did not want to answer the question.

Because of this coding issue, there is an unresolved debate in the survey research literature over which response format provides the most *reliable* data for self-administered surveys. Also, in mixed-mode surveys, or when comparing results across surveys of different modes, it is important to understand which format generates the most comparable (i.e. *valid*) results across phone, web and mail (Smyth *et al.* 2008). For instance, the only option for telephone interviews is to ask forced-choice questions. So is a check-all-that-apply format a valid substitute when the same question is self-administered? Or consider another example of face-to-face surveys: it is possible for an interviewer to either read respondents a list of options in forced-choice format or to display a notecard with the full list of options and ask them to indicate all that apply.[1] Thus, in mixed-mode surveys with a face-to-face component, data can be recorded in checkbox format or forced-choice format. Which format should online respondents of the same survey be presented with?

Many experiments comparing the two formats have been conducted with a randomised experimental design. We start by reviewing these experiments and then conduct several meta-analyses to summarise their collective results.

Results from randomised experiments comparing check-all-that-apply with forced-choice answer formats

Locating experimental studies

Several experiments comparing the results of check-all-that-apply versus forced-choice response options have been conducted. To find all published and unpublished papers for our review, we consulted the following databases searching for the keywords 'mark-all' and 'check-all' in the title and abstract: Ebsco, Warc, Psynet, WebSm, Pubmed, Acm, Ieee, Ssrn, Marketresearch.com and Ingentaconnect. To avoid publication bias (Dickersin 2005), we also searched the proceedings of the American Statistical Association, Joint Statistical Meeting, Survey Research Methods section; the past ten years of conference programmes of the American Association for Public Opinion Research (AAPOR); the past ten years of conference programmes of the General Online Research (GOR); and the past ten years of proceeds and publications of the European Society for Opinion and Marketing Research (ESOMAR). We also performed a Google and Google Scholar search with the same keywords. For each relevant paper found in these searches, we looked at the references cited and obtained any new titles. This 'snowballing' method was updated, and the full list of papers was finalised in October 2013. The effort yielded a total of 18 works: seven peer-reviewed journal publications, three conference proceedings, five conference papers, one conference poster,

[1] For example, in the British Household Panel Survey, showcards are used to ask which appliances are present in each household: 'Would you look at this card please and tell me if you have any of the items listed in your (part of the) accommodation?' (p. 15, showcard on page H50). Online at: www.iser.essex.ac.uk/bhps/documentation/pdf_versions/questionnaires/bhpsw11q.pdf.

one book chapter and one working paper in which check-all-that-apply answers were compared to forced-choice answers. The outcome of the search revealed that these experiments were mostly conducted in the US, with few notable exceptions (Thomas & Klein 2006; Nicolaas *et al.* 2011). Given that we used only English-language keywords, we cannot guarantee that other studies have not been published in other languages.

Paper-and-pencil survey experiments

The first bona fide experiment on the general population was conducted in an exit poll during the 1992 US Presidential Election (Mitofsky & Edelman 1995). On that occasion, voters were randomised to receive an exit poll questionnaire with some factual and opinion items in either a check-all-that-apply or a yes–no format. The yes–no version increased the estimate of average support for those items by about seven percentage points (p. 95). In another experiment, Rasinski and colleagues (1994) randomly distributed paper-and-pencil questionnaires to high-school seniors in either a check-all-that-apply or yes–no format for the same set of questions. They found that significantly fewer response options were selected in the check-all-that-apply condition (average of 4.0) when compared to the yes–no condition (average of 4.6) across three sets of response options ranging from 4 to 20 items.

Mirroring these results, Stern *et al.* (2007, 2012) conducted a survey experiment by mailing questionnaires to residents of a community in North America and found a significantly lower number of endorsed items for checkboxes (average of 3.01) versus yes–no answers (average of 3.20) in one set of six response options. In a randomised paper-and-pencil experiment with university students, Smyth *et al.* (2006) obtained a significantly lower number of 'yes' responses to 15 items in a check-all-that-apply format (average of 2.6) rather than yes–no format (average 3.8). In a paper questionnaire administered to campers in British Colombia, respondents were randomised to different versions of the same questionnaire asking about camp-related information and camping equipment used (Dyck & Moore 2008). The authors found statistically significant differences in the expected direction (more items reported in the yes–no format) for some question sets.[2]

[2] That the difference was not significant for all groups may be due to the nature of the questions. For questions regarding an event that is very salient (e.g. 'For each of the following types of camping shelters, please tell us if you did or did not use each one during your stay in the campground', asked immediately after the camping took place), asking the question as check-all or as yes–no produced the same amount of endorsements (Ziv question: Was it precisely the same amount of endorsement, or approximate? We should note if it's approximate).

The largest paper-and-pencil experiment was conducted in the 2006 American Community Survey (ACS) content test (Ericson & Nelson 2007), where 62,900 respondents were evenly randomised to a series of items asking about their health coverage either in a check-all-that-apply format or a yes–no format. The results confirmed the previous finding of higher endorsement rates on the yes–no response format, this time with a general population, probability-based sample.

Web survey experiments

Smyth *et al.* (2006) and Thomas and Klein (2006) were the first to extend these findings to web surveys. In the Smyth study, a sample of Washington State University students was randomly assigned to either a check-all-that-apply or a yes–no format in two web surveys about campus life. The forced-choice option elicited significantly more responses than the check-all-that-apply option. The overall mean of endorsed items was 3.3 in the checkboxes format versus 4.1 in the yes–no format, out of three sets of response options ranging from 10 to 15 items (web experiment 1). Similarly, the second experiment found significantly lower endorsements for checkboxes. Thomas and Klein (2006) conducted five separate randomised experiments with respondents from the Harris Interactive panel, some with US-only participants and some with international participants. The results were similar to the previous studies: the average score for the check-all-that-apply format was always lower than the yes–no condition. However, in their study, respondents were required to answer each item of the yes–no condition before they could advance to the next screen (p. 240), which may have confounded the question format effect. All the other studies reviewed here, including our experiment, do not force answers in the yes–no condition nor remind the respondents for missing items.

In a web experiment on a sample of Washington State University students, Smyth *et al.* (2008) obtained a significantly higher average number of endorsed items for yes–no questions (4.74) than for check boxes (4.19) out of four sets of response options ranging from nine to 13 items. Nicolaas and colleagues (2011) found similar results from two randomised web experiments. The first experiment yielded an average of 4.4 endorsed items (out of 8) in the yes–no format, compared with an average of 2.9 endorsed items in the checkboxes format. The second experiment produced similar results: an average of 5.6 endorsed items (out of a possible 8) for the yes–no format condition compared to an average of 3.9 endorsed items for the checkboxes condition. These authors also conducted the same experiment on CAPI with very similar results.

Building on these results, Dykema *et al.* (2011) conducted a randomised experiment with university students and obtained a higher average number of endorsed items in the yes–no condition (8.3) compared with the check-all-that-apply condition (7.7). Finally, as part of a monthly experiment carried out on the Harris Interactive panel conducted for 25 months from 2006 to 2008, the employment section questions of the ACS were asked randomly in either a check-all-that-apply or in a yes–no format (Thomas *et al*. 2009). When asked about their full-time employment status, 52% reported being employed in the randomly assigned yes–no condition, compared with 46% in the check-all-that-apply condition.

Tsuchiya and Hirai (2010) used ‘applies’ and ‘does not apply’ instead of the classic yes–no wording. Their paper suggests that any binary response option, not just ‘yes’ and ‘no’ responses, still elicits higher endorsement rates than a checkbox. For eight lists of four items each, the forced-choice format enjoyed an average level of support 16.4% higher than the check-all-that-apply format. Smyth *et al*. (2005, 2006) also used variations to yes–no question wordings in other experimental conditions of their studies.

Finally, Callegaro *et al.* (2012, study 1) obtained similar results in an international web survey of small business advertisers. A total of 18,301 customers were asked about their usage of nine customer support channels either in a check-all-that-apply format (one-third of respondents) or a yes–no format (two-thirds of respondents). The average number of endorsed items was 2.67 for the yes–no condition versus 2.06 for the check-all-that-apply condition (a statistically significant difference). The same authors conducted a follow-up study asking customers whether they sought customer support of five different types, with about 1,200 respondents per experimental condition (study 2). Customers in the yes–no condition reported seeking an average of 1.85 types of support, compared with an average of 1.53 types for customers assigned to the check-all-that-apply condition (difference was statistically significant).

Mixed-mode experiments with telephone interviews

We identified three mixed-mode experiments comparing the results of a telephone survey with an identical web or mail survey, both with yes–no response options. In a mode experiment conducted in the 1993 National Survey of College Graduates (Mooney & Carlson 1996), respondents were allocated to either a telephone mode or to a mail mode. Six batteries of up to 14 yes–no questions per battery were spaced throughout the questionnaire. When comparing the results obtained by mail and telephone, four out of the six batteries did show statistically significant differences, with a higher average number of endorsed items for the telephone mode (the small effect size was statistically significant due to the very large sample size).[3] The authors explained the mode effect by noting that, in the CATI environment, interviewers read each item and make sure each answer is recorded, while in the mail self-administered mode the respondent might just 'stop marking responses at any point' (p. 617).

In the first study to be found in the literature, conducted both on the web and using CATI, undergraduates at Washington State University were randomised to either a telephone survey or a web survey (Smyth *et al.* 2008). Both surveys obtained a response rate of 59%. The authors found no difference in endorsement rates for yes–no options and concluded that yes–no response options were not prone to mode effects. On the other hand, Nicolaas *et al.* (2011) found a higher endorsement of yes–no options in a telephone version when compared with a web version of the same questionnaire. Respondents were randomly allocated to the two versions, but because of different response rates (69% by phone and 47% by web), the results were computed with controls for non-response in an OLS regression model.

In a telephone interview, *respondents* are typically presented questions only in a forced-choice format (e.g. yes–no), because it is unfeasible to ask them to check-all-that-apply in an auditory mode. However, the *interviewer* in a phone interview may be presented with the question battery on their CATI screen or script as either a forced-choice grid or as checkboxes, and there is some evidence that interviewers using checkboxes are less likely to read each item individually than interviews using yes–no options (Feindt *et al.* 1997).

Hypotheses explaining the question format treatment effects

The authors working on this topic tend to agree on the *theory of deeper cognitive processing* as an explanation for the main finding (i.e. more items endorsed in the forced-choice versus check-all-that-apply format). We present these explanations first, and then we propose an alternative and complementary theory that is not generally contemplated in the

[3] Sample sizes for the telephone group, depending on the question, varied from 585 to 3,193 respondents, while for the mail mode, they varied from 15,316 to 92,206 (Mooney & Carlson 1996, Table 1, p. 618).

literature to explain this particular treatment effect: *acquiescence response bias*. Finally, we comment on the exclusion of acquiescence response bias explanation by Smyth *et al*. (2006) as a potential explanatory factor.

Deeper cognitive processing

Sudman and Bradburn (1982) recognised an advantage of forced-choice answers over check-all-that-apply: they demand more thought. 'Respondents [in the forced-choice format] have to consider each adjective and decide whether it applies or does not apply to them' (p. 168). Later, Rasinski and colleagues (1994) hypothesised that the check-all-that-apply format 'might encourage a satisficing response strategy' (p. 403). Bradburn and colleagues (2004) proposed that forced-choice answers elicit *deeper cognitive processing* in asking respondents to consider each item singularly, thus possibly reducing satisficing strategies (Krosnick 1999). This theory has since been embraced by many other scholars (Smyth *et al*. 2006; Thomas & Klein 2006; Nicolaas *et al*. 2011).

Acquiescence response bias

We suggest an alternative and complementary theory that can explain the main findings: the well-known 'acquiescence response bias', i.e. 'the tendency for survey respondents to agree with statements regardless of their content' (Holbrook 2008, p. 3). Acquiescence response bias has been identified and discussed since the early days of survey research. For example, Schuman and Presser (1981) dedicated a chapter to this topic, called 'The acquiescence quagmire', in their questionnaire design book.

Most of the research on acquiescence bias has focused on agree/disagree items in experiments showing that respondents agreed, to a certain extent, with pairs of mutually exclusive statements. For example, Krosnick (1999, p. 552) reports a summary of 40 such agree/disagree studies where the average correlation among mutually exclusive (also called negatively worded) statements (e.g. *I enjoy socialising* versus *I don't enjoy socialising*) was only –0.22, when it should be about –1.0, thus providing strong evidence of acquiescence response bias. With regard to yes–no items, Krosnick and Presser (2010) concluded that the effects they found were very similar to agree/disagree items effects. For instance, when asked about factual questions (questions where there is a correct answer), respondents tend to answer them rightly more often when the correct response is 'yes' rather than 'no' (e.g. Larkins & Shaver 1967). For an example of acquiescence bias in market research, we refer the reader to Martin *et al.* (2011).

Acquiescence bias has been explained in three different ways (Krosnick 1999; Krosnick & Presser 2010): psychologists explain it as a *personality trait* for some respondents who are more predisposed than others to appear agreeable; sociologists explain it as the result of the *higher social status* of interviewers (representing the researchers) who, for some respondents, elicit the tendency to endorse assertions; lastly, acquiescence is explained by the notion of *satisficing* (Krosnick 1991). In the case of weak satisficing, it is easier to respond 'yes' because of the confirmatory bias hypothesis (Krosnick *et al.* 1996), which states that people start their memory search by looking first for reasons to agree, and search for reasons to disagree later in the response process. Agreeing is faster than disagreeing. In the case of strong satisficing, it is much easier to agree because it is considered polite – agreeing is more socially acceptable.

Acquiescence response bias ruled out by Smyth et al.*, and our counterargument*

Smyth *et al.* (2005, 2006) ruled out acquiescence response bias as a potential explanation for the finding of higher endorsement rates for questions in a forced-choice format compared with the check-all format. The authors had two more experimental conditions in addition to what has been described thus far. In these other treatment conditions, 'Don't know' and 'No opinion' were each added as a third available response option beside 'Yes' and 'No'.[4] The results from these experiments showed that when 'Don't know' was added, the difference in percentages in the distribution of yes–no items (the % who said yes in yes–no vs the % who said yes in yes–no–don't know) in both conditions was relatively unchanged in 10 of the 12 items. The results are much different under the 'No opinion' condition. Adding this category substantially changed the percentage of respondents endorsing 'No', lowering it by an average of 15.2 percentage points across the 15 items. Thus, the 'No opinion' option was drawing responses from the 'No' category. The authors use these two extra experiments to argue that if 'neutral or undecided respondents are acquiescing by choosing "yes" to avoid being disagreeable, we would expect to see these third categories drawing responses from the "yes" category' (p. 74), which is not the case.

[4] Tables 7 and 8 in Smyth *et al.* (2005, pp. 21–22).

We argue that this line of reasoning does not rule out acquiescence response bias as an explanation for higher endorsement in forced-choice questions. First, the experiment was conducted by comparing the condition with ‘Yes’ and ‘No’ response options with the condition, which added ‘No opinion’, in forced-choice format only. The parallel experiment where a ‘No opinion’ response option would appear next to the checkbox was not conducted. For this reason we cannot measure if the rate of ‘No opinion’ was higher (or lower) in the forced choice or check-all condition. Second, regardless of whether acquiescence bias is explained by weak or strong satisficing (Krosnick 1999), both a quicker search (in the respondent’s memory) for agreement criteria and greater social rewards for agreement predict that respondents will ‘Agree’ with statements regardless of their content. Neither predicts a higher rate of 'No opinion'.

Analyses of the randomised experiments

To deepen our knowledge of the topic, we performed summary analyses using the publications’ original tables or by obtaining the necessary data from the authors. Six analyses were performed, although the number of experiments per analysis differed depending on the availability of the data.

1. We calculated the average increase in item endorsement of the forced-choice condition over the check-all-that-apply condition.
2. For each format, we computed a rank order of answers within the battery of questions, from the answer that obtained the highest endorsement to the answer that obtained the lowest endorsement. We then analysed the correlation between the two rank orders as a measure of changes within the rank by format.
3. We computed how much longer it took respondents to answer forced-choice questions compared with questions in the check-all-that-apply format.
4. We studied primacy effects by question format.
5. We calculated the percentage of respondents who treated a forced-choice question as if it were asked in a check-all-that-apply format.
6. We examined break-off rates by condition.

To perform the first analysis, we first reduced the various tables in these papers[5] to common summary measures that could be directly

[5] We excluded the paper by Dyck and Moore (2008) due to the low sample size of their experiment.

compared. We captured the average percentage of endorsements in the check-all-that-apply tables and in the yes–no tables for each study. This resulted in 17 separate tables of averages. We treated each table as independent if the study was conducted with different subjects, and we combined different tables from the same study if they used the same subjects. We assumed independence of studies in the meta-analysis from a statistical point of view.[6] We performed a meta-analysis using OpenMeta (Wallace *et al.* 2012). A random effects meta-analysis (Borenstein *et al.* 2010) yielded a statistically significant point estimate of an odds ratio of 1.417 ($p < 0.001$) with a lower bound of 1.317 and an upper bound of 1.525. An odds ratio of 1 means no effect. This result shows that the forced-choice format increased endorsement rates substantially, with an overall average increase of a factor of 1.42.

The second analysis investigated whether the relative order (or rank) of an endorsed item changes depending on whether the question is asked in a check-all-that-apply or in a forced-choice format. For both the check-all-that-apply condition and the forced-choice condition, we assigned an order value to each item that captured the rank of the frequency of endorsement for that item – an approach similar to that used by Thomas and Klein (2006). We then computed the Pearson correlation between the two rankings, with a perfect correlation (1.0) indicating exactly the same order between the two conditions. The results from a meta-analysis of correlations using Comprehensive Meta Analysis (Borenstein *et al.* 2011) showed an overall cumulative correlation value of 0.97 with a random effect model ($p < 0.001$). Thomas and Klein (2006) reported a similar value ($r = 0.98$, $p < 0.001$) across all of their studies.

Next, in our third meta-analysis, we investigated how long it took respondents to answer both types of question format. To make the numbers comparable, we divided the reported average time per battery of questions by the number of items in the battery. On average, respondents took almost twice as long to answer question batteries in a yes–no format (192%) compared to the same questions in a check-all format. For example, a battery of ten items completed in 20 seconds (about two seconds per item) in the check-all-that-apply format would take about 38 seconds to complete in a forced-choice format. However, the standard deviation of the increase is high due to the varying topics and number of items used in the experiments, the population of interest, how each author treated outliers in the computation of the mean answer time, and the fact that

[6] Because we found only one study (Tsuchiya & Hirai 2010) where language departed from 'yes'/'no' language (using instead, 'applies to me'/'does not apply to me'), we excluded it from the analysis

forced-choice items were mandatory in the web survey (Thomas & Klein 2006). If we exclude the outliers (Thomas & Klein 2006; Dykema *et al.* 2011), the average increase is even greater: 264%. One reason for the longer period of time spent on forced-choice items is mechanical: they require the respondent to click either 'yes' or 'no' for each row, while check-all-that-apply grids demand a click on only a fraction of the items (Thomas & Klein 2006). But the size of this effect may also reflect some difference in psychological engagement as well.

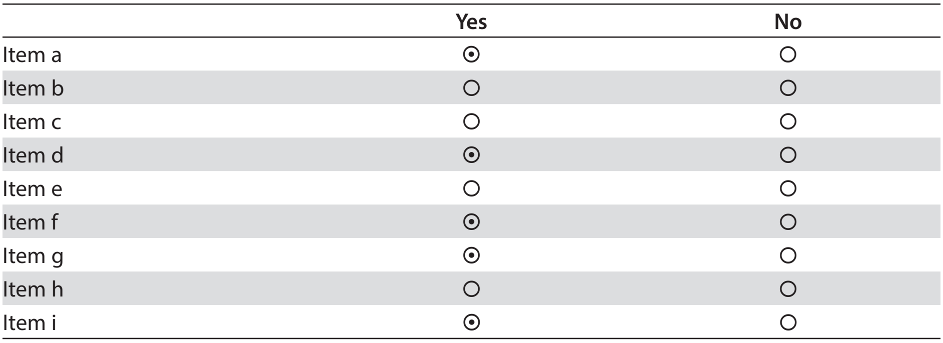
Another question we investigated is whether one of these two question formats is more prone to order or primacy effects than the other. For example, are items presented further down the list less likely to be endorsed for check-all versus forced-choice formats? In our fourth meta-analysis, we attempted to answer this question, and the results are mixed. In experiment 2, Thomas and Kline (2006) randomised the order of the list and kept track of the order of presentation of the items for a list of five and a list of ten items. They found that items presented later in the list are endorsed less, but there was no significant response format interaction. These results echo the paper-and-pencil experiment by Rasinski *et al.* (1994), where the order of the items was reversed for half the respondents. The authors found a significant order effect (items presented early were selected more often) but no interaction with the format of the question. However, Thomas and Klein (2006) in their 'experiment 4' (conducted in eight different countries with a list of five items)[7] found no order effect for check-all-that-apply, while they found increased endorsement for yes–no items presented later in the list. Lastly, Smyth *et al.* (2005) did not find any primacy effects for the check-all or yes–no question formats.[8] They did, however, find primacy effects for respondents who answered the survey in below-average time for check-all-that-apply in comparison to yes–no answers, which is an indication of satisficing. From the above evidence, it is difficult to conclude that one format is more prone to primacy effects than another.

A fifth question is if respondents treat a forced-choice response format as a check-all-that-apply format by marking only 'yes' responses without checking any 'no' responses. An example of such a pattern is depicted in Figure 1.

Patterns like this present the same challenges of interpretation that apply to the check-all-that-apply format. Although blank rows could be interpreted as item non-response, in practice when there are many

[7] Australia, Canada, France, Germany, Italy, Spain, UK and US.

[8] This is a longer version of the same study published a year later as Smyth *et al.* (2006), but contains more detailed analysis and tables.



	Yes	No
Item a	◉	○
Item b	○	○
Item c	○	○
Item d	◉	○
Item e	○	○
Item f	◉	○
Item g	◉	○
Item h	○	○
Item i	◉	○

Figure 1 Example of a respondent treating a yes–no set as check-all-that-apply

respondents answering in a similar fashion, it is difficult to make the assumption that all of the item non-response should be excluded from subsequent analysis. In a yes–no grid, non-response may simply be an indication of 'not yes'.

Two studies conducted with university students estimated the number of respondents using a yes–no grid as a check-all-that-apply format. In the first study, only 1% of respondents treated a yes–no matrix as a check-all-that-apply grid (Smyth *et al*. 2008), while in a second study across 24 yes–no treatments in three surveys, 2.7% of respondents treated a yes–no matrix as a check-all-that-apply grid (Smyth *et al*. 2006). However, the results from these papers were based on very salient survey topics conducted on college students with a high level of engagement, as demonstrated by high response rates of 56–59%. Callegaro *et al*. (2012) found much higher percentages of respondents treating a yes–no as a check-all-that-apply: 16.9% in study 1, with significant variation by country, and 3.2% in study 2. The results from study 1 might be due to the question wording used in the experiment.[9] If we exclude study 1 as an outlier due to the particular question wording, it appears that only a very small minority of respondents treat a yes–no matrix as a check-all-that-apply grid, although we advise readers to measure this effect in their datasets.

Finally, our sixth analysis revolved around the question of whether one response format produces a higher break-off rate than the other and, if so, how much higher? The limited evidence suggests that there are not dramatic differences. For example, in their study featuring four grids with a number of items ranging from four to seven, Dykema *et al*. (2011) found a slightly higher (but not statistically significant) break-off rate in the yes–no format (1.6%)

[9] 'Which, if any, of the following have you used for further information or help with AdWords in the last six months?'

rather than the check-all-that-apply format (0.7%). Callegaro *et al*. (2012), in two studies utilising one grid with eight items, found similar results: a 2.5% versus 1.5% break-off rate (significant at $p < 0.05$) in study 1, and a 0.6% vs 0.59% break-off rate (n.s.) in study 2. With only two studies investigating break-off in our inventory, there is limited evidence suggesting no difference in break-off rates between the two formats. Moreover, given that Peytchev (2009) showed that grids are associated with a higher level of break-offs than other question types, the reader might wonder if longer grids can trigger higher break-off rates than shorter grids. Unfortunately from the above data, we cannot validate this hypothesis.

Test–retest reliability and validity studies

Despite the substantial body of research on check-all-that-apply and yes–no formats, only a handful of studies have examined the reliability and validity of each approach. Specifically, two studies used re-interviews to compare the reliability of each method (Groves *et al*. 2009, pp. 282–284).

In the first study, about 1,000 respondents from the School and Staffing Survey (SASS) conducted in 1988 and 1991 in paper format were re-interviewed by phone by the US Census CATI laboratory (Feindt *et al*. 1997). Some questions had changed between waves, from a check-all-that-apply to a yes–no format. The authors computed an index of inconsistency (Forsman & Schreiner 1991) between the original interview and the re-interview data. The index of inconsistency for the re-interview questions in the check-all-that-apply format was greater than for the questions asked originally in the yes–no format, indicating that the yes–no format yielded more reliable responses.

Similar results were found in the above-mentioned ACS large-scale experiment (Ericson & Nelson 2007). For each of the original respondents who provided a phone number, a telephone re-interview was attempted, with a response rate of 76.1% across both the control and experimental groups. The yes–no format produced more consistent (implying more reliable) responses across the interviews, compared to the check-all-that-apply format, for questions on health status coverage.

The problem with these studies is that only the yes–no format in the telephone interview was cleanly compared with its paper version. Thus, there is a missing comparison: check-all-that-apply in the first interview with check-all-that-apply in the phone re-interview. This missing comparison weakens the researchers' claim that the yes–no format is more reliable than the check-all-that-apply response format.

In addition to reliability, researchers should also assess measurement validity in evaluating the quality of a response format (Rasinski *et al*. 1994; Smyth *et al*. 2008; Thomas & Klein 2006). In their camping study, Dyck and Moore (2008) obtained validation data for the camping shelter types via parking operators who observed and recorded the type of camping shelter each camping party used (e.g. tent or van). Results showed no statistically significant difference between the validation data and the questionnaire for either format, perhaps because of the high saliency of the questions. Dykema *et al.* (2011) also attempted to assess validity of the two question formats through an internal validity check. After asking respondents a series of 22 check-all or yes–no questions regarding media habits and social activities, they later asked in the same survey how often the respondents followed current events in the news and how often they have participated in social activities. Their hypothesis was that the 'better' question format should predict a stronger correlation with the internal criterion validity measure. In the end, their analysis showed that neither response format was superior in terms of criterion validity. This might be because the criterion validity question (how often) contained vague quantifiers, which may have produced measurement error that obfuscated the results. Finally, Callegaro *et al*. (2012, study 2) were able to match three of the five product help-seeking activities to internal data (for example, if a respondent said he/she chatted with a service representative during the past three months, that customer identification number was matched with the internal chat record). The results were inconclusive: the differences in false positive rates (when the respondent said he/she did the activity but no validations record was found) and false negative rates (when the respondent said he/she did not do the activity but a validation record was found) between the two conditions were not statistically significant.

The mixed evidence of studies conducted so far leaves the reader with some doubts about a potential higher test–retest reliability of forced choice versus check-all-that-apply, and no firm answer to the overarching question, 'Which response format elicits answers closer to the "true value"?' Reliability and validity should be measured together to give the full picture of the quality of a response format.

Discussion: two competing or complementary theories to explain the findings?

In this section, we evaluate the *deeper cognitive processing* and *acquiescence response bias* hypotheses. We will draw on the first four main findings from our systematic review to understand if one hypothesis better predicts our empirical results. Because the relevant studies for the subsequent three discussions are either inconclusive (primacy effects), relevant for only one format (respondents treating a forced choice as a check-all), or statistically (and substantively) insignificant (break-off rates), we do not consider them as evidence one way or the other.

Higher endorsement rates for forced-choice vs check-all

The deeper cognitive processing theory explains the finding by posing that more attention placed on the item makes it more likely information that can verify the statement will be retrieved from memory. The theory of acquiescence bias, which explains the same findings primarily for its mechanism, which is the tendency to agree with the statement regardless of content because it is easier to do so, or more socially accepted, explains the same findings. Based on the studies conducted so far, as well as our review and meta-analyses, we conclude that these two theories are not mutually exclusive, and indeed may both explain the higher endorsement rates for yes–no items. Deeper cognitive processing is the prevalent theory used by the authors of these studies. However, given that the magnitude of acquiescence bias is substantial, as summarised by Krosnick and Presser (2010), we are not convinced from the study of Smith *et al.* (2005, 2006) that acquiescence bias should be ruled out as an explanation of higher endorsement rates for yes–no question formats.

The rank order of the items does not change by question format

For this particular finding, neither theory seems to predict whether the order of items should change by questionnaire format. We may hypothesise that the effect should be constant throughout a battery of questions, regardless of explanatory theory. This may be why the rank order does not change across formats.

The time spent completing questions in the forced-choice format is about double that in the check-all format

Because answering yes–no questions requires a greater number of clicks in comparison to check-all-that-apply, it is difficult to disentangle these mechanics from other (cognitive and motivational) elements of answering the questionnaire. Proponents of the deeper processing theory (Smyth *et al.*

2006) propose that respondents spend more time 'on the forced-choice format independent of this extra mechanical response step' (p. 72). In fact, the respondents (college students) in their study spent two and a half times as long answering the forced-choice battery compared with the check-all format. Nicolaas *et al.* (2011) were able to measure time latency in their study with a sample of UK general population respondents. The treatment effect was even higher than in Smyth *et al.* (2006), about three times as long for the yes–no format than the check-all-that-apply format. An eye-tracking study, coupled with paradata, might be able to shed some light on how much of the treatment effect is caused by mechanical demands rather than deeper cognitive engagement.

In summary, our literature review found that effects are remarkably consistent across modes (paper and pencil or web), topics, populations and countries, and that both theories – deeper cognitive processing (Smyth *et al.* 2008) and acquiescence response bias (Krosnick & Presser 2010) – can account for these findings.

Conclusions and future research agenda

A change in response format can produce quite different estimates. What are the implications for survey measurement? The good news is that if researchers are interested in the relative order of the items, asking the question in one format or the other does not make a difference. On the other hand, researchers interested in accurate point estimates will be disappointed that choosing one format or the other will produce at times substantially different estimates, with the check-all-that-apply format always providing a lower estimate than the yes–no response format. Perhaps more frustrating is that the lack of conclusive validation studies makes it difficult to discern which method comes closer to hitting the mark.

We compared the current theory of deeper cognitive processing with the alternative theory of acquiescence response bias to explain differences across response formats. Both theories can account for the differing results between the two formats, and neither seems clearly superior to the other. One of the limitations of this study is that, while we identified 18 research studies, it is possible, as with any academic review, that we missed other experiments evaluating these formats given we included only English-language keywords in our search.

The overarching goal of this paper is to spur more research on these two kinds of question format, because they do elicit quite different estimates.

We hope the reader can use this paper as a starting point to test new hypotheses and the alternative theories that can account for some of the results and to compare the validity of yes–no and check-all-that-apply formats. Validation data can be difficult to obtain, difficult to process and expensive. At the same time, these data are invaluable if we are to know which response format provides answers closer to the 'true value'.

Appendix: Computation of rank order using Stern *et al.* (2007, p. 128) as an example

	Check-all	Forced-choice	Rank order check-all	Rank order forced-choice
Voted in the 2004 general election	83.5	84.1	1	1
Donated money to a community group	70.9	81.3	2	2
Signed a petition	65.7	70.4	3	3
Attended public hearings	37.8	41.5	4	4
Attended a public meeting	34.0	37.0	5	5
Participated in a strike	2.7	3.7	6	7
None of the above	1.8	6.3	7	6

Pearson's correlation between the rank order of forced-choice vs the rank order check-all: 0.964

Acknowledgements

We would like to thank the Political Psychology Research Group (PPRG) at Stanford for their valuable comments during a presentation of an early draft of this paper.

References

Ares, G., Barreiro, C., Deliza, R., Giménez, A. & Gámbaro, A. (2010) Application of a check-all-that-apply question to the development of chocolate milk desserts. *Journal of Sensory Studies*, **25**, s1, pp. 67–86.

Birnbaum, M.H. (2001) *Introduction to Behavioral Research on the Internet*. Upper Saddle River, NJ: Prentice Hall.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothstein, H.R. (2010) A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, **1**, 2, pp. 97–111.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothstein, H.R. (2011) *Comprehensive Meta Analysis*. Englewood, NJ: Biostat.

Bradburn, N.M., Sudman, S. & Wansink, B. (2004) *Asking Questions. The Definitive Guide to Questionnaire Design for Market Research, Political Polls, and Social and Health Questionnaires*, revised edn. San Francisco, CA: Jossey Bass.

Callegaro, M., Murakami, M.H., Henderson, V., Tepman, Z. & Dong, Q. (2012) Yes–no vs options in web surveys: what form is closer to benchmarks? Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Orlando, FL.

Dickersin, K. (2005) Publication bias: recognizing the problem, understanding its origins and scope, and preventing harm, in Rothstein, H.R., Sutton, A.J. & Borenstein, M. (eds) *Publication Bias in Meta-analysis. Prevention, Assessment and Adjustments*. Chichester: Wiley, pp. 11–33.

Dyck, B.W. & Moore, D.L. (2008) Validating a check-all and forced-choice question in a paper survey of provincial park campground users. Paper presented at the 63rd Annual Conference of the American Association for Public Opinion Research, New Orleans, LA.

Dykema, J., Schaeffer, N.C., Beach, J., Lein, V. & Day, B. (2011) Designing questions for web surveys: effects of check-list, check-all, and stand-alone response formats on survey reports and data quality. Paper presented at the 66th Annual Conference of the American Association for Public Opinion Research Phoenix, AZ. Available online at: http://uwsc.wisc.edu/2011fd/DykemaEtAl_2011_AAPOR.pdf (accessed 15 August 2014).

Ericson, L. & Nelson, C.T. (2007) A comparison of forced-choice and mark-all-that-apply formats for gathering information on health insurance in the 2006 American Community Survey Content Test. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA. Available online at: https://fcsm.sites.usa.gov/files/2014/05/2007FCSM_Ericson-VI-A.pdf (accessed 15 August 2014).

Feindt, P., Schreiner, I. & Bushery, J. (1997) Reinterview: a tool for survey quality improvement. *Proceedings of the Joint Statistical Meeting, Survey Research Methods Section*. Washington, DC: American Statistical Association, pp. 105–110.

Forsman, G. & Schreiner, I. (1991) The design and analysis of reinterview: an overview, in Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. & Sudman, S. (eds) *Measurement Errors in Surveys*. New York: Wiley, pp. 279–301.

Groves, R.M., Fowler, F.J. Jr, Couper, M.P., Lepkowski, J.M., Singer, V. & Tourangeau, R. (2009) *Survey Methodology*, 2nd edn. Hoboken, NJ: Wiley.

Holbrook, A. (2008) Acquiescence response bias, in Lavrakas, P.J. (ed.) *Encyclopedia of Survey Research Methods*. Vols. 1–2. Los Angeles, CA: Sage, Vol. 1, pp. 3–4.

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

Krosnick, J.A. (1999) Survey research. *Annual Review of Psychology*, **50**, 1, pp. 537–567.

Krosnick, J.A. & Presser, S. (2010) Question and questionnaire design, in Marsden, P.V. & Wright, J.D. (eds) *Handbook of Survey Research*, 2nd edn. Howard House, UK: Emerald, pp. 263–313.

Krosnick, J.A., Narayan, S. & Smith, W.R. (1996) Satisficing in surveys: initial evidence. *New Directions for Evaluation*, **70**, pp. 29–44.

Lado, J., Vicente, E., Manzzioni, A. & Ares, G. (2010) Application of a check-all-that-apply question for the evaluation of strawberry cultivars from a breeding program. *Journal of the Science of Food and Agriculture* **90**, 13, pp. 2268–2275.

Larkins, A.G. & Shaver, J.P. (1967) Matched-pair scoring technique used on a first-grade yes–no type economics achievement test. *Utah Academy of Science, Arts and Letters: Proceedings*, **44**, pp. 229–242.

Martin, W.C., Engelland, B.T. & Collier, J.E. (2011) Assessing the impact of acquiescence response bias on marketing data. *Marketing Management Journal*, **21**, 1, pp. 31–46.

Mitofsky, W.J. & Edelman, M. (1995) A review of the 1992 VRS exit poll, in Lavrakas, P.J. & Traugott, M.W. (eds) *Presidential Polls and the News Media*. Boulder, CO: Westview Press, pp. 81–100.

Mooney, G.M. & Carlson, B.L. (1996) Reducing mode effects in 'mark all that apply' questions. *Proceedings of the Joint Statistical Meeting, Survey Research Methods Section*. Washington, DC: American Statistical Association, pp. 614–619.

Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A. & Lynn, P. (2011) Is it a good idea to optimise question format for mode of data collection? Results from a mixed modes experiment. ISER working paper 2011–31. Available online at: www.iser.essex.ac.uk/publications/working-papers/iser/2011-31.pdf (accessed 15 August 2014).

Parente, M.E., Manzoni, A.V. & Ares, G. (2011) External preference mapping of commercial antiaging creams based on consumers' responses to a check-all-that-apply question. *Journal of Sensory Studies*, **26**, 2, pp. 158–166.

Peytchev, A. (2009) Survey breakoff. *Public Opinion Quarterly*, **73**, 1, pp. 74–97.

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

Schuman, H. & Presser, S. (1981) *Question & Answers in Attitude Surveys. Experiment of Question Form, Wording, and Context*. San Diego, CA: Academic Press.

Smyth, J.D., Christian, L.M. & 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, pp. 103–113.

Smyth, J.D., Dillman, D.A., Christian, L.M. & Stern, M.J. (2005) Comparing check-all and forced-choice question formats in web surveys. The role of satisficing, depth of processing, and acquiescence in explaining differences. Social and Economic Sciences Research Center, Pullman, Washington. Available online at: www.sesrc.wsu.edu/dillman/papers/2005/comparingcheckall.pdf (accessed 15 August 2014).

Smyth, J.D., Dillman, D.A., Christian, L.M. & Stern, M.J. (2006) Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, **70**, 1, pp. 66–77.

Stern, M.J., Dillman, D. & Smyth, J.D. (2007) Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods*, **1**, 3, pp. 121–138.

Stern, M.J., Smyth, J.D. & Mendez, J. (2012) The effects of item saliency and question design on measurement error in a self-administered survey. *Field Methods*, **24**, 1, pp. 3–27.

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

Thomas, R.K. (2011) Effects of response format on requalification for recontact studies. Paper presented at the 66th Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

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

Thomas, R.K., Dillman, D.A. & Smyth, J.D. (2009) Response format effects on measurement of employment. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Washington, DC.

Tsuchiya, T. & Hirai, Y. (2010) Elaborate item count questioning: why do people underreport in item count response? *Survey Research Methods*, **4**, 3, pp. 139–149.

Wallace, B., Dahabreh, I.J., Trikalinos, T., Lau, J., Trow, P. & Schmid, C.H. (2012) Closing the gap between methodologists and end-users: R as a computational back-end. *Journal of Statistical Software*, **49**, 2, pp. 1–15.

About the authors

Mario Callegaro is senior survey research scientist at Google, London, where he works on numerous survey projects in terms of questionnaire design, sampling, weighting and data collection methods. Mario holds an

MS and a PhD in Survey Research and Methodology from the University of Nebraska, Lincoln. His current research areas are web survey design, smartphone surveys, telephone/cell phone surveys and questionnaire design, on which he has published numerous papers, book chapters and conference presentations. Mario has just published an edited book on the quality of data obtained with online panels (Wiley 2014), and a monograph on web surveys methodology is due in 2015 (Sage).

Michael H. Murakami is the Senior Director of Research and Insights at Latinum Network, where his team conducts survey research, statistical analysis and qualitative research. Most recently, he was a quantitative marketing manager at Google, Inc., where he specialised in quantitative analysis, research design and survey methodology. Prior to Google, he was a professor of political science at the University of Illinois and Georgetown University. He also served as a Post-doctoral Fellow at Yale University's Center for the Study of American Politics, and an APSA Congressional Fellow in Washington, DC. He received his BA from Harvard University and his PhD from the University of California, Berkeley.

Ziv Tepman earned his BA in Economics and MS in Management Science and Engineering from Stanford University. He has published a book examining household spending patterns using survey data, and has presented at survey research conferences on the topic of response scales in customer satisfaction surveys. Ziv has launched surveys in multiple organisations, including Stanford University and Google. During his time at Google, Ziv conducted customer research studies on online advertisers through survey instruments and interviews, and continues to concentrate on market research projects.

Vani Henderson manages the Customer Research & Insights Team in the Global Customer Services organisation at Google. Her work focuses on leveraging survey research and Big Data analytics to deliver insights about customers. Prior to joining Google, Vani was a senior strategic planner at DraftFCB. She holds a PhD in communication from the Annenberg School at the University of Pennsylvania.

Address correspondence to: Mario Callegaro, Google UK Ltd, Belgrave House, 76 Buckingham Palace Road, London SW1W 9TQ, UK.

Email: callegaro@google.com