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

Patrick Sturgis, University of Southampton, p.sturgis@soton.ac.uk

Joel Williams, TNS-BMRB, Joel.Williams@tns-bmrb.co.uk

Ian Brunton-Smith, University of Warwick, i.r.brunton-smith@surrey.ac.uk

Jamie Moore, University of Southampton, j.c.moore@soton.ac.uk

Running header: fieldwork effort, response rate, bias

Word count: 8617

Author note

Patrick Sturgis is Professor in the Department of Social Statistics and Demography, University of Southampton. Joel Williams is Director of Methodology with TNS UK, Ian Brunton-Smith is Professor in the Department of Sociology, University of Warwick. Jamie Moore is Senior Research Fellow with the Southampton Statistical Sciences Research Institute.

The authors gratefully acknowledge the support of the Economic and Social Research Council through the grant for the National Centre for Research Methods (NCRM; grant reference: ES/L008351/1)

Address correspondence to Patrick Sturgis, University of Southampton, Department of Social Statistics and Demography, Southampton SO17 1BJ United Kingdom; email: p.sturgis@soton.ac.uk

**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 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 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 pressure throughout the world and as scholars and funders question the relevance of traditional survey methods in an 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; 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 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[1]. 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, a substantial proportion of total fieldwork costs can be deployed in obtaining interviews from the minority group of 'hardest-to-get' respondents. And, while call-back and refusal conversion strategies are generally effective in pushing the response rate up by sometimes substantial margins, how successful they are in reducing nonresponse bias is less clear. 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 including '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' 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

[1] In the UK, surveys including the British Social Attitudes survey, the British Election Study, the Wellcome Trust Monitor, and Understanding Society now use this approach.

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 modeling 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 contain variables that are 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 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 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 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 common 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 refusal conversion.

Using the former approach, Keeter et al (2000) compared estimates from an RDD survey conducted over 5 days 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 topics, 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 '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 survey estimates. In this paper we extend the simulation 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 these 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 541 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. Additional calls at an address can be made for a number of different reasons, the most common being that no contact is made with a household member or eligible respondent but it can also happen because an appointment has been made at a previous call, or because a refusal conversion is being attempted. Unfortunately, the coding of the call record data across surveys does not allow us to account for these different reasons for additional calls in our analysis. The APD treats all percentage differences as equivalent, irrespective of the size of the discrepancy relative to the final sample estimate[2]. We therefore also calculated the relative percentage difference (RPD) by scaling the absolute difference to the final sample proportion. Although differing in some respects, the basic pattern of results is substantively the same using either the APD or the RPD. To conserve space, we therefore present the results for the RPD in the appendix. For categorical items with k response

[2] For example, a difference between 10% at call 1 and 5% after all calls has the same APD as a difference between 55% at call 1 and 50% after all calls, despite the first discrepancy being much larger in relative terms than the second.

options, we derived k-1 APD/RPD estimates, where the omitted category is the one with the lowest frequency. 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 option as a separate binary outcome. These procedures resulted in a total of 1,243 estimates of the APD/RPD measured after 1, 2, 3 and 5 calls respectively[3]. This produces highly skewed distributions, so we normalized them by taking the natural log. 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 differences in proportions have two attractive properties. First, all estimates are placed on the same metric, which renders a combined analysis feasible to implement. Second, differences in proportions are more intuitively interpretable than are differences between means or between standardized variables. We also 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 at each call number. The calibration weights use Office for National Statistics mid-year population totals for age, sex, and region plus contemporary Labour Force Survey data on highest qualification, working status, and ethnic group.

[3] We include these estimates in the online appendix for those who wish to replicate or extend our analysis.

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 behaviors, attitudes, or beliefs. We distinguish between beliefs and attitudes 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 1,243 APD/RPD estimates from all 541 questions. This approach enables us to produce a pooled estimate of the 'average' difference 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 that have particularly large (or small) differences at each call number. The meta-analysis is implemented within a multi-level modeling framework (Goldstein, 2011) which adjusts for the dependency 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 three level model, with the APD/RPD at each call number (level 1), within response category (level 2), within question (level 3).

We include dummy indicators of the surveys as fixed effects at level 3 of the model. This controls for differences between surveys that might also affect the magnitude of the APD/RPD, such as target population, survey topic, level of incentive, and so on. An alternative specification would be to treat differences between surveys as a random effect at level 4. This would have the advantage that observed differences between surveys could be used as predictors in the model rather than being conditioned out. However, the number of surveys is too small to make a random effect specification at the survey level feasible (Bryan and Jenkins, 2015).

Following the procedure described in Hox (2010: 208), we incorporate the sampling error of each estimate by including the sampling variances of the APD/RPD at level 1. This accounts for the fact that estimates of APD/RPD are based on different sample sizes across surveys and calls and should not, therefore, be given an equal weight in the combined estimates. The sampling distribution of the estimates is assumed to be normal with known variance (Hedges and Olkin, 1985).[4] Let $\hat{\pi}_{jkl}$ denote the estimated proportion for response category $k$ of question $l$ at call $j$, and $\hat{\pi}_{Ckl}$ the estimated proportion from the final achieved sample. Define a response variable $Y_{jkl}$ as the logged APD/RPD $Y_{jkl} = log\left|\hat{\pi}_{jkl} - \hat{\pi}_{Ckl}\right|$ or logged RPD $Y_{jkl} = log\left(\left|\hat{\pi}_{jkl} - \hat{\pi}_{Ckl}\right|/\hat{\pi}_{Ckl}\right)$. We use the multilevel model:

$$Y_{jkl} = \beta_0 + \beta_1 Call2_{jkl} + \beta_2 Call3_{jkl} + \beta_3 Call5_{jkl} + \mathbf{x}'_{jkl}\beta_4 + w_l + v_{kl} + u_{jkl} + e_{jkl} \quad (1)$$

where $\beta_0$ is the mean APD/RPD at call 1, with $\beta_1$ to $\beta_3$ indexing how the APD/RPD changes with each additional call. $\mathbf{x}_{jkl}$ is a vector of covariates with coefficients $\beta_4$. The covariates include the survey fixed effects, question response format (binary, categorical, multi-coded categorical), question type (behavioral, belief, attitudinal), and interactions between the call number dummies and other covariates. The coefficients of the fixed effects indicate whether these characteristics are associated with larger or smaller differences across all calls relative to the full sample; their interactions with call number denote whether their magnitude changes as call attempts increase. The quantities $w_l \sim N(0, \sigma_w^2)$, $v_{kl} \sim N(0, \sigma_v^2)$, $u_{jkl} \sim N(0, \sigma_u^2)$ and $e_{jkl} \sim N\left(0, \hat{o}_{jkl}^2\right)$ are normally distributed random

[4] To calculate the sampling variances we use the following procedures. For, logged APD, denote $\hat{\delta} = \hat{\pi}_{jkl} - \hat{\pi}_{Ckl}$, with estimated variance $\hat{\psi} = v\hat{a}r(\hat{\pi}_{Ckl}) + v\hat{a}r\left(\hat{\pi}_{jkl}\right) - 2c\hat{o}v\left(\hat{\pi}_{Ckl}, \hat{\pi}_{jkl}\right)$. Assuming $\hat{\delta}$ to be approximately normal, its absolute value follows the "folded normal" distribution, with estimated variance $v\hat{a}r\left(\left|\hat{\delta}\right|\right) = \hat{\psi} + \hat{\delta}^2 - \mu_{|\hat{\delta}|}^2$, where $\mu_{|\hat{\delta}|} = \sqrt{\frac{2}{\pi}}\hat{\psi}^{1/2}e^{-\hat{\delta}^2/2\hat{\psi}} + \hat{\delta}\left[1 - 2\Phi\left(-\frac{\hat{\delta}}{\hat{\psi}^{1/2}}\right)\right]$ and $\Phi$ denotes the cumulative distribution function of the standard normal distribution. Using the delta method, the estimated variance of $log\left|\hat{\delta}\right|$ is then $v\hat{a}r\left(log\left|\hat{\delta}\right|\right) = \frac{v\hat{a}r(|\hat{\delta}|)}{\hat{\delta}^2}$. For logged RPD, denote $\hat{\delta} = \frac{\hat{\pi}_{jkl} - \hat{\pi}_{Ckl}}{\hat{\pi}_{Ckl}}$, with estimated variance $\hat{\psi} = \frac{\hat{\pi}_{jkl}^2}{\hat{\pi}_{Ckl}^2}\left[\frac{v\hat{a}r(\hat{\pi}_{jkl})}{\hat{\pi}_{jkl}^2} + \frac{v\hat{a}r(\hat{\pi}_{Ckl})}{\hat{\pi}_{Ckl}^2} - \frac{2c\hat{o}v(\hat{\pi}_{Ckl}, \hat{\pi}_{jkl})}{\hat{\pi}_{jkl}\hat{\pi}_{Ckl}}\right]$. The remaining steps are then identical to the logged APD because $\left|\hat{\delta}\right| = \left|\frac{\hat{\pi}_{jkl} - \hat{\pi}_{Ckl}}{\hat{\pi}_{Ckl}}\right| = \frac{\left|\hat{\pi}_{jkl} - \hat{\pi}_{Ckl}\right|}{\hat{\pi}_{Ckl}}$.

effects which are assumed to be uncorrelated with one another and with the covariates. The random effect variances $\sigma_w^2$, $\sigma_v^2$, and $\sigma_u^2$ are estimable parameters and $\hat{\sigma}_{jkl}^2$ are the estimated sampling variances of $Y_{jkl}$ which are taken as known quantities in (1).

**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 differs 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 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 (the equivalent estimates for the RPD are included in appendix table A1). After only one call, there is an average difference across all questions of 1.6 percentage points between the response distribution at that stage of fieldwork and the final achieved sample distribution (model 2). This average difference is surprisingly small, considering the notional response rates to the six surveys vary between 7% and 22% at this stage. As the number of calls (and hence the response rate) increases, the magnitude of the difference decreases markedly, falling to just 1 percentage point after 2 calls and to 0.4 percentage points after 5 calls.[5] A similar reduction is evident when considering the RPD, with an average 3.9 point difference at call 1, falling to 2.5% after 2 calls, 1.8% after 3 calls and 1% by the fifth call.

[5] Calculated as $e^{(0.41-0.44)}$ at call 2, $e^{(0.41-0.82)}$ at call 3, and $e^{(0.41-1.43)}$ at call 5.

Turning to the random effect estimates, an 'empty' model containing no covariates (model 1) shows that the number of calls accounts for just 25% of the total variability in the APD, with response category and question type contributing 35% and 40%, 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 estimated question level APDs with 95% confidence intervals.[6] The absolute differences for half of the questions (50%) cannot be reliably distinguished from zero after the first call. By the 5th call attempt, this has dropped to 45% of all items, reflecting the fact that although the difference estimates are smaller, the sample size is larger and the precision of the estimates greater. 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 percentage points and three further questions have a difference larger than 4 percentage points. These are all behavioral items, two of which ask about use of the internet (in the BCS and CLS), one which asks how many hours people typically spend away from home during the day (BCS), and one which asks 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 than others. However, even the item with the largest APD of 6.6 percentage points after the first call, has been reduced to just 1.4 percentage points by the 5th call. Using the RPD there are more 'extreme' values at each tail of the distribution, with the largest being 29% at call 1 (appendix figure A1). However, like the APDs, large discrepancies are rare and very few discernible differences are evident after five calls.

[6] The estimated question level APDs are calculated by summing the question level shrunken residuals and the mean APD at each call number. Estimates are derived separately for each call number using the untransformed data.

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 percentage points for post-stratified estimates after the first call (table A2). 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 percentage points by the second call, 0.5 points by the third call and 0.3 points by the 5$^{th}$ call. The same is true of the RPD, with a difference of 3.2% at call 1, falling to 0.9% by call 5 (table A3). Considering the estimated question level APDs at each call number (Figure 3), we still note a small number of items with higher than average APDs, however these are all markedly smaller than the unweighted estimates (the equivalent estimated question level RPDs are included in figure A2).

FIGURE 3 HERE

Model 3 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 only equates to an APD of 1.9 percentage points in the TP survey at call 1, compared to just under 1 percentage point 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 3 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. A similar picture is evident when considering the RPD estimates.[7]

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 4). However, including the interaction between call number (model 5) 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 and effectively disappears by call 3. Both attitudinal and belief questions also have somewhat larger APDs at the 5th call compared to behavioral items, though 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 to interviews. While such procedures are clearly effective as a means of increasing the 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

[7] The larger RPD for single-coded ordinal and single coded categorical items reflects the use of k-1 response categories in our model. Single coded binary variables and multi-coded categorical variables have k=2, therefore the removal of the lowest response category results in all proportions being >0.5. This tends toward lower RPDs than single coded ordinal and single coded categorical items, where proportions range from 0-1. APD estimates are not affected by this.

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 facilitate extensions of this nature, we are making the data set of effect size estimates and the code for model fitting available in the online appendix.

Our results show that over 1,243 effect sizes from 541 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 percentage points after the second call, and to only 0.3 percentage points by the 5$^{th}$ 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 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, a less than 1 percentage point 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 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. 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, strictly, 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. In any event, even without this strong assumption, our estimates are upper bounds on the bias reduction that can be achieved by additional calls.

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 be more representative of the wide range of individual and social characteristics that tend to be measured in household surveys. Second, because all 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 Peytcheva 2008).

Nonetheless, there are some limitations to the research design we have adopted here which should be acknowledged. The first relates to our measures of effect size. We noted earlier that converting all survey variables into binary categories is somewhat arbitrary and different procedures could be employed which would likely 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 used relative as well as absolute measures of effect size across call number by scaling the absolute difference to the

final sample proportion. Absolute measures are, we believe, more intuitively interpretable but tend to under-emphasize the size of differences at the lower and upper bounds. As should be expected, using a relative measure of difference changes the shape of the distribution such that there are more 'extreme' values in each tail. However, the substantive conclusions remain the same as for the absolute differences; large discrepancies are rare after three calls, with very few discernible differences evident after that.

Other limitations to the generality of our findings are that we have only considered face-to-face personal interview surveys and, while there is no strong reason to assume that the pattern of results would be very different for other modes, extrapolation of these findings to self-completion and telephone interview surveys should be done with caution. We have also aggregated the various different reasons for repeated calls at an address – non-contact, initial refusal, appointments, and so on - into a single category. Future research could usefully evaluate whether the patterns we have observed here vary according to the different reasons that additional calls are made. It is also the case that consideration of these effects for different population sub-groups might reveal a different pattern. Given the similarities in the marginal distributions at each call across the variables considered here, we anticipate that differences would actually be less pronounced in the bivariate and multivariate cases, though additional analysis is required to assess this.

We should also 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, our findings 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.

Bryan, Mark L. and Jenkins, Stephen P. (2015) Multilevel modelling of country effects: a cautionary tale *European Sociological Review*. ISSN 0266-7215

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.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-analysis*. San Diego, CA: Academic Press.

Hox, J. 2010. *Multilevel Analysis. Techniques and Applications.* (2nd Ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

Hunter, J. E., and F. L. Schmidt. 2004. *Methods of Meta-analysis* (2nd Ed.). Newbury Park, CA: Sage.

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.

King Gary, 1986. "How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science." American Journal of Political Science, 30: 666–687, August.

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 SURVEY DESIGNS AND CHARACTERSTICS

| | **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 MULTILEVEL MODELS, LOGGED ABSOLUTE PERCENTAGE DIFFERENCE, FINAL SAMPLE - CALL C (DESIGN WEIGHTED)

| | Model 1 | | | Model 2 | | | Model 3 | | | Model 4 | | | Model 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** |
| Constant | -0.27* | 0.03 | 0.76 | 0.41* | 0.04 | 1.51 | -0.03 | 0.09 | 0.97 | -0.08 | 0.10 | 0.92 | -0.08 | 0.10 | 0.92 |
| Call number (ref: 1 call) | | | | | | | | | | | | | | | |
| Up to 2 calls | | | | -0.44* | 0.02 | 0.64 | -0.44* | 0.02 | 0.64 | -0.44* | 0.02 | 0.64 | -0.41* | 0.02 | 0.66 |
| Up to 3 calls | | | | -0.82* | 0.02 | 0.44 | -0.82* | 0.02 | 0.44 | -0.82* | 0.02 | 0.44 | -0.81* | 0.02 | 0.45 |
| Up to 5 calls | | | | -1.43* | 0.02 | 0.24 | -1.43* | 0.02 | 0.24 | -1.43* | 0.02 | 0.24 | -1.48* | 0.02 | 0.23 |
| Questionnaire (ref: BCS) | | | | | | | | | | | | | | | |
| BES | | | | | | | 0.47* | 0.12 | 1.60 | 0.49* | 0.12 | 1.63 | 0.49* | 0.12 | 1.64 |
| SFL | | | | | | | 0.53* | 0.11 | 1.70 | 0.58* | 0.12 | 1.79 | 0.58* | 0.12 | 1.79 |
| TP | | | | | | | 0.59* | 0.13 | 1.80 | 0.60* | 0.13 | 1.83 | 0.60* | 0.13 | 1.83 |
| CLS | | | | | | | 0.26* | 0.09 | 1.30 | 0.29* | 0.09 | 1.33 | 0.29* | 0.09 | 1.33 |
| NSW | | | | | | | 0.00 | 0.10 | 1.00 | 0.02 | 0.10 | 1.02 | 0.02 | 0.10 | 1.02 |
| Variable (ref: Multi-coded) | | | | | | | | | | | | | | | |
| Single coded categorical | | | | | | | 0.20 | 0.11 | 1.23 | 0.21 | 0.11 | 1.24 | 0.21 | 0.11 | 1.23 |
| Single coded ordinal | | | | | | | 0.30* | 0.08 | 1.36 | 0.25* | 0.09 | 1.29 | 0.25* | 0.09 | 1.29 |
| Single coded binary | | | | | | | 0.38* | 0.13 | 1.46 | 0.42* | 0.13 | 1.51 | 0.41* | 0.13 | 1.51 |
| Question (ref: Behavior) | | | | | | | | | | | | | | | |
| Attitude | | | | | | | | | | 0.06 | 0.09 | 1.06 | 0.01 | 0.09 | 1.01 |
| *Up to 2 calls | | | | | | | | | | | | | -0.01 | 0.04 | 0.99 |
| *Up to 3 calls | | | | | | | | | | | | | 0.05 | 0.04 | 1.05 |
| *Up to 5 calls | | | | | | | | | | | | | 0.18* | 0.04 | 1.20 |
| Belief | | | | | | | | | | 0.30* | 0.12 | 1.35 | 0.38* | 0.12 | 1.46 |
| *Up to 2 calls | | | | | | | | | | | | | -0.15* | 0.05 | 0.86 |
| *Up to 3 calls | | | | | | | | | | | | | -0.18* | 0.05 | 0.84 |
| *Up to 5 calls | | | | | | | | | | | | | -0.02 | 0.05 | 0.98 |
| RANDOM EFFECTS | | | | | | | | | | | | | | | |
| Question | 0.31 | 0.04 | | 0.29 | 0.04 | | 0.25 | 0.03 | | 0.24 | 0.03 | | 0.25 | 0.03 | |
| Response category | 0.35 | 0.03 | | 0.39 | 0.03 | | 0.39 | 0.03 | | 0.39 | 0.03 | | 0.39 | 0.03 | |
| Difference at time t | 0.22 | 0.01 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | |
| -2*loglikelihood | 15400 | | | 12287 | | | 12223 | | | 12216 | | | 12168 | | |
| Questions | 541 | | | 541 | | | 541 | | | 541 | | | 541 | | |
| Effects | 1243 | | | 1243 | | | 1243 | | | 1243 | | | 1243 | | |
| Effects*calls | 4972 | | | 4972 | | | 4972 | | | 4972 | | | 4972 | | |

*=p<0.05; B=logit coefficient; S.E. = standard error; exp(B) = ratios of estimated absolute differences.

**FIGURE 1 RESPONSE RATE PER CALL NUMBER FOR ALL SIX SURVEYS**

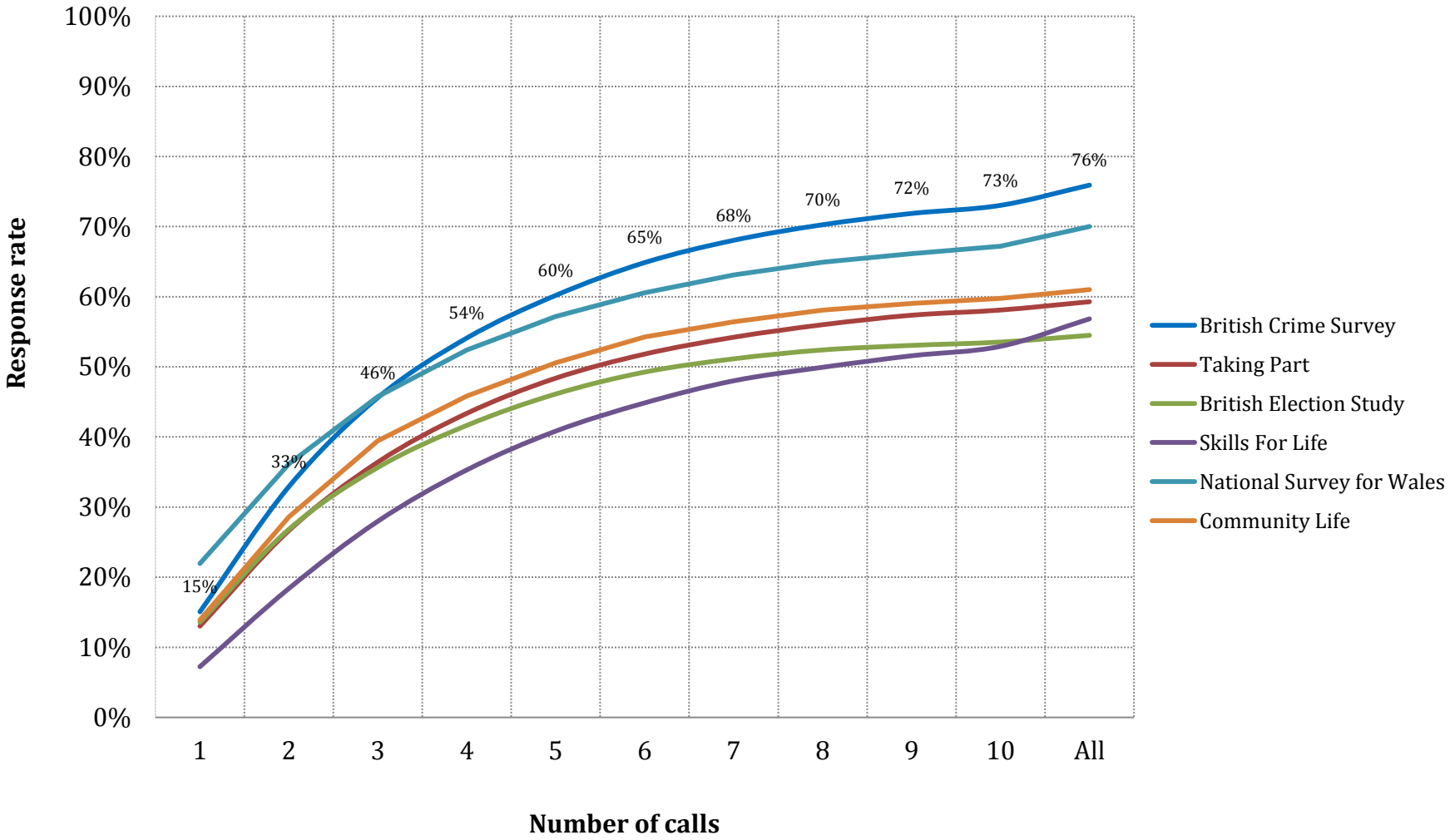


**FIGURE 2 ESTIMATED ABSOLUTE PERCENTAGE DIFFERENCE BY QUESTION (DESIGN WEIGHTED)**

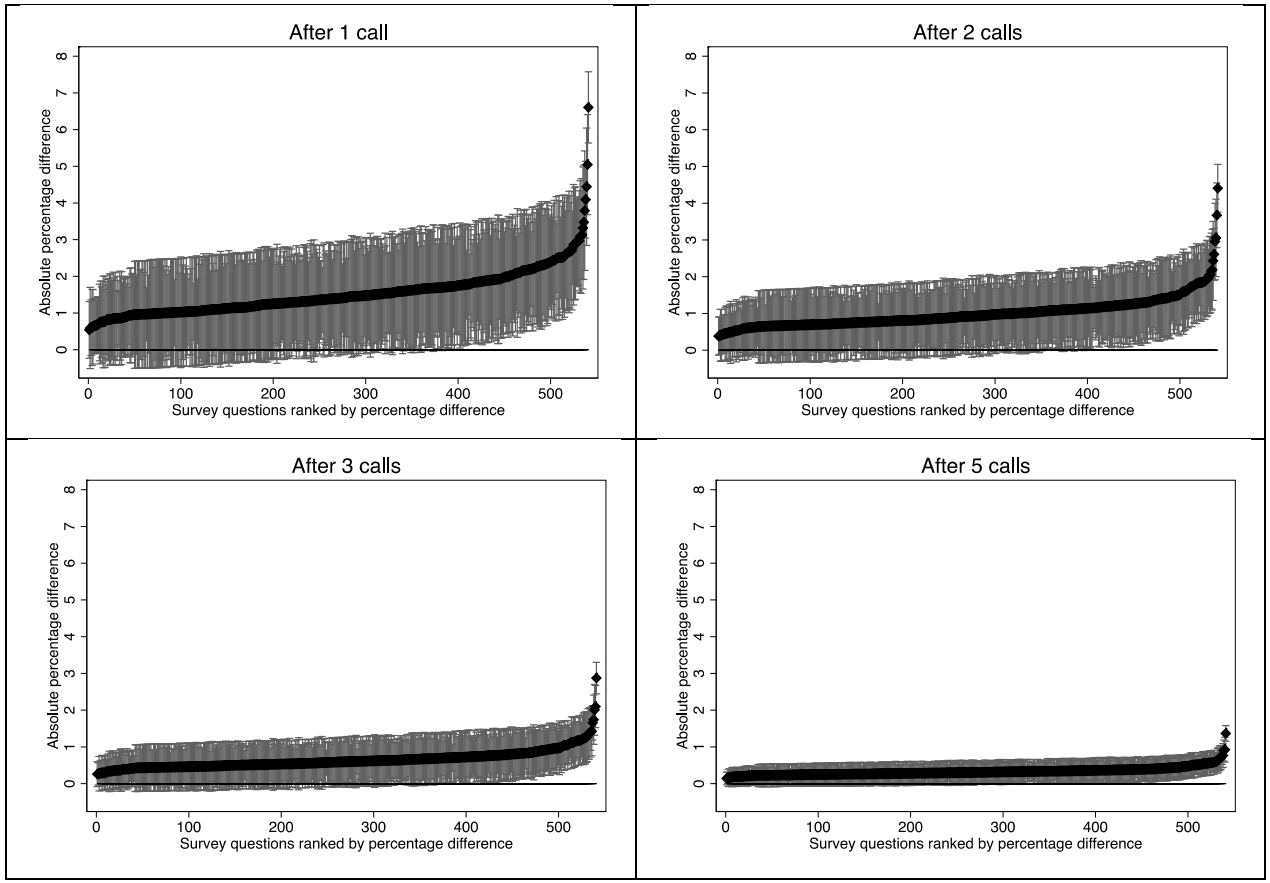


**FIGURE 3 ESTIMATED ABSOLUTE PERCENTAGE DIFFERENCE BY QUESTION (CALIBRATION WEIGHTED)**

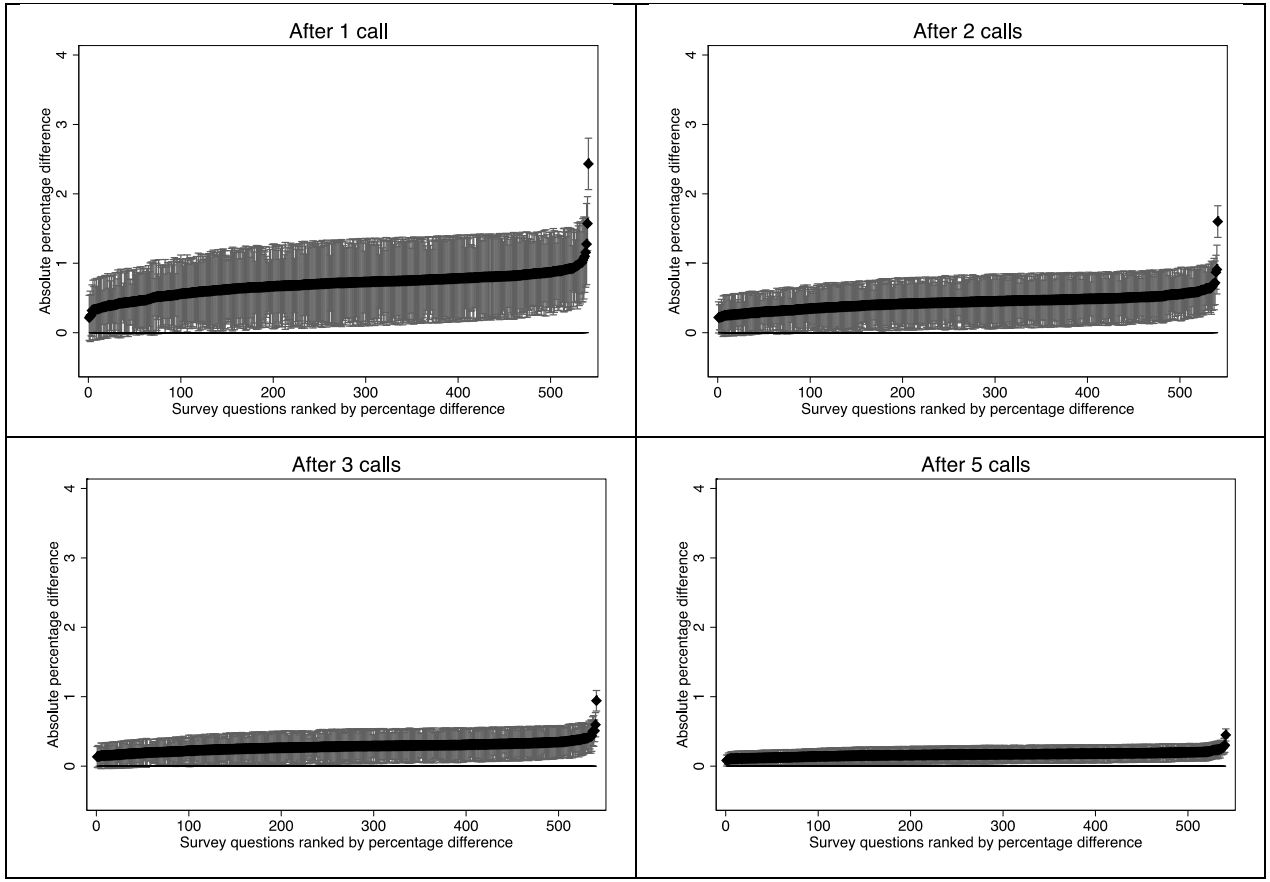


TABLE A1 MULTILEVEL MODELS, LOGGED RELATIVE PERCENTAGE DIFFERENCE, FINAL SAMPLE - CALL C (DESIGN WEIGHTED)

| | Model 1 | | | Model 2 | | | Model 3 | | | Model 4 | | | Model 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** |
| Constant | 0.67* | 0.05 | 1.95 | 1.33* | 0.05 | 3.79 | -0.04 | 0.10 | 0.96 | -0.09 | 0.10 | 0.92 | -0.09 | 0.10 | 0.92 |
| Call number (ref: 1 call) | | | | | | | | | | | | | | | |
| Up to 2 calls | | | | -0.44* | 0.02 | 0.64 | -0.44* | 0.02 | 0.64 | -0.44* | 0.02 | 0.64 | -0.42* | 0.02 | 0.66 |
| Up to 3 calls | | | | -0.82* | 0.02 | 0.44 | -0.82* | 0.02 | 0.44 | -0.82* | 0.02 | 0.44 | -0.80* | 0.02 | 0.45 |
| Up to 5 calls | | | | -1.42* | 0.02 | 0.24 | -1.42* | 0.02 | 0.24 | -1.42* | 0.02 | 0.24 | -1.45* | 0.02 | 0.23 |
| Questionnaire (ref: BCS) | | | | | | | | | | | | | | | |
| BES | | | | | | | 0.35* | 0.12 | 1.42 | 0.34* | 0.13 | 1.41 | 0.34* | 0.13 | 1.41 |
| SFL | | | | | | | 0.71* | 0.12 | 2.03 | 0.74* | 0.12 | 2.09 | 0.74* | 0.12 | 2.09 |
| TP | | | | | | | 0.73* | 0.14 | 2.08 | 0.73* | 0.14 | 2.07 | 0.73* | 0.14 | 2.07 |
| CLS | | | | | | | 0.47* | 0.10 | 1.60 | 0.49* | 0.10 | 1.64 | 0.49* | 0.10 | 1.64 |
| NSW | | | | | | | 0.23* | 0.11 | 1.26 | 0.24* | 0.11 | 1.26 | 0.24* | 0.11 | 1.26 |
| Variable (ref: Multi-coded) | | | | | | | | | | | | | | | |
| Single coded categorical | | | | | | | 1.92* | 0.12 | 6.84 | 1.93* | 0.12 | 6.86 | 1.92* | 0.12 | 6.85 |
| Single coded ordinal | | | | | | | 1.86* | 0.09 | 6.45 | 1.82* | 0.10 | 6.14 | 1.81* | 0.10 | 6.13 |
| Single coded binary | | | | | | | 0.51* | 0.14 | 1.67 | 0.55* | 0.14 | 1.73 | 0.55* | 0.14 | 1.73 |
| Question (ref: Behavior) | | | | | | | | | | | | | | | |
| Attitude | | | | | | | | | | 0.09 | 0.09 | 1.10 | 0.05 | 0.10 | 1.06 |
| *Up to 2 calls | | | | | | | | | | | | | -0.01 | 0.04 | 0.99 |
| *Up to 3 calls | | | | | | | | | | | | | 0.03 | 0.04 | 1.03 |
| *Up to 5 calls | | | | | | | | | | | | | 0.16* | 0.04 | 1.17 |
| Belief | | | | | | | | | | 0.15 | 0.13 | 1.16 | 0.23 | 0.13 | 1.26 |
| *Up to 2 calls | | | | | | | | | | | | | -0.14* | 0.05 | 0.87 |
| *Up to 3 calls | | | | | | | | | | | | | -0.18* | 0.05 | 0.84 |
| *Up to 5 calls | | | | | | | | | | | | | -0.04 | 0.06 | 0.96 |
| RANDOM EFFECTS | | | | | | | | | | | | | | | |
| Question | 1.01 | 0.09 | | 1.04 | 0.09 | | 0.25 | 0.04 | | 0.26 | 0.04 | | 0.26 | 0.04 | |
| Response category | 0.42 | 0.03 | | 0.46 | 0.03 | | 0.51 | 0.03 | | 0.50 | 0.03 | | 0.50 | 0.03 | |
| Difference at time t | 0.21 | 0.01 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | |
| -2*loglikelihood | 15908 | | | 12828 | | | 12436 | | | 12434 | | | 12398 | | |
| Questions | 541 | | | 541 | | | 541 | | | 541 | | | 541 | | |
| Effects | 1243 | | | 1243 | | | 1243 | | | 1243 | | | 1243 | | |
| Effects*calls | 4972 | | | 4972 | | | 4972 | | | 4972 | | | 4972 | | |

*=p<0.05; B=logit coefficient; S.E. = standard error; exp(B) = ratios of estimated absolute differences.

TABLE A2 MULTILEVEL MODELS, LOGGED ABSOUTE PERCENTAGE DIFFERENCE, FINAL SAMPLE - CALL C (CALIBRATION WEIGHTED)

| | Model 1 | | | Model 2 | | | Model 3 | | | Model 4 | | | Model 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** |
| Constant | -0.52* | 0.03 | 0.59 | 0.24* | 0.04 | 1.27 | -0.24* | 0.08 | 0.78 | -0.29* | 0.09 | 0.75 | -0.26* | 0.09 | 0.77 |
| Call number (ref: 1 call) | | | | | | | | | | | | | | | |
| Up to 2 calls | | | | -0.53* | 0.02 | 0.59 | -0.53* | 0.02 | 0.59 | -0.53* | 0.02 | 0.59 | -0.52* | 0.03 | 0.59 |
| Up to 3 calls | | | | -0.93* | 0.03 | 0.40 | -0.93* | 0.03 | 0.39 | -0.93* | 0.03 | 0.39 | -0.96* | 0.03 | 0.38 |
| Up to 5 calls | | | | -1.46* | 0.03 | 0.23 | -1.46* | 0.03 | 0.23 | -1.46* | 0.03 | 0.23 | -1.56* | 0.04 | 0.21 |
| Questionnaire (ref: BCS) | | | | | | | | | | | | | | | |
| BES | | | | | | | 0.97* | 0.10 | 2.65 | 0.95* | 0.11 | 2.58 | 0.94* | 0.11 | 2.57 |
| SFL | | | | | | | 0.55* | 0.10 | 1.73 | 0.57* | 0.10 | 1.76 | 0.57* | 0.10 | 1.76 |
| TP | | | | | | | 0.52* | 0.11 | 1.69 | 0.51* | 0.11 | 1.66 | 0.50* | 0.11 | 1.65 |
| CLS | | | | | | | 0.31* | 0.08 | 1.36 | 0.33* | 0.08 | 1.39 | 0.33* | 0.08 | 1.39 |
| NSW | | | | | | | 0.06 | 0.09 | 1.06 | 0.05 | 0.09 | 1.05 | 0.05 | 0.09 | 1.05 |
| Variable (ref: Multi-coded) | | | | | | | | | | | | | | | |
| Single coded categorical | | | | | | | 0.19 | 0.10 | 1.20 | 0.19* | 0.10 | 1.21 | 0.20* | 0.10 | 1.22 |
| Single coded ordinal | | | | | | | 0.33* | 0.07 | 1.39 | 0.27* | 0.08 | 1.31 | 0.27* | 0.08 | 1.31 |
| Single coded binary | | | | | | | 0.07 | 0.11 | 1.08 | 0.12 | 0.12 | 1.13 | 0.12 | 0.12 | 1.13 |
| Question (ref: Behavior) | | | | | | | | | | | | | | | |
| Attitude | | | | | | | | | | 0.16* | 0.08 | 1.17 | 0.05 | 0.08 | 1.05 |
| *Up to 2 calls | | | | | | | | | | | | | 0.04 | 0.05 | 1.04 |
| *Up to 3 calls | | | | | | | | | | | | | 0.14* | 0.05 | 1.15 |
| *Up to 5 calls | | | | | | | | | | | | | 0.24* | 0.06 | 1.27 |
| Belief | | | | | | | | | | 0.09 | 0.11 | 1.09 | 0.15 | 0.12 | 1.16 |
| *Up to 2 calls | | | | | | | | | | | | | -0.16 | 0.09 | 0.85 |
| *Up to 3 calls | | | | | | | | | | | | | -0.19* | 0.09 | 0.83 |
| *Up to 5 calls | | | | | | | | | | | | | 0.05 | 0.09 | 1.06 |
| RANDOM EFFECTS | | | | | | | | | | | | | | | |
| Question | 0.25 | 0.03 | | 0.27 | 0.03 | | 0.18 | 0.03 | | 0.19 | 0.03 | | 0.19 | 0.03 | |
| Response category | 0.24 | 0.02 | | 0.27 | 0.02 | | 0.25 | 0.02 | | 0.25 | 0.02 | | 0.25 | 0.02 | |
| Difference at time t | 0.17 | 0.01 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | |
| -2*loglikelihood | 16242 | | | 14083 | | | 13939 | | | 13935 | | | 13902 | | |
| Questions | 541 | | | 541 | | | 541 | | | 541 | | | 541 | | |
| Effects | 1243 | | | 1243 | | | 1243 | | | 1243 | | | 1243 | | |
| Effects*calls | 4972 | | | 4972 | | | 4972 | | | 4972 | | | 4972 | | |

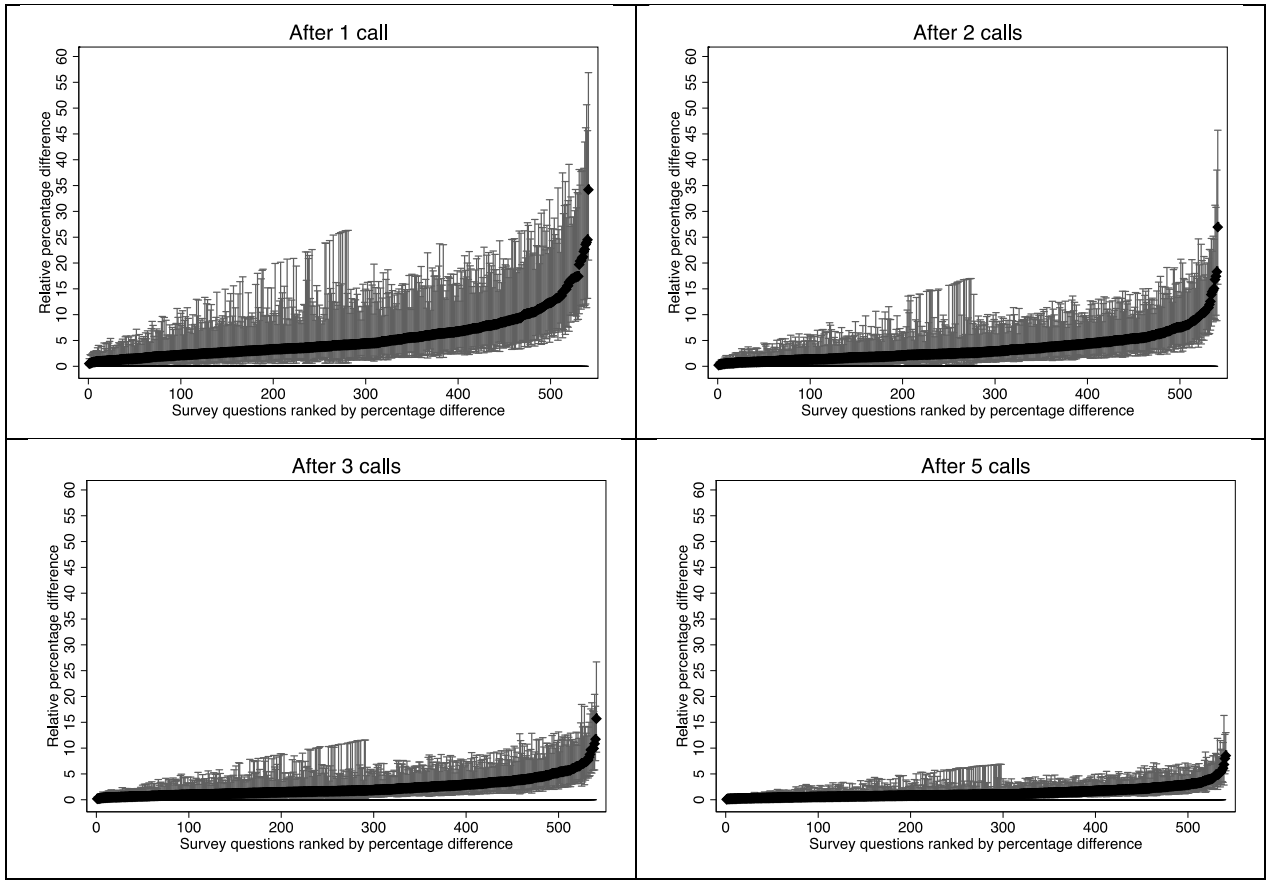
*=p<0.05; B=logit coefficient; S.E. = standard error; exp(B) = ratios of estimated absolute differences.

TABLE A3 MULTILEVEL MODELS LOGGED RELATIVE PERCENTAGE DIFFERENCE, FINAL SAMPLE - CALL C (CALIBRATION WEIGHTED)

| | Model 1 | | | Model 2 | | | Model 3 | | | Model 4 | | | Model 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** | **B** | **S.E.** | **exp(B)** |
| Constant | 0.45* | 0.05 | 1.56 | 1.18* | 0.05 | 3.25 | -0.28* | 0.09 | 0.75 | -0.30* | 0.09 | 0.74 | -0.27* | 0.10 | 0.76 |
| Call number (ref: 1 call) | | | | | | | | | | | | | | | |
| Up to 2 calls | | | | -0.52* | 0.03 | 0.59 | -0.53* | 0.03 | 0.59 | -0.53* | 0.03 | 0.59 | -0.53* | 0.03 | 0.59 |
| Up to 3 calls | | | | -0.93* | 0.03 | 0.39 | -0.93* | 0.03 | 0.39 | -0.93* | 0.03 | 0.39 | -0.96* | 0.04 | 0.38 |
| Up to 5 calls | | | | -1.45* | 0.03 | 0.24 | -1.45* | 0.03 | 0.23 | -1.45* | 0.03 | 0.23 | -1.55* | 0.04 | 0.21 |
| Questionnaire (ref: BCS) | | | | | | | | | | | | | | | |
| BES | | | | | | | 0.82* | 0.11 | 2.27 | 0.77* | 0.11 | 2.15 | 0.76* | 0.11 | 2.14 |
| SFL | | | | | | | 0.71* | 0.11 | 2.04 | 0.69* | 0.11 | 1.99 | 0.69* | 0.11 | 2.00 |
| TP | | | | | | | 0.65* | 0.13 | 1.92 | 0.62* | 0.13 | 1.85 | 0.61* | 0.13 | 1.84 |
| CLS | | | | | | | 0.55* | 0.09 | 1.74 | 0.55* | 0.09 | 1.74 | 0.55* | 0.09 | 1.74 |
| NSW | | | | | | | 0.31* | 0.09 | 1.36 | 0.28* | 0.09 | 1.32 | 0.28* | 0.09 | 1.32 |
| Variable (ref: Multi-coded) | | | | | | | | | | | | | | | |
| Single coded categorical | | | | | | | 2.01* | 0.10 | 7.43 | 2.01* | 0.10 | 7.44 | 2.01* | 0.10 | 7.48 |
| Single coded ordinal | | | | | | | 1.92* | 0.08 | 6.84 | 1.88* | 0.09 | 6.53 | 1.88* | 0.09 | 6.55 |
| Single coded binary | | | | | | | 0.22 | 0.13 | 1.24 | 0.25 | 0.13 | 1.29 | 0.25 | 0.13 | 1.29 |
| Question (ref: Behavior) | | | | | | | | | | | | | | | |
| Attitude | | | | | | | | | | 0.15 | 0.08 | 1.17 | 0.05 | 0.09 | 1.05 |
| *Up to 2 calls | | | | | | | | | | | | | 0.04 | 0.05 | 1.04 |
| *Up to 3 calls | | | | | | | | | | | | | 0.12* | 0.06 | 1.13 |
| *Up to 5 calls | | | | | | | | | | | | | 0.24* | 0.06 | 1.28 |
| Belief | | | | | | | | | | -0.12 | 0.11 | 0.89 | -0.07 | 0.12 | 0.93 |
| *Up to 2 calls | | | | | | | | | | | | | -0.14 | 0.09 | 0.87 |
| *Up to 3 calls | | | | | | | | | | | | | -0.16 | 0.09 | 0.85 |
| *Up to 5 calls | | | | | | | | | | | | | 0.07 | 0.09 | 1.07 |
| RANDOM EFFECTS | | | | | | | | | | | | | | | |
| Question | 0.94 | 0.09 | | 1.00 | 0.09 | | 0.13 | 0.03 | | 0.12 | 0.03 | | 0.12 | 0.03 | |
| Response category | 0.49 | 0.04 | | 0.51 | 0.04 | | 0.51 | 0.03 | | 0.51 | 0.03 | | 0.51 | 0.03 | |
| Difference at time t | 0.16 | 0.01 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 | |
| | 17025. | | | 14910. | | | 14378. | | | 14369. | | | 14340. | | |
| -2*loglikelihood | 205 | | | 883 | | | 539 | | | 679 | | | 94 | | |
| Questions | 541 | | | 541 | | | 541 | | | 541 | | | 541 | | |
| Effects | 1243 | | | 1243 | | | 1243 | | | 1243 | | | 1243 | | |
| Effects*calls | 4972 | | | 4972 | | | 4972 | | | 4972 | | | 4972 | | |

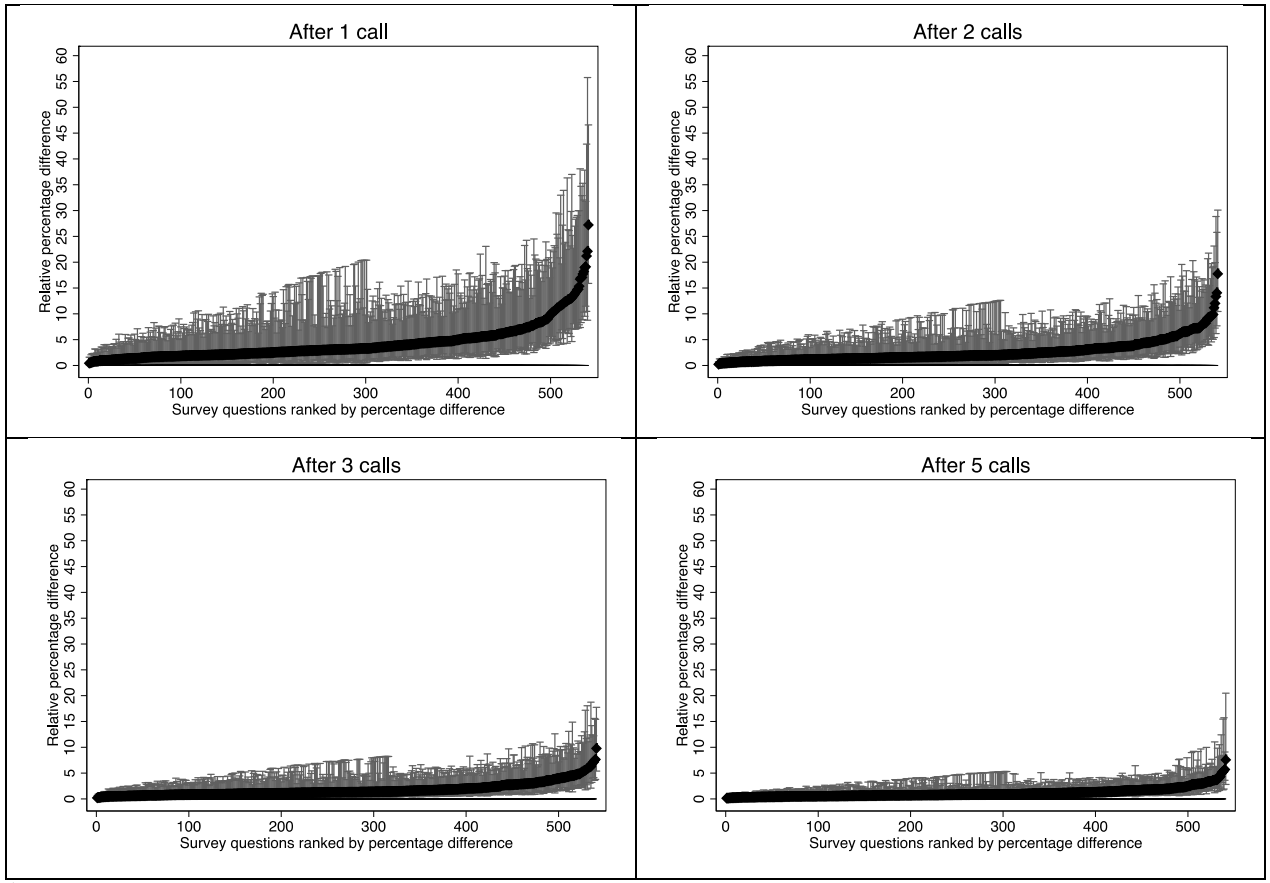
*=p<0.05; B=logit coefficient; S.E. = standard error; exp(B) = ratios of estimated absolute differences.

**FIGURE A1 ESTIMATED RELATIVE PERCENTAGE DIFFERENCE BY QUESTION (DESIGN WEIGHTED)[1]**




[1] Estimated using log-transformed data

**FIGURE A2 ESTIMATED RELATIVE PERCENTAGE DIFFERENCE BY QUESTION (CALIBRATION WEIGHTED)[1]**




[1] Estimated using log-transformed data