National Centre for Research Methods Working Paper

1/16

Fieldwork effort, response rate, and the distribution of survey outcomes: a multi-level meta-analysis

Joel Williams , Patrick Sturgis, Ian Brunton-Smith, Jamie Moore

NCRM National Centre for Research Methods

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

Fieldwork effort, response rate, and the distribution of survey outcomes: a multi-level meta-analysis

Joel Williams, TNS-BMRB

Patrick Sturgis, University of Southampton

Ian Brunton-Smith, University of Surrey

Jamie Moore, University of Southampton

ABSTRACT

We assess how survey outcome distributions change over repeated calls made to addresses in face-to-face household interview surveys. We consider this question for 559 survey variables, drawn from six major face-to-face UK surveys which have different sample designs, cover different topic areas, and achieve response rates which vary between 54% and 76%. Using a multi-level meta-analytic framework, we estimate for each survey variable, the expected difference between the point estimate for a proportion at call *n* and for the full achieved sample. We find that most variables are surprisingly close to the final achieved sample distribution after only one or two call attempts and before any post-stratification weighting has been applied. The mean expected difference from the final sample proportion across all 559 variables after 1 call is 1.6%, dropping to 0.7% after 3 calls, and to 0.4% after 5 calls. These estimates vary only marginally across the six surveys and the different types of questions examined. Our findings add further weight to the body of evidence which questions the strength of the relationship between response rate and nonresponse bias. In practical terms, our results suggest that making large numbers of calls at sampled addresses and converting 'soft' refusals into interviews are not cost-effective means of minimizing survey error.

INTRODUCTION

Random probability surveys face two key threats to their long-term viability: high and increasing costs, and low and declining response rates (Groves 2011; Peytchev 2013). These twin pressures, evident for at least the past two decades, have been felt all the more acutely in recent years as the rapid and pervasive emergence of opt-in internet panels has introduced a substantially quicker, more flexible, and most importantly, cheaper alternative to the survey marketplace (Callegaro et al. 2014). As response rates for random probability surveys edge ever lower, stakeholders increasingly question whether the substantial costs required to achieve high response rates can be justified. What, many survey commissioners ask, do they gain in data quality by choosing a low response rate random survey over a carefully designed quota sample? Or, indeed, by choosing a high over a low response rate design? Questions such as these have become more pressing as survey budgets face greater downward pressures throughout the world and as scholars and funders question the relevance of traditional survey methods in the age of linked administrative, transactional, and other forms of 'big data' (Couper 2013; Savage and Burrows 2007).

Costs and response rates are not, of course, independent of one another; as maintaining response rates at recent historical levels becomes ever more challenging, data collection agencies allocate larger shares of available resources to strategies which seek to ensure that contractual fieldwork targets are met (Curtin, Presser, and Singer 2000). A standard means of reaching response rate targets is to require interviewers to make repeated visits to addresses until contact is made and, hopefully, an interview is achieved. For face-to-face interview surveys this often involves making a minimum of six or more calls at different times of the day, and on different days of the week, before an address can be classified as a non-contact (Legleye et al. 2013). It also necessitates sending interviewers back to contacted but initially un-cooperative households, with the aim of persuading householders to undertake an interview, so-called 'refusal conversion'. It is now common for this group of 'soft-refusers' to be offered additional pecuniary incentives to participate at this stage of fieldwork, which raises the cost of achieving these final interviews still higher. Not only does this 'squeezing' of initially unproductive households raise ethical questions about the voluntary nature of research participation and the right to privacy (AAPOR

2014), there are also grounds for believing that gains in sample representativeness, if they exist, may be offset by the poorer quality data provided by the reluctant and unengaged respondents who may be recruited into the sample at this stage (Kreuter, Muller, and Trappmann 2010; Roberts, Allum, and Sturgis 2014).

In the UK context, and we suspect elsewhere, up to forty percent of total fieldwork costs can now be deployed in obtaining interviews from the twenty percent 'hardest-to-get' group of respondents. And, while call-back and refusal conversion strategies are generally effective in pushing the headline response rate up by sometimes substantial margins, it is less clear how successful they are in reducing nonresponse bias. Historically, it has been held that obtaining interviews with the 'hard-to-get' respondents is a key benefit of high quality (and therefore expensive) random household surveys. For, it is precisely these sorts of less accessible and less cooperative individuals who are likely to be under-represented in quota samples and in random samples with low response rates. There are reasons to believe, however, that gains in sample representativeness which accrue to obtaining interviews with 'hard-to-get' respondents may be less impressive than has often been assumed (Curtin, Presser, and Singer 2000).

This is for two primary reasons. First, the group of 'hard-to-get' respondents generally represents a comparatively small fraction of the total sample size (though how small depends, of course, on how 'hard-to-get' respondents are defined). Therefore, in order to have anything other than a trivial impact on a point estimate for the population, the 'hard-to-get' respondents must be substantially different on the survey outcome of interest, compared to the group of respondents who have already been interviewed. Second, and relatedly, there is growing evidence that the propensity to respond to survey invitations is only weakly correlated with many, even most, of the sorts of characteristics that are routinely measured in surveys (Groves 2006; Peytchev 2013). Put differently, the 'hard-to-get' respondents appear to be quite similar on many observable characteristics to the group of respondents who are cooperative after only a small number of call attempts. This points to an intriguing possibility;

that it might be possible to reduce survey costs significantly, without noticeably exacerbating nonresponse bias, simply by 'tolerating' lower response rates as the norm.

For this possibility to come to pass, commissioners, analysts, and other stakeholders in, and consumers of survey data, will need to have good grounds for believing that response rate genuinely is a poor indicator of nonresponse bias, so deeply ingrained has the belief become that survey estimates should not be trusted below some arbitrary threshold (Groves 2006). More and better evidence is therefore required on the question of how sensitive survey estimates are to variation in response rates and how this sensitivity varies across survey designs, modes, question types, and so on.

Our objective in this paper is to add to the body of evidence which seeks to address this key question. We analyze change in the distribution of survey outcomes over repeated calls for more than 550 variables taken from six major face-to-face surveys fielded in the UK between 2010 and 2014. The surveys have response rates between 54% and 76%, which was the typical range for high quality random surveys in the UK during this period, and cover a wide range of topic areas. For every variable, we compare the estimate at each call with the estimate for the full sample. We do this within a meta-analytic, multi-level modelling framework (Hox 2002), which enables us to account for differences in sample sizes across surveys and calls. It also allows modeling of difference estimates as a function of call number, survey, and question types, and to include interactions between these characteristics. The novel contribution of our approach is its comprehensiveness; while scholars have compared response distributions over call numbers and other indices of fieldwork effort, few have done so over such a large number of variables, across multiple surveys, and using a model-based approach.

The remainder of the paper is structured as follows. We first present a brief review of studies which have sought to elucidate the relationship between response rate and the distribution of survey outcomes. We then describe the data on which our analyses are based, including how different questions are coded, and detail the statistical model that we use to analyze them. We then present the key findings from our analyses, before concluding with a consideration of the implications of our findings, both for

understanding of the relationship between fieldwork effort and the distribution of survey outcomes, and for survey practice.

RESPONSE RATE AND SAMPLE REPRESENTATIVENESS

In principle, the research design best suited to addressing how response rates are related to the distribution of survey outcomes is one where an external, 'gold-standard' criterion is available for both respondents and nonrespondents. The criterion can then be compared to the variable measured in the survey to gauge nonresponse bias (and other forms of error) as the response rate to the survey increases. It is almost tautological to observe, however, that such designs are 'more honoured in the breach than in the observance'. For, not only do most of the characteristics that are routinely measured in surveys lack any external referent, if a well-measured criterion variable were available, there would be no good reason for attempting to measure it (less well) in a survey. Such studies are, therefore, rare and those that exist are unlikely to be representative of the range of individual and household characteristics that are measured in surveys. More commonly, investigators are able to compare respondents and nonrespondents on variables which are available on the sample frame, or which are linked to the sample frame from an external source. Other designs include those in which a follow-up survey is carried out amongst a sample of refusers, or where responses to screener questions can be compared between subsequent respondents and nonrespondents.

Merkle and Edelman (2002) used data from the 2000 US Presidential election exit poll to measure the association between voter-level (within-precinct) response rate to the exit poll and the magnitude of the difference in the Democratic and Republican vote shares compared to the poll estimate. Despite wide variability in response rates and the size of the bias across precincts, they were not able to reject the null hypothesis of zero correlation. Groves (2006) considers 319 bias estimates drawn from 30 studies and finds only a weak correlation (.33) between response rate and the magnitude of bias in point estimates. In a subsequent meta-analysis, which expanded on the set of studies included in the 2006 research, Groves and Peytcheva (2008) find no association between response rate and relative absolute nonresponse bias across 959 estimates from 59 different studies. In both these syntheses, the authors

found considerably more heterogeneity in bias estimates within than between studies, which indicates in itself that bias is more a function of questions and topic area than it is of studies and, therefore, response rates. The same general conclusion has been drawn in subsequent studies which have used composite 'representativity' indicators on a range of international data sets (Schouten, Cobben, and Bethlehem 2009); respondent samples differ from population values on many variables but these differences are only weakly related to the headline response rate. Sometimes a weakly negative relationship is observed, such that increases in response rate *reduce* sample representativeness (Bethlehem, Cobben, and Schouten 2011; Fuchs, Bossert, and Stukowski 2013).

The key strength of these studies – the ability to compare respondents and nonrespondents on the same variable – is also their primary limitation, which is to say that variables that are observed on nonrespondents are uncommon and unlikely to be representative of the universe of characteristics measured in surveys. The conclusions that can be drawn from these findings may, therefore, be limited to the quite narrow range of questions and topic areas which they are able to consider. It is also not uncommon for external variables to be measured using a different mode or procedure than the one used to undertake the survey, which means that these estimates of bias conflate nonresponse and measurement error. A different approach to addressing the same question involves comparing the distribution of survey outcomes at different levels of response rate. This can be done by examining response distributions across surveys which administer the same questionnaire at the same time to the same population, but which employ different fieldwork procedures and achieve different response rates. Or it can be implemented by 'simulating' different response rates within a single survey by truncating the sample according to indicators of fieldwork effort such as call number, or indicators of refusal conversion.

Using the former approach, Keeter et al (2000) compared estimates from an RDD survey conducted over 5 days and which achieved a response rate of 36 percent, to one conducted over 2 months with a response rate of 61 percent. Across 91 survey variables, covering a range of demographic, attitudinal, and behavioral topic areas, they found only 14 significant differences with an average discrepancy of

just 2 percentage points. Adopting the simulation strategy, Curtin, Presser, and Singer (2005) assessed the effect of excluding respondents who required higher numbers of calls and refusal conversion on the Survey of Consumer Attitudes between 1979 and 1996. They found that, although the 'hard to get' respondents differed significantly from the rest of the sample, excluding them had very little effect on survey estimates, despite their exclusion leading to a substantial drop in the response rate. Lynn and Clarke (2002) applied a similar procedure to three major UK face-to-face interview surveys and found that 'hard to get' respondents - defined in terms of being initially uncooperative and requiring a large number of calls – exhibited a number of significant differences from 'easy to get' respondents, with differences more pronounced on demographic than on attitudinal outcomes. However, while the two groups of respondents were clearly different on a broad range of characteristics, it is not clear from the analyses reported what effect, if any, the inclusion/exclusion of the 'hard to get' respondents had on the final survey estimates. A commensurate approach was adopted and similar conclusions drawn in the context of an RDD survey of sexual behaviors in France (Legleye et al. 2013).

In summary, the existing literature shows that response rate appears to have only a weak association with nonresponse bias in the limited range of contexts in which it has been possible to assess this relationship. A small but growing number of studies have also shown that, although 'hard to get' respondents do differ from more cooperative and easier to contact sample members, their exclusion from the final sample has little impact on the distribution of survey estimates. In this paper we extend this basic design by comparing response distributions for a large and diverse pool of survey variables as the number of calls to sampled addresses increases.

DATA

Our data are drawn from six random face-to-face interview surveys covering a range of topic areas and with differing sample designs and response rates (ranging from 54% to 76%). These are the British Crime Survey (2011); the British Election Study (2010); the Skills for Life Survey (2010-11); the Taking Part Survey (2011); the Community Life Survey (2013-14); and the National Survey for Wales (2013-14). Table 1 summarizes the key features of the six surveys. From each survey, we took all non-demographic variables that were asked of all respondents and excluded any questions administered only

to sub-groups on the basis of previous answers. We did not include demographic variables relating to age, sex, and working status because they were used to construct the post-stratification weights, so including them would have complicated the comparisons we make between weighted and un-weighted estimates. This resulted in a total of 559 survey questions across the six surveys.

TABLE 1 HERE

Each question was first transformed into a set of binary categorical variables (if it was not already a binary categorical variable in its initial state). We then calculated the absolute percentage difference (APD) between the proportion in each category at each call number and the proportion in the final achieved sample. For categorical items with k response options, we derived k-1 APD estimates, where the omitted category is the one with the lowest number of responses. So for example, with a binary response (k=2) coded as 0 or 1 we calculated:

1. The difference in the proportion scoring 1 after a single call and the final proportion scoring 1 after all calls
2. The difference between the proportion scoring 1 after 2 calls and the final proportion scoring 1 after all calls
3. The difference between the proportion scoring 1 after 3 calls and the final proportion scoring 1 after all calls
4. The difference between the proportion scoring 1 after 5 calls and the final proportion scoring 1 after all calls

Where respondents were able to select more than one response option, we treated each response option as a separate binary outcome.

These procedures resulted in a total of 1,250 estimates of the APD measured after 1, 2, 3 and 5 calls respectively. This produces a highly skewed distribution of, so we normalized by taking the natural log of the absolute differences. Transforming all variables into a set of proportions is somewhat arbitrary

and it would also be possible to make comparisons which preserve the 'natural' metric of the outcomes by, for instance, taking the absolute difference between means/medians, or by using differences in standardized variables. However, we believe that the use of absolute proportions has two attractive properties. First, all estimates are placed on the same metric, which renders a combined analysis feasible to implement. Second, proportions are more intuitively interpretable than are differences between means or between standardized variables. We compare results for design weighted and for post-stratified estimates, where the design weight is combined with a calibration weight and the calibration weight is recalculated for each call number.

Questions were classified according to response format and question type. We distinguish between questions according to whether their response format is: categorical, ordinal, binary, or multi-coded. We also coded questions according to whether they require respondents to report on behaviors, attitudes, or beliefs. We distinguish between attitudes and beliefs where the former are statements about the current, future, or past state of the world which could, in principle, have a right or wrong answer, while attitudes are a cognitive/affective evaluation of a stimulus object.

Multi-level meta-analysis

We use a meta-analytic framework to analyze the 1,250 APD estimates from all 559 questions. This approach enables us to produce a pooled estimate of the 'average' APD in the response distribution at each call number. The pooled estimate is a weighted average of the percentage difference for each question within each survey, which serves to smooth effects of questions which have particularly large (or small) differences at each call number. The meta-analysis is implemented within a multi-level modelling framework (Goldstein, 2011). This allows us to adjust for the dependency which is induced by including multiple estimates from the same question (for k-1 categories) across multiple call attempts. It also enables the inclusion of covariates which can be used to adjust for unobserved differences between surveys and to model differences as a function of call number and question characteristics. We specify a four level model, with the APD at each call number (level 2), within

response category (level 3), within question (level 4). Following Hox (2002), we also include a standard error for each of the APD estimates as the only variable at level 1[1] and the level 1 variance is constrained to unity because it is a known value. The model has the following form:

$$Log(\overline{x}_c - \bar{x}) = \beta_0 + \beta_1 Call2_{jkl} + \beta_2 Call3_{jkl} + \beta_3 Call5_{jkl} + \mathbf{x}'_{jkl}\boldsymbol{\beta} + w_l + v_{kl} + u_{jkl} + e_{ijkl} S.E_{jkl} \quad (1)$$

where $\bar{x}_c - \bar{x}$ is the APD at call c and at the final call. β_0 is the mean APD at call 1, with β_1 to β_3 indexing how the APD changes with each additional call. $\mathbf{x}_{ij}$ is a vector of covariates with coefficients $\boldsymbol{\beta}$. Covariates are included to indicate which survey the estimate is from, response format (binary, categorical, multi-coded categorical), and question type (behavioral, belief, attitudinal). The coefficients of the fixed effects show whether these different characteristics are associated with larger or smaller differences across all calls relative to the full sample; their interactions with call number test whether their magnitude changes as call attempts increase. w_l, v_{kl}, u_{jkl} are question-level, response category-level, and call number-level random effects, respectively. The random effects and residuals are assumed to be uncorrelated with one another and with the covariates, and to be normally distributed with zero means and constant variances, $w_l \sim N(0, \sigma_w^2)$, $v_{kl} \sim N(0, \sigma_v^2)$, and $u_{jkl} \sim N(0, \sigma_u^2)$.

RESULTS

Figure 1 shows the response rates for the six surveys at each call number. As we would expect, the response rates increase steadily from the first over subsequent calls, although the exact trajectory varies across surveys. The British Crime Survey has the highest response rate at every call, reflecting the generally high public interest in crime and disorder, the Skills for Life survey is the most burdensome of the six surveys, requiring respondents to undertake a range of cognitive tests. This means that the first contacts with a household often involve making appointments to return to undertake the long

[1] This is calculated using the following formula: $\mathrm{S.E} = \sigma^2 + \sigma_c^2 - \left(2 \cdot \sigma_c^2 \cdot \left({}^{n_c}/_{n}\right)\right)$, where σ^2 is the variance of the total sample point estimate, σ_c^2 is the variance of the call number point estimate at call c, n_c is the sample size at call number c, and n is the total sample size.

interview at a later date and this is reflected in the lower response rate for this surveys earlier in the call pattern.

FIGURE 1 HERE

Table 2 presents the coefficient estimates for the multi-level meta-analysis of the (logged) APD estimates. After only one call, we observe an average difference across all questions of 1.6% between the response distribution at that stage of fieldwork and the final achieved sample distribution. This average difference is surprisingly small, considering that the notional response rates to the six surveys vary between 7% and 22% at this stage. As the number of calls increases, the magnitude of the difference decreases quite markedly, falling to just 1% after 2 calls and to 0.4% after 5 calls. In addition to these fixed effects coefficients, it is also informative to consider the random effect estimates. Results from an 'empty' model containing no fixed covariates (not reported in table 2) reveal that number of calls accounts for just 25% of the overall variability in the AAD, with response category and question contributing 35% and 40% of the total variability, respectively. This echoes the findings of Groves and Peytcheva (2008), who also found that differences in response rates account for only a small fraction of the variability across nonresponse bias estimates, with the majority of the variance accounted for at the question level.

TABLE 2 HERE

Figure 2 shows the question level posterior Bayes estimates (Bryk and Raudenbush, 1992) with 95% credible intervals. Estimates are derived separately for each call number using the untransformed data. The absolute differences for more than half of the questions (54%) cannot be reliably distinguished from zero, even after the first call. By the 5th call attempt, this has risen to nearly 60% of all items. At each call attempt we do, however, note a small minority of items (on the far right of the graph) that have significantly higher absolute differences compared to the mean. This is particularly evident after 1 call, where the largest difference is 6.6% and three further questions have a difference larger than 4%.

These are all behavioral items, with two items asking about use of the internet (in the BCS and CLS), one asking how many hours people typically spend away from home during the day (BCS), and one asking about the number of visits made to the pub in the last month (BCS). Thus, while the average difference is remarkably close to the final distribution, even after 1 call, this masks considerable heterogeneity across variables, with some more affected by increases in response rates as the number of calls increases than others. However, even the item with the largest aPD of 6.6% after the first call, has been reduced to just 1.4% by the 5th call.

FIGURE 2 HERE

As should be expected, the size of these effects is reduced by calibration weighting, with a model estimated APD of 1.4% for post-stratified estimates after the first call. The size of the gap between the design weighted and the post-stratified difference estimates declines as the response rate increases, with an average difference of 0.8% by the second call, 0.5% by the third call and 0.3% by the 5th call. Looking at the posterior Bayes estimates at each call number (Figure 3), we still note a small number of items on the right hand side of the graph with higher than average APDs, however these are all markedly smaller than the unweighted estimates.

FIGURE 3 HERE

Model 2 includes covariates for the survey and question characteristics. No significant differences are evident between the BCS and NSW, though we do see significantly larger percentage differences for variables taken from the BES, SFL, TP, and the CLS. This is most apparent for the TP survey, where the APD is, on average, almost twice as large as the BCS. However, reflecting the small initial differences across survey items, this coefficient still only equates to an APD of 1.9% in the TP survey at call 1, compared to just under 1% for the BCS. It is likely that the survey fixed effects are reflective of topic content differences between the surveys, with some topic areas being more sensitive to variation in response rates and these topic areas being more prevalent in some surveys than others. We did attempt

to test this by coding questions according to their substantive topic area. However, there proved to be insufficient overlap of content areas over surveys to identify their independent effects. Model 2 also shows that differences tend to be larger for single coded items, with APDs up to 50% larger than they are for categorical items which ask respondents to provide multiple answers to the same question. As with the effects of surveys, it is difficult to say whether this effect is actually due to response type and format per se, or whether it is to do with a covariance between the topic area of the question and the sort of response format typically employed.

Modest differences are also evident between behavioral questions and questions which measure psychological variables, though this is not constant across call numbers/response rates. The APD for attitudinal items, and items measuring beliefs does not differ significantly from behavioral items (model 3). However, including the interaction between call number (model 4) and question type, we find larger differences for items measuring beliefs, after the first call attempt. However this difference is reduced by the 2nd call attempt and effectively disappears by call 3. Both attitudinal and belief questions also have somewhat larger APDs at the 5th call compared to behavioral items. However, the overall pattern for the covariates is one of moderate variability in what are already small differences.

DISCUSSION

As the confluence of downward pressures on citizens' availability and willingness to participate in surveys show no signs of abating, the cost of undertaking random probability surveys is reaching levels that many funders consider to be prohibitive. A substantial part of fieldwork costs is now expended on efforts to maintain response rates at levels which are considered, by historical standards, acceptable. These strategies are varied but primarily involve requiring interviewers to make repeated calls to sampled addresses until contact is made and to return to un-cooperative households to convert initial refusals into interviews. While such procedures are clearly effective as a means of increasing the headline response rate by what can sometimes be substantial amounts, their effect on the quality of survey estimates is considerably less well understood. In this study we have sought to go some way to addressing this lacuna by examining how the distributions of survey variables change, compared to the

final sample, as the number of calls (and therefore the response rate) increases. The approach we have adopted can be extended to include additional covariates covering characteristics of surveys and items such as topic area, mode, time, country, and other features of survey design. To encourage extensions of this nature, we are making the data set of effect estimates we have used, the code for model fitting, and a macro for generating effect size estimates from surveys available at this website (weblink here).

Our results show that over 1,250 effect sizes – defined as the absolute percentage difference - from 559 questions, drawn from six different surveys and covering a range of different topic areas and question types, the average APD for post-stratified estimates after the first call is just 1.4 percentage points. This falls to 0.8% after the second call, and to only 0.3% by the 5th call. An implication of these findings is that, had fieldwork ceased after the first call, the resulting estimates would have been very similar to those obtained after all calls, albeit with a significantly reduced sample size. Stopping after 5 calls would yield estimates which are practically indistinguishable from the full sample data. Of course, these figures average over a large number of different variables and there is a minority of these where the differences are more pronounced, particularly at the first few calls, when the response rate is low. Yet even for these variables, the estimates after the fifth call show, on average, less than a 1% difference compared to the final sample. Some question types are more sensitive to variability in response rate than others; questions eliciting respondents' beliefs show larger differences between the full sample and the sample after one call, relative to behavioral and attitudinal items. However, this effect dissipates quickly, with no statistically distinguishable differences by question type after only three calls.

These results have, we believe, at least two important implications for survey practice. The first is that the often substantial efforts expended by fieldwork agencies to attain arbitrarily high response rate targets do not appear to be cost-effective when assessed in terms of their effect on univariate response distributions. Potentially significant sums could be saved, or expended more effectively, if a limit were placed on the number of calls interviewers make at each address and if refusal conversions were curtailed or stopped entirely. The second is that response rates to these surveys are not strongly related to the distribution of survey outcomes. We cannot make the stronger claim that response rates are

weakly related to nonresponse *bias* because we do not observe the survey outcomes for the nonrespondents to the six surveys. This means that our estimates can only be taken as measures of nonresponse bias if we are prepared to assume that the final sample estimates are themselves free of nonresponse bias. While it is not uncommon for estimates taken from high response rate surveys to be used as criteria for bias assessment (Yeager et al 2011; Erens et al 2014), this is a strong assumption, albeit one that will always be necessary for psychological and attitudinal variables.

Moreover, our approach has some notable advantages, compared to studies for which observations are available for both respondents and nonrespondents. First, we consider a much larger and more diverse pool of items than is common in most studies of nonresponse bias, as we are able to include all survey variables, rather than only those that happen to be available for respondents and nonrespondents alike. It is likely that the kinds of variables which meet this latter criterion will be skewed toward particular topic areas and particular kinds of surveys, such as health surveys conducted via mail self-completion (Groves, 2006). For this reason, our findings and conclusions should, we believe, be more representative of the wide range of individual and social characteristics that tend to be measured in household surveys. Second, because all the comparisons we make are between variables measured within the same survey, these estimates do not confound nonresponse bias and measurement error, as is the case for studies which use external criterion variables that were administered in a different mode to assess nonresponse bias (Groves and Petcheyva).

Nonetheless, there are some limitations to the research design we have adopted here which should be acknowledged. The first relates to our use of APD as the measure of effect size. We acknowledged earlier that converting all survey variables into binary categories is somewhat arbitrary and different choices regarding where to place thresholds on continuous variables and so on would produce somewhat different results to those presented here. However, the arbitrary nature of this procedure must be pitted against the more intuitively interpretable metric that binary indicators provide compared to standardized variables (King 1986). We have repeated our analyses using the relative absolute difference by scaling the absolute difference to the final sample proportion. As should be expected,

using relative absolute values changes the shape of the distribution such that there are more 'extreme' values at each tail of the distribution. However, the substantive conclusions remain the same as for the absolute differences; large discrepancies are comparatively rare for the first one or two calls, with very few discernible differences evident after five calls.

Another limitation to the generality of our findings is that we have only considered univariate estimates for the general population. A consideration of these effects across different population sub-groups as well as for bivariate and multivariate distributions might reveal different patterns. Given the similarities in the marginal distributions at each call across these variables, we anticipate that the differences will be even less pronounced in the multivariate case, though additional analysis is required to assess this speculation. It is also important to be clear that we cannot assume that this pattern of estimates is what would be observed were a cap to be imposed on the number of calls that interviewers are permitted to make to each address. The implementation of a cap on calls might induce interviewers to adopt different strategies and approaches, for example devoting maximum effort to the 'easiest to interview households' and this might lead to larger (or smaller) average differences than we have shown here. It is also the case, of course, that a capped design would yield a smaller achieved sample size than an uncapped design so would generally require a larger issued sample in order to achieve the desired level of precision for key estimates. This is, itself, likely to have implications for the cost-effectiveness and practicality of fieldwork operations. For example, it might require a larger number of interviewers per survey, which might also affect the quality of the survey, if the additional interviewers were less experienced or skilled. We therefore urge caution in interpreting these results as indicating that capped designs will produce straightforward reductions in the cost of fieldwork, without unduly affecting the quality of survey estimates. Despite these caveats, we believe that our findings and the methodological approach we have adopted add further weight to the body of evidence which questions simplistic assumptions about the relationship between response rates and the distribution of survey outcomes.

References

AAPOR. 2014. *Current Knowledge and Considerations Regarding Survey Refusals*

Bethlehem, JG, F Cobben, and B Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: Wiley.

Callegaro, Mario , Reginald P. Baker, Jelke Bethlehem, Anja S. Goritz, Jon A. Krosnick, and Paul J. Lavrakas. 2014. *Online Panel Research: A Data Quality Perspective*. Wiley.

Couper, Mick P. 2013. Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods* 7 (3): 145-156.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly* 64: 413-428.

———. 2005. Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly* 69 (1): 87-98.

Fuchs, Marek, Dayana Bossert, and Sabrina Stukowski. 2013. Response Rate and Nonresponse Bias - Impact of the Number of Contact Attempts on Data Quality in the European Social Survey. *Bulletin de Me´thodologie Sociologique* 117: 26-45.

Groves, Robert. 2006. Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly* 70 (5): 646-75.

———. 2011. Three Eras of Survey Research. *Public Opinion Quarterly* 75: 861-71.

Groves, Robert , and E. Peytcheva. 2008. The Impact of Nonresponse Rates on Nonresponse Bias. *Public Opinion Quarterly* 72 (2): 167-89.

Hox, J. 2002. *Multilevel Analysis.*

Techniques and applications. . Mahwah, N.J: Lawrence Erlbaum Associates. .

Keeter, Scott, Carolyn Miller, Andrew Kohu, Robert Groves, and Stanley Presser. 2000. Public Opinion Quarterly. *Consequences of Reducing Nonresponse in a Large National Telephone Survey* 64: 125-48.

Kreuter, Freuke , G Muller, and Mark Trappmann. 2010. Nonresponse and Measurement Error in Employment Research. Making Use of Administrative Data. *Public Opinion Quarterly* 74 (5): 880-906.

Legleye, Stephane, Geraldine Charrance, Nicolas Razafindratsima, Aline Bohet, Nathalie Bajos, and Caroline Moreau. 2013. Improving Survey Participation: Cost Effectiveness of Callbacks to

Refusals and Increased Call Attempts in a National Telephone Survey in France. *Public Opinion Quarterly* 77 (3).

Lynn, Peter, and Paul Clarke. 2002. Separating Refusal Bias and Non-contact Bias: Evidence from UK National Surveys. *Journal of the Royal Statistical Society. Series D (The Statistician)* 51 (3): 319-33.

Merkle, D., and M. Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, eds. R Groves, D A Dillman, J. Eltinge and R J A Little. New York: Wiley. 243-258.

Peytchev, Andy. 2013. Consquences of Survey Nonresponse. *Annals of the American Academy of Political and Social Science* 645 (1): 88-111.

Roberts, C., N. Allum, and P. Sturgis. 2014. "Non-response and measurement error in online panels based on probability samples - are efforts to recruit reluctant panelists worth it?" In *Online panel research: a data quality perspective*, ed. M. Baker Callegaro, R. Bethlehem, J. Göritz, A., Krosnick, J & Lavrakas, P.: Wiley.

Savage, Mike, and Roger Burrows. 2007. The coming crisis of empirical sociology. *Sociology* 41 (5): 885-899.

Schouten, Barry, Fannie Cobben, and Jelke Bethlehem. 2009. Indicators for the representativeness of survey response. *Survey Methodology* 35 (1): 101-13.

TABLE 1

	British Crime Survey	**Taking Part**	**British Election Study**	**Community Life**	**National Survey for Wales**	**Skills for Life**
Population	England & Wales 16+	England 16+	Great Britain 18+	England 16+	Wales 16+	England 16-65
Timing	2011	2011	2010	2013-14	2013-14	2010-11
Sample size	46,785	10,994	1,811	5,105	9,856	7,230
Response rate	76%	59%	54%	61%	70%	~57%
Incentives?	Stamps (U)	Stamps (U) +£5 (C)	£5-10 (C)	Stamps (U) +£5 (C)	None	£10 (C)

U=unconditional incentive, C= conditional incentive

Table 2 Parameter estimates for multi-level models fitted to logged absolute percentage differences between the final sample and call n

	B	S.E.	EXP(B)	B	S.E.	EXP(B)	B	S.E.	EXP(B)	B	S.E.	EXP(B)
Constant	0.452*	0.035	1.571	-0.006	0.096	0.994	-0.045	0.100	0.956	-0.048	0.100	0.953
Call number (ref: 1 call)												
Up to 2 calls	-0.441*	0.016	0.643	-0.441*	0.016	0.643	-0.441*	0.016	0.643	-0.412*	0.018	0.662
Up to 3 calls	-0.820*	0.017	0.440	-0.820*	0.017	0.440	-0.819*	0.017	0.441	-0.804*	0.019	0.448
Up to 5 calls	-1.443*	0.018	0.236	-1.443*	0.018	0.236	-1.443*	0.018	0.236	-1.476*	0.021	0.229
Questionnaire (ref: BCS)												
BES				0.512*	0.123	1.669	0.503*	0.126	1.654	0.504*	0.127	1.655
SFL				0.337*	0.108	1.401	0.363*	0.109	1.438	0.366*	0.109	1.442
TP				0.643*	0.127	1.902	0.654*	0.128	1.923	0.652*	0.128	1.919
CLS				0.260*	0.099	1.297	0.273*	0.100	1.314	0.272*	0.100	1.313
NSW				0.130	0.111	1.139	0.124	0.112	1.132	0.124	0.112	1.132
Variable type (ref: Multi-coded categorical)												
Single coded categorical				0.218*	0.088	1.244	0.209*	0.089	1.232	0.209*	0.089	1.232
Single coded ordinal				0.399*	0.094	1.490	0.359*	0.099	1.432	0.353*	0.099	1.423
Single coded binary				0.313*	0.158	1.368	0.345*	0.160	1.412	0.341*	0.160	1.406
Question type (ref: Behaviour)												
Attitude							0.072	0.081	1.075	0.034	0.084	1.035
*Up to 2 calls										-0.014	0.038	0.986
*Up to 3 calls										0.047	0.039	1.048
*Up to 5 calls										0.149*	0.042	1.161
Belief							0.166	0.115	1.181	0.251*	0.118	1.285
*Up to 2 calls										-0.149*	0.047	0.862
*Up to 3 calls										-0.178*	0.050	0.837
*Up to 5 calls										-0.028	0.054	0.972
RANDOM EFFECTS												
Question	0.313	0.038		0.286	0.035		0.287	0.035		0.290	0.036	
Response category	0.331	0.026		0.324	0.026		0.324	0.026		0.322	0.026	
Difference at time t	0.002	0.001		0.002	0.001		0.002	0.001		0.000	0.001	
SE	1	0		1	0		1	0		1	0	
-2*loglikelihood:	13946			13892			13890			13855		
Questions	559			559			559			559		
Effects	1250			1250			1250			1250		
Effects*calls	5000			5000			5000			5000		

*=p<0.05; B=logit coefficient; S.E. = standard error; EXP(B)=odds ratio.

FIGURE 1 RESPONSE RATE PER CALL NUMBER FOR ALL SIX SURVEYS

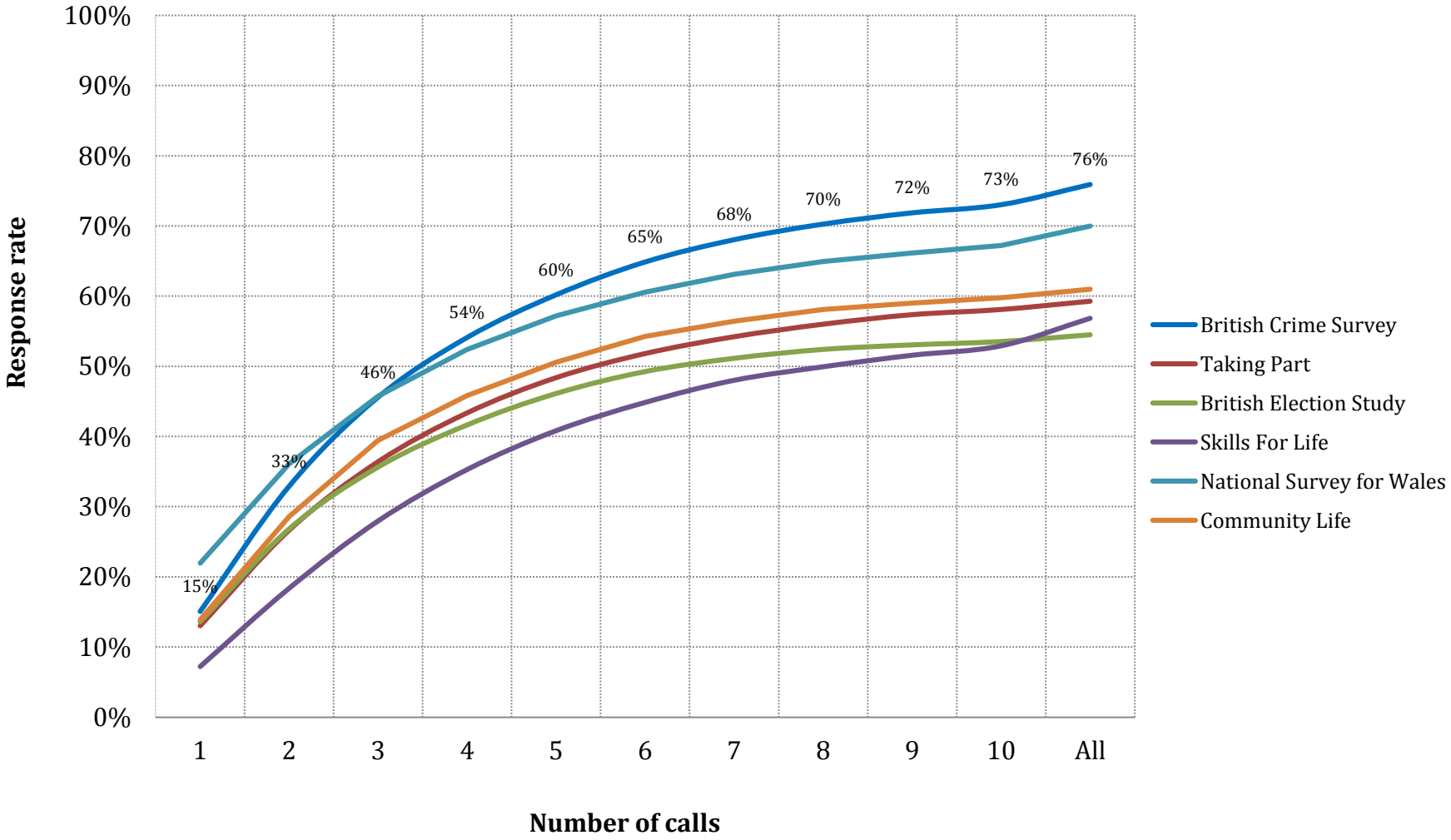

FIGURE 2 ESTIMATED ABSOLUTE PERCENTAGE DIFFERENCE BY QUESTION (DESIGN WEIGHTED)

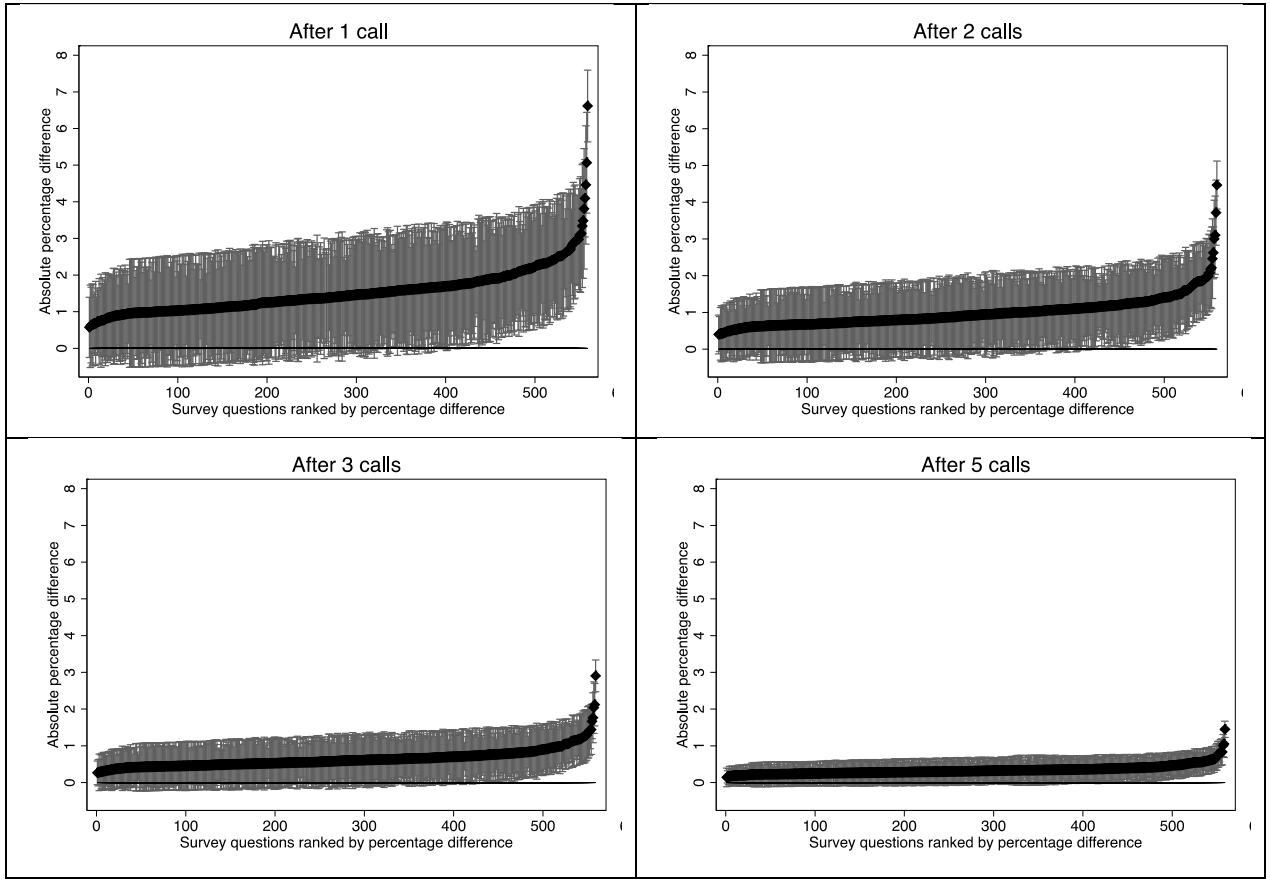

FIGURE 3 ESTIMATED ABSOLUTE PERCENTAGE DIFFERENCE BY QUESTION (CALIBRATION WEIGHTED)