Is success in obtaining contact and cooperation correlated with the magnitude of interviewer variance?

Ian Brunton-Smith[1], Patrick Sturgis[2] and Joel Williams[3]

[1], Department of Sociology, University of Surrey, Guildford, Surrey, UK. Tel: +44 (0)1483 68 6965, fax: +44 (0)1483 68 69 email: i.r.brunton-smith@surrey.ac.uk

[2] Division of social statistics, University of Southampton, Southampton, UK. Tel: +44 (0)23 8059 4082, fax: +44 (0)23 8059 3858, email: p.sturgis@soton.ac.uk

[3] TNS - BMRM, 6 More London Place, London, UK. Tel: +44 (0)20 7656 5778, email: joel.williams@tns-TNS-BMRB.co.uk

Word count : 6492

RUNNING HEADER: CONTACT, COOPERATION AND INTERVIEWER VARIANCE

Ian Brunton-Smith is a lecturer in the Department of Sociology, University of Surrey, UK.

Patrick Sturgis is a Professor in the School of Social Sciences, University of Southampton, UK

Joel Williams is Head of Survey Methods, TNS-BMRB, London. We thank Chris Skinner, Gabi Durrant, three anonymous reviewers and the editors, whose comments improved earlier versions of this paper. Any errors or omissions are entirely our own. This work was funded through grants provided by the Economic and Social Research Council (Ian Brunton-Smith - PTA-033-2005-00028, and Patrick Sturgis - RES-576-47-5001). Data collection for the British Crime Survey was funded by the UK Home Office. Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author(s) and do not necessarily reflect the views of the funding organizations.

Address correspondence to: Ian Brunton-Smith, Department of Sociology, University of Surrey, Guildford, Surrey, UK GU2 7XH. Tel: +44 (0)1483 68 6965, fax: +44 (0)1483 68 email: i.r.brunton-smith@surrey.ac.uk

Abstract

Evidence is now beginning to accumulate which shows that interviewer attitudes, personality, and behavior are predictive of success in achieving contact and cooperation with sampled households. A less frequently explored possibility, however, is that these same characteristics might also be the source of variability in the extent to which interviewers follow best practice in the implementation of standardized interviewing. That is to say, there may be a correlation between interviewer-induced nonresponse bias and measurement error. In this paper we provide the first empirical investigation of the direction and magnitude of the relationship between interviewer skill in obtaining contact and cooperation and correlated interviewer error. Drawing on face-to-face interview data from a large, multi-stage probability sample of the British population, we use cross-classified multilevel models with a complex error structure to examine how the interviewer variance component varies as a function of historical measures of interviewer skill in obtaining contact and cooperation. Our results show that, across a broad range of variables, interviewers with a history of obtaining poor rates of contact and cooperation exhibit higher levels of correlated interviewer error than their better-performing colleagues. For cooperation, we find some evidence of a u-shaped relationship, with the least *and* the most successful interviewers having the largest interviewer variance component.

There are a large number of opportunities throughout the various stages of fieldwork for interviewers to contribute to survey error (Couper and Groves, 1992; Groves, 1989; Groves and Couper, 1998). Perhaps the most obvious of these relates to the bias that can be introduced through differential nonresponse; if some interviewers employ strategies and behaviors which result in non-contacts and refusals where a more skilled interviewer might have obtained an interview, then estimates will be more biased if this group of nonrespondents differs from respondents on the survey variables of interest. It is for this reason that survey agencies devote considerable resources to understanding the features of successful doorstep approaches (Campanelli et al., 1997; de Leeuw et al., 1998; Groves and Couper, 1998) and to training field-forces in their effective implementation.

Interviewers can also contribute to survey error through the manner in which they administer questions to respondents (Cannell et al., 1981; Fowler and Mangioine, 1990; Freeman and Butler, 1976; Groves 1989; Marquis and Cannell, 1969). These interviewer-induced measurement errors are thought to arise through interviewers diverting from the questions as they are written, not following the interviewer instructions (relating, for example, to probing and show cards), and 'helping' respondents to understand and formulate responses to difficult questions (Biemer and Lyberg 2003; Cannell et al., 1981; Kish 1962; Mangione et al. 1992; O'Muircheartaigh, 1976; Schnell and Kreuter, 2005). Although these practices will sometimes result in systematic differences between the true population value and the survey estimate, in the vast majority of cases, external criteria are not available to allow their characterisation as biases in this manner. It is more straightforward, however, to determine the combined effect of these idiosyncrasies of individual interviewer behaviour on the *precision* of estimates. Because each interviewer will tend to divert from standardized procedures in the same way over repeated interviews (e.g. he or she always reads out a particular question incorrectly), the end result is an increase in within-interviewer homogeneity of responses (Biemer, 2010; Kish, 1962; Mahalanobis 1946). Because this introduces an additional source of variability into the population estimator, the eventual upshot is larger standard errors. Existing studies have shown that the interviewer variance component can substantially reduce the precision of survey estimates, although the exact magnitude of the effect depends heavily on both the

nature of the question and the size of interviewer workloads (Collins and Butcher 1982; Couper and Groves, 1992; Mangione et al., 1992).

Traditionally, survey methodologists have focused on such observational and non-observational errors in isolation from one another. More recently, however, attention has increasingly turned to a consideration of the possibility that these error types may covary due to a common underlying cause (Groves, 2006; Groves and Lyberg, 2010; Olson, 2006). Where there is a common cause of observational and non-observational errors, attempts to reduce one type of error may have unintended consequences for the other. Whether these consequences are harmful or benign depends on the direction of the correlation induced by the common cause. If a positive correlation is apparent, then efforts to reduce one type of error will also serve, whether intentionally or not, to reduce the other. However, where the common cause results in a negative correlation, then efforts to reduce one error type will be off-set by increases in the other and might even result in a *larger* total survey error (Kreuter et al., 2010).

To date scholars have focused on respondents as the (unwitting) agents of this potential 'error trade-off'. For instance, Kominska et al. (2010) show that respondents who display reluctance to participate in a survey are more likely to use a 'satisficing' response style (Krosnick, 1991), while Tourangeau et al. (2010) consider whether increasing the proportion of respondents in the sample who are not interested in politics, via a monetary incentive, results in higher rates of misreporting of electoral turnout (see also Fricker and Tourangeau 2010; Groves et al., 2006; Groves and Peytcheva, 2008; Kreuter et al 2010; Sakshaug et al 2010). The underlying rationale in these studies is that efforts to reduce nonresponse bias by converting initial refusals into successful interviews may be counter-productive if the 'converted refusals' end up providing error-ridden responses. Of course, it may also be the case that poor data quality arising from these situations is also partially due to interviewers, who may rush through the questionnaire as a way of minimizing the burden on reluctant respondents.

By way of contrast, however, no existing study has yet empirically investigated whether interviewer skill in locating and persuading sample members to agree to a survey request might also be related to the degree of interviewer-induced measurement error. This is a surprising lacuna, because there has long been anecdotal

evidence within survey agencies that interviewers who are skilled at gaining cooperation may be less committed to the principles of standardized interviewing (Groves and Lyberg 2010). There are also theoretical reasons to believe that at least some of the factors that have been shown to underpin interviewer success in obtaining contact and cooperation might also result in the kinds of behaviors that lead to within-interviewer correlated error in survey outcomes. For instance, there is growing evidence to suggest that various aspects of interviewer personality, attitudes, and behavior are predictive of success in making contact with sample members and obtaining interviews once contact is made (Blom et al., 2010; Couper and Groves, 1992; Durrant et al, 2010; Hox and de Leeuw, 2002; Singer et al., 1983; Snijkers et al., 1999). Such evidence is of practical as well as of theoretical interest because it points to ways in which survey agencies might develop recruitment strategies and training programs in order to produce a panel of interviewers who possess the characteristics that are likely to result in interviews with sample members and, as a result, minimize nonresponse bias in survey estimates.

Yet these same dispositional and behavioral characteristics might also give rise to a 'conversational' interviewing style, which increases the variance of estimators. For instance, it is plausible that a disposition that promotes flexible tailoring on the doorstep and an ability to maintain an interaction with a stranger (Groves and Couper, 1998; Groves and McGonagale, 2001; Sturgis and Campanelli, 1998) might also result in higher rates of deviation from the questionnaire script, of differential probing effort, or of helping respondents with interpretation of difficult questions. A conceptual model setting out how such a negative correlation between interviewer induced nonresponse and measurement error might arise is specified as a path diagram in Figure 1.

INSERT FIGURE 1 HERE

In Figure 1, an interviewer trait, agreeableness, is shown (along the upper chain of the diagram) to have a positive effect on the extent to which interviewers engage in tailoring behaviour on the doorstep (Groves and Couper, 1998). Agreeableness is one of the dimensions in the 'Big Five' personality inventory, with higher scoring individuals being more empathetic, trusting and cooperative (Digman 1990). Tailoring results in higher rates of cooperation for these interviewers and this, in turn, leads to a reduction in the bias of the estimator in

question.[1] On the lower chain of the diagram, we see that this same disposition also results in interviewers diverting from the principles of standardized interviewing, such as failing to read the questions as written and clarifying the meaning of words and questions for respondents. This might happen if, for instance, interviewers who score higher on the agreeableness dimension dislike the formal and rather stylized interactional format of the standardized interview. Diverting from the questions as written results in a heightened within-interviewer variance component which, in conjunction with the size of the interviewer assignment, serves to increase the design effect due to interviewers. Thus, a negative covariance between nonresponse and measurement error has been induced by the common cause. A negative correlation of this nature implies that recruiting and training interviewers to possess these characteristics as a strategy for reducing nonresponse bias might be offset, or even outweighed, by a concomitant increase in measurement error (Groves and Lyberg 2010).

Of course, the causal model set out in Figure 1 is speculative and the true direction of the covariance between interviewer-induced nonresponse and measurement error may not be negative. Other traits, such as conscientiousness, for example, might be expected to underlie both success in obtaining contact and cooperation and sticking closely to interviewer instructions. Put differently, conscientiousness is a trait which we should expect to lead interviewers to conduct all aspects of their work *according to how they were trained*. In this case, the induced covariance between nonresponse and measurement error would be positive. So, because there are good grounds for assuming that the true relationship between interviewer-induced nonresponse and measurement error might be positive or negative, we do not specify a hypothesis about the direction of this association but treat this as a purely empirical question at this stage.

Data

We draw our data from the 2005-2006 round of the British Crime Survey (BCS), a nationally representative victimisation survey of adults aged 16 years or over, living in private residential accommodation in England and Wales. We restrict our focus to the sample from England, because linked census data that are required for our

[1] This, of course, requires the additional assumption that for at least some variables in the survey there is a correlation between the cooperation rate and the magnitude of nonresponse bias, which we believe to be reasonable.

model specification are not available for Wales. Data is collected using a multistage, stratified sample design, in which a sample of postcode sectors is first drawn with probability proportional to size, before a sample of 32 households is selected in each sector and assigned to an interviewer. At each address, the interviewer randomly selects an individual to take part in the survey (see Grant et al., 2006). In 2005-2006, a total of 57,795 households in England were issued, with a total response rate of 73%,[2] which provides us with an analytical sample of 42,350 individuals nested within 419 interviewers.

Measuring Interviewer skill in obtaining contact and cooperation

It is widely accepted that interviewers vary in their ability to convert eligible sample members into interviews (Groves and Couper 1998; Sturgis and Campanelli 1998). However, the raw response rate an interviewer obtains is problematic as a mesure of interviewer skill in this task because it is confounded with a range of factors that are themselves predictive of whether a sampled household will be converted to a successful interview. Indeed, because survey agencies often deliberately allocate the most difficult cases to their most experienced interviewers, it is possible that the 'best' interviewers will sometimes have some of the lowest raw response rates. Similarly, because some areas (e.g. in cities) yield consistently lower response rates than others, a more skilled interviewer may obtain a lower response rate than a less skilled interviewer, simply as a result of socio-economic differences between the areas in which they are working. For these reasons, we use measures of interviewer success in obtaining response which adjust for the difficulty of the assignments. Additionally, the skills required to obtain high cooperation rates are likely to be rather different in nature from those that underpin success in making contact with householders. While gaining cooperation is essentially a matter of using persuasive verbal and non-verbal interactional techniques (Morton-Williams, 1993) and understanding and reacting appropriately to the sorts of objections householders are likely to raise (Campanelli et al., 1997; Groves and Couper, 1998), making contact is more a case of organizing one's time effectively, following best-practice guidelines and being persistent. Because these different skill-sets are unlikely to have exactly the same effect on the way interviewers administer questionnaires, we consider the relationship between within-interviewer variance and success in obtaining contact and cooperation separately.

[2] AAPOR Response Rate 1.

To correct for the 'difficulty' of an interviewer's assignment, we specify interviewer skill in obtaining contact and cooperation, respectively, as the difference between the achieved contact /cooperation rate across all his or her assignments for the BCS for the period April 2004 to March 2007[3] and the 'expected' contact and cooperation rates, given the profile of the cases issued to them. Where a household was initially unproductive with one interviewer but was then reissued to a more experienced colleague for refusal conversion, these were excluded from the analysis. To calculate the expected contact and cooperation rate for each interviewer's assignment of cases, we used CHAID analysis (Kass, 1980).[4] The outcome in the CHAID analysis was a case-level binary indicator of whether contact/cooperation was made at the address. The predictor variables were a range of area and household characteristics that have been found to be predictive of nonresponse in the existing literature (Campanelli et al 1997; Durrant and Steele, 2009; Groves and Couper, 1998). These were: the Police Force Area and ACORN[5] group of each address, interviewer observations about the local housing conditions, whether the household is in a neighbourhood watch area, the type of accommodation, whether the property has an answer phone, whether incentive stamps were used,[6] and whether the interviewer was required to ask about the ethnic status of individuals living in neighbouring addresses. A minimum size of 1,000 cases was specified for the response propensity groups.

This procedure resulted in 64 separate contact propensity groups, with probabilities ranging from .48 to .99. For cooperation, the procedure resulted in 80 propensity groups, with probabilities ranging from .41 to .94. The contact and cooperation measures were then constructed in the same way. For each interviewer, the expected rate was calculated as the mean of the propensities for the CHAID groups to which their eligible issued households were allocated. Taking the ratio of the observed to the expected rate for each interviewer's eligible

[3] Note that this covers an additional two years of fieldwork on the BCS than we use to produce our estimates of interviewer variance. For our estimates of interviewer variance, we use only data from the 2005-06 fieldwork period.
[4] An alternative approach would be to use predicted probabilities from a logistic regression model. While the results are essentially the same using either method, we prefer the CHAID approach due to its more flexible ability to detect higher-order interactions between the predictor variables.
[5] ACORN is a neighbourhood classification scheme developed by CACI Ltd. that classifies households according to the demographic, employment, and housing characteristics of the surrounding neighbourhood. There are five main ACORN groups, which are given the following descriptive labels: Wealthy Achievers; Urban Prosperity; Comfortably Off; Moderate Means; and Hard Pressed (http://www.caci.co.uk/acorn).
[6] In an attempt to improve survey response rates, a random subset of all households selected for interview was issued with a book of 6 first class stamps along with the letter inviting them to participate in the survey (for a summary of evidence on the use of incentives see Simmons and Wilmot, 2004).

case load provides us with our measure of interviewer skill in obtaining contact and cooperation, adjusted for what can be thought of as a 'difficulty tariff' for the workload each interviewer was assigned.

Estimating Interviewer Variance

To estimate the interviewer variance component we use a cross-classified multi-level model (Rasbash and Goldstein, 1994). In order to determine whether interviewers with different levels of success in obtaining contact and cooperation exhibit response variances of differing magnitudes, this incorporates a complex variance term for a level 2 variable (Goldstein, 2003). In statistical terms, this means that rather than estimating a global variance for all interviewers, we specify a random effect on the interviewer-level variable. Assuming, for simplicity of explication, a binary interviewer-level cooperation success variable, denoted , the model has the following form:

[1]

if

if

Adopting the notation of Rasbash and Goldstein (1994), refers to survey outcome, *y*, measured for the i^{th} respondent within the cross-classification of interviewer j_1 and area j_2. is the intercept, which is allowed to vary across interviewers and areas. However, unlike the standard cross-classified model, the variability about this intercept is modelled separately for each category of the binary interviewer-level variable with residual errors, and . These are assumed IID, with mean zero and variances and , and denote the interviewer variance for each level of the interviewer success variable, . All covariances in the interviewer variance-covariance matrix, , are constrained to zero, reflecting the fact that these are mutually exclusive

categories. Also in the random part of the model are the error terms [ ] and [ ], denoting the residual variability across areas and individuals, with variances [ ] and [ ] respectively.

Because, in this study, respondents were not randomly allocated to interviewers, the variance components in equation 1 will be biased if respondent characteristics are systematically different across interviewer workloads. That is to say, using this approach, variance in a survey outcome might be attributed to the behavior of interviewers, when it actually arises from the fact that certain kinds of interviewers (e.g. men) tend to be allocated certain types of respondents (e.g. those residing in inner cities), who are more similar to one another on the survey outcome of interest than they are to the general population. The potential for bias of this nature can be mitigated by controlling for the observed characteristics of respondents, interviewers and areas – the [ ], [ ], [ ], terms in equation 1. It is this approach that we use here. However, as with all estimators that rely on statistical control for unbiasedness, the approach is only successful if the full range of necessary covariates is included in the model. And, of course, whether this has been achieved will not generally be known. We consider the robustness of our inferences to the possibility of unobserved variable bias in the discussion section of the paper.

The model in equation 1 is only implemented for categorical variables therefore we use quintile groups rather than the continuous measures of interviewer success, with the bottom quintile group representing interviewers with the worst performance. Although forming quintiles results in a certain loss of information relative to the continuous measures, it does have the analytical benefit of greater flexibility in detecting potential non-linearities between the interviewer success variables and the magnitude of interviewer variance. Five groups proved to be the maximum that it was possible to use before experiencing irresolvable convergence problems during the model fitting stage.[7]

[7] Analyses not reported here show the same pattern of results in obtained using tercile and quartlle categorisations. These are available from the corresponding author upon request.

It is important to note that we do not include the interviewer success variable as a fixed effect in this model specification. This is because our aim is to use the contact and cooperation success quintiles to *partition* rather than to *explain* the global interviewer variance; we are interested in the difference in interviewer variances between these groups before, not after, it has been controlled for in the fixed part of the model.[8] Including the success quintiles as level-2 fixed effect dummy variables would tell us how estimates of item means vary over the quintile groups, when our interest is in differences in variances. We do, however, control for all other interviewer characteristics available to us, in order to be confident that any differences we observe are due to the interviewer success variables rather than some other characteristic with which they are correlated. Post-estimation, it is straightforward to calculate the within-interviewer correlation for each category of the interviewer level variable. For the first category of [ ], we have:

[ ] [2]

Analytical Strategy

The model specified in equation 1 was estimated for all 36 eligible items in the BCS for both the contact and cooperation measures of success, resulting in a total of 72 separate models. Eligible items were defined as all those which were administered to the full sample, which could be treated as a continuous outcome in a linear model, and which required some degree of input from the interviewer in the form of either probing and/or use of show-cards. Factual items with no probing or show-cards were excluded at this stage because existing research has shown that the degree of interviewer variance is generally close to zero on factual items that require no probing or show-cards (Schnell and Kreuter, 2005). Models were estimated within the generalised linear mixed modelling framework (lme4) in R version 2.12.1 using a restricted maximum likelihood estimation procedure (Bates and Maechler, 2010).

Interviewer characteristics

[8] Note, however, that including the interviewer success quintiles in the model as level 2 fixed effects produces essentially the same pattern of results as are presented here. These are available from the corresponding author upon request.

At the interviewer level, we include fixed effects for all those characteristics for which measures were available: gender, age (in years), ethnicity (white/non-white), and a measure of experience (a count of the total number of months the interviewer has worked on the BCS). This allows us to separate the effect of our doorstep success measure from some of the characteristics with which it might be correlated and which also influence intra-interviewer response error.

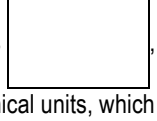
Area-level characteristics

As our area-level identifier, we use the Middle Layer Super Output Area (MSOA) geography (Martin, 2001) rather than the postcode sector. MSOAs contain, on average, 5,000 households and have been designed with the intention that they are more homogenous in size and social structure, remain stable over time, and with a view to maintaining 'natural' boundaries at a small area level. We therefore prefer MSOA to postcode sectors because they are a more meaningful spatial unit at which to specify 'area' level variance but also because it is possible to attach a rich variety of variables from the census and other sources to provide more powerful control for the non-random allocation of respondents to interviewers across areas.[9]

At the MSOA level, 21 different variables were merged in from the 2001 census of England and Wales covering a broad range of social, economic, demographic, and structural characteristics. These were combined using a principal components analysis with orthogonal rotation to generate a series of summary indices about each local area. This yielded a 5-component solution, with the five components representing the area level of economic deprivation, urbanisation, population migration, age structure and housing structure. We also include a measure of the level of ethnic diversity of each MSOA, assessed using the Herfindahl concentration formula (Hirschman, 1964)[10].

Respondent characteristics

[9] Repeating our analysis using postcode sectors with fewer controls shows essentially the same pattern of results that are presented here. These are available from the corresponding author upon request.

[10] The Herfindahl concentration formula is calculated as  , where s_i is the proportion of ethnic group, *i*, in each geographical unit and N is the total number of geographical units, which in our case are MSOA. It has a theoretical range of 0-1. Values can be interpreted as the probability that two individuals drawn randomly from the same geographical unit will be from a different ethnic group.

To ensure that each interviewer assignment is broadly comparable, at the individual level we control for respondent gender, age, ethnicity, and education level. Because the interviewer is also responsible for collecting basic demographic information about each respondent and, as such, these measurements are also subject to a degree of interviewer error, we are restricted here to the core set of individual characteristics least likely to be affected by the interviewer. Given the requirement for MSOA to be internally homogenous on key socio-economic characteristics, we expect that many other differences in interviewer assignments will be captured by our range of area level measures. Of course, if the composition of respondents across interviewer success quintiles differs notably on other individual characteristics, we may still be overestimating any differences in the variances associated with each groups.

Results

We estimate the model in equation 1 for all 36 in-scope items, partitioning the interviewer variance component as a function of the measures of interviewer success in making contact with eligible households and obtaining cooperation, conditional on contact. Because this yields a total of 72 models, each containing a large number of parameters, space precludes presentation of the full set of results here. Instead, we first present the complete results for a single exemplar item, before moving on to summary statistics derived from the complete set of analyses. The within-interviewer correlations for all 72 models are included in the online appendix to this article (see online appendix 1). Table 1 shows the parameter estimates for an attitudinal item assessing respondents' self-reported health status, measured on a 5-point scale ranging from 'very good (1)' to 'very bad (5)'. We can see that the respondent and area level fixed effects for this item generally conform to expectations regarding the correlates of self-reported health with women, younger people, and those with higher educational qualifications reporting better health. At the area level, higher levels of economic disadvantage, population mobility and areas with more flats and terraced housing are associated with lower health ratings. Of the interviewer variables, only gender is significant, with male interviewers more likely to obtain lower ratings of heatlh.

INSERT TABLE 1 HERE

Our primary interest, however, is not in the fixed but in the random part of the model. Looking at the random effects for each of the interviewer success quintiles, we can see that for the cooperation measure there is little variation between the top 4 quintile groups (range=.015-.020) and, indeed, these are not significantly different from one another ($p<0.05$).[11] However, the bottom quintile shows a significantly larger variance ($p<0.05$), which is approximately double the magnitude of the other quintile groups (0.033). The same general pattern is observed for the contact measure, with the bottom quintile group having a variance which, at 0.035, is significantly larger than the remaining groups. On the contact measure, there is some indication of a more linear relationship between success and the magnitude of interviewer variance.

For this item, then, our working hypothesis is supported: interviewer success in obtaining contact and cooperation is strongly associated with the magnitude of interviewer variance. With regard to the direction of this relationship, it is those interviewers who are *least* successful on both contact and cooperation measures who exhibit considerably larger variances than the remaining quintile groups. Assuming an average interviewer assignment of 101 respondents across the data collection period (based on a sample of 42,288 divided equally across respondents and 419 interviewers), the design effect due to interviewers alone is 52% higher in this group when considering contact, and 67% higher when considering co-operation, compared to the average across the other four interviewer quintiles. If the lower than expected contact and cooperation rates achieved by these interviewers results in a greater degree of nonresponse bias on this variable, then the bottom quintile group are making a disproportionate contribution to total survey error.

This, however, is the pattern for only one question. An examination of the pattern across the full set of items in the online appendix (see online appendix 1) reveals that, for the contact measure, the bottom quintile group has the largest intra-interviewer correlation for 25 of the 36 items, of which 15 are significantly larger at the 95% level of confidence.[12] The corresponding figure for the cooperation measure is 20 items having the largest intra-interviewer correlation in the bottom quintile group, of which 13 are statistically significant differences compared

[11] It is not possible to obtain standard errors for the variance component estimates. To test for differences between quintile groups, we use chi square difference tests between nested models (Bollen, 1989).
[12] We refer here to the intra-interviewer correlations calculated using equation 3, as this enables straightforward comparisons between items.

to the remaining groups. So, although the same pattern is not observed across all 36 items, we find support for our hypothesis that there is an association between how successful interviewers are in gaining contact and cooperation and the magnitude of the within-interviewer error.

In terms of the direction of this association, the predominant pattern is for the interviewers who were least successful in obtaining contact and cooperation to have the largest error. Additionally, for the contact measure, there is evidence of a linear downward trend in the intra-interviewer correlation as we move from the least to the most successful interviewer quintile group; on 19 items, the most successful quintile group has the smallest correlation, of which 14 are statistically significantly smaller than the other groups. On 10 items, the bottom quintile group has the largest correlation *and* the top quintile group has the smallest. The intra-interviewer correlations for these 10 items are displayed as a line graph in Figure 2. Although there is some variability in the pattern across the intermediate groups, all 10 items show a large and consistent drop in the size of the correlation between the top and the bottom contact success groups.

FIGURE 2 HERE

To provide an indication of the average magnitude of these differences, Table 2 presents the mean intra-interviewer correlation for each success quintile across all 36 items, separately for the contact and cooperation measures. For the contact success measure, there is a clear downward gradient across the success quintiles, dropping from an average of 0.059 in the bottom group to 0.034 in the top quintile group. For the cooperation success measure, the bottom quintile group also has the largest average error by some margin, although the relationship could certainly not be characterized as linear. Indeed, there is some evidence of a curvilinear distribution across the success quintiles for the cooperation measure. This implies that in the case of cooperation, for some items, the least *and* the most successful interviewers are yielding the largest within-interviewer correlated errors.

TABLE 2 HERE

To explore this possibility further, we examined the extent to which items exhibited the smallest intra-interviewer correlation on the middle quintile group (quintile group 3) on the cooperation measure. In total, twelve items exhibited this curvilinear pattern and 8 of these were statistically significant differences ($p<0.05$). These 8 items are plotted in the line graph in Figure 3. The predominant pattern for these items is for the bottom quintile group to have the largest correlated error, for this to decline over the ensuing quintiles, before rising again in quintiles 4 and 5, although not quite to the same level as in the bottom group. This pattern suggests that, for a substantial minority of items, both of our initial speculations regarding the potential common causes of interviewer-induced nonresponse and measurement error find some support. On the one hand, interviewers who show poor performance in obtaining contact and cooperation also perform poorly with regard to the application of standardized interviewing. On the other, interviewers who surpass expectations in their level of success in persuading householders to participate in the survey also exhibit significantly larger correlated errors than interviewers who are less successful on this key dimension of interviewer performance.

FIGURE 3 HERE

Before discussing the implications of our findings, it is necessary to consider whether the differential variance components across the interviewer success quintiles are not due to interviewer behavior but to non-random allocation of respondents to interviewers and/or differential nonresponse across areas (Hox, 1994; West and Olson, 2010To do this, we compared distributions on a range of background variables across the five quintile groups for both contact and cooperation. This showed there to be no evidence of differential selection of respondent groups into the interviewer success quintiles on these observed variables (see online appendix 2). It is possible, of course, that there is still non-random selection into quintile groups on variables that we have not observed. As an additional check, therefore, we also compared the intra-interviewer correlations for our 36 in-scope items to those estimated on 7 additional questions, which were factual in nature and which required no probing or show-cards (Table 3). If ostensible effects arising from differences between interviewers are actually due to an uneven distribution of household characteristics across interviewers, we should anticipate this effect to be more or less constant across question types.

TABLE 3 HERE

In contrast to the items that involve probing and show-cards, the factual items with no interviewer involvement (beyond reading the question aloud) show considerably smaller within-interviewer correlated errors and no difference across quintile groups in the magnitude of the intra-interviewer correlation. We contend that this pattern of results makes the 'area-compositional confounding' account of the regularities we have shown here considerably less plausible and parsimonious than one based on the idiosyncratic behavior of interviewers during the administration of the questionnaire.

Discussion

The job of a survey interviewer comprises several different tasks, each of which requires a rather different set of aptitudes and skills in order to be implemented optimally. Our motivation in this paper has been to investigate the possibility that the sorts of interviewer characteristics which underpin success in contacting and persuading households to agree to a survey request might also be related to the way in which an interviewer administers the questionnaire, once an interview takes place. Put more succinctly, we evaluate whether there might be *common causes* of interviewer-induced nonresponse bias and measurement error (Groves, 2006; Groves and Lyberg, 2010; Olson, 2006).

Across a diverse range of questions we have found support for the idea that there *is* a link between an interviewer's level of success in obtaining interviews, on the one hand, and the degree of measurement error in the data they obtain, on the other. For both contact and cooperation, those interviewers with the lowest levels of success exhibited larger within-interviewer correlated error than their more successful colleagues. On some variables the interviewer variance component was more than 50% higher for these interviewers than for those enjoying greater levels of achievement. Unfortunately, we have no direct measure of nonresponse bias for any variable in the data set and cannot, therefore, calculate the mean squared error for particular estimates across the interviewer success groups. However, it is clear that where response rate and nonresponse bias are

correlated for particular variables in the data set, this group of interviewers makes a wholly disproportionate contribution to total survey error.

Although the predominant trend across the items examined was for the worst performing group of interviewers to have the largest variance component, there were also clear differences in the patterns observed for the contact and cooperation success measures, respectively. These differences are potentially informative about the nature of the underlying causal mechanisms. For contact, the majority of items showed a clear downward trend in the intra-interviewer correlations, with the worst interviewers having the largest correlated errors and the best interviewers the smallest. For cooperation, the pattern across items was more heterogeneous and the general trend was essentially non-linear. For a clear majority of items, the primary difference across the cooperation success quintiles was between the worst performing group and the rest, with no strong differences between the remaining 4 quintile groups. However, on a significant minority of items, a u-shaped pattern was evident across the groups, with the smallest variances in the middle group and the largest in the bottom and the top groups.

In our assessment, these differences in the pattern of effects across the contact and cooperation success measures are likely to reflect the fact that success in each domain is driven by a rather different set of underlying aptitudes and abilities. Maximising contact is essentially a matter of being well-organised, following best-practice guidelines, and persistence. If an interviewer is well-organised and conscientious when carrying out the part of their job that involves making contact with households, they are also likely to approach the survey interview in the same manner. Which is to say that they will adhere to the principles of standardized interviewing, in which they were trained. Thus, we should expect the relationship between contact success and the magnitude of interviewer variance to be approximately linear in nature, insofar as conscientiousness can be considered a continuous latent dimension.

Obtaining cooperation from reluctant respondents also requires conscientiousness and a willingness to follow training procedures and guidelines. However, it also calls for a diverse range of verbal and non-verbal skills, such as an ability to put people at ease and to maintain an interaction, to be knowledgeable about the content of the survey and to anticipate likely objections, based on the differing observable characteristics of householders

(Campanelli et al., 1997; Groves and Couper, 1998). And it is these latter characteristics which might also plausibly be the cause of a more 'conversational' style of interviewing. So, for the cooperation measure, we should expect to see larger intra-interviewer correlations at the bottom *and* at the top of the cooperation success distribution, which is exactly what we do observe on nearly a third of the items examined.

A clear limitation of our analysis for addressing these causal issues is that we do not actually have measurements of the putative 'common cause' variables at the interviewer-level but must infer their existence from the dependency that we do observe between nonresponse and measurement error. We know from the existing literature that interviewer beliefs, attitudes and behavior are predictive of nonresponse (Blom et al., 2010; Durrant and Steele, 2009; Hox and De Leeuw, 2002; Pickery et al., 2001) and measurement error (Singer et al., 1983; Olson and Peytchev, 2007), and we have speculated that traits such as conscientiousness and agreeableness (Digman, 1990) might plausibly account for the patterns of association we have found between these two sources of error. In future research, we will examine which interviewer characteristics are simultaneously predictive of *both* success in obtaining response and interviewer variance.

The heterogeneity of our results across items begs the obvious question of what it is about some items that makes them susceptible to a particular relationship with contact and cooperation success, while others are not. From our analysis of the characteristics of the 36 items, there is no obvious indication that the shape of the relationship varies as a function of the level of interviewer involvement required, the number or format of the response alternatives, or of any other identifiable characteristics of the items. All 36 items required either probing, the use of show-cards, or both so the nature of the relationship does not appear to depend, in any straightforward manner, on the extent of interviewer effort or involvement required. Again, we intend to address the underlying causes of this variability in future work, by increasing the pool of items under consideration and by including a broader range of item characteristics in the analysis.

Our findings have important implications for survey practice. While survey agencies have long focused on interviewer performance in achieving contact and cooperation as a means of improving survey quality, efforts to monitor how interviewers administer face-to-face questionnaires have been considerably less embedded in

standard practice. This imbalance undoubtedly reflects the relative difficulty and costs of monitoring performance in each area, with attempts to record and evaluate questionnaire administration on a routine basis facing a number of logistical barriers. Not least of these is the fact that monitoring how interviewers perform this task might very well alter the behavior it seeks to record. Our findings suggest that measures of success in obtaining contact and cooperation, of the kind we have developed here, can be used as diagnostic indicators of *both* of these key aspects of interviewer performance. By applying this method, it is possible to identify a group of interviewers who, via a programme of monitoring and training, could yield significant improvements in the overall quality of a survey, by increasing the rate at which they convert issued addresses into completed interviews and, crucially, by raising their adherence to standardized procedures of questionnaire administration.

An alternative explanation of our results is that, rather than interviewer behavior being the cause of the differential variances across quintile groups, what we are observing is some kind of selection effect: interviewers in the bottom success quintile are allocated, or end up achieving as a result of differential nonresponse, systematically different kinds of respondents to interviewers in the remaining quintile groups. We believe this is unlikely for several reasons. First, our pattern of results is very similar to that which has been found using interpenetrating sample designs on face-to-face surveys in the past. We find, on average, slightly more than half of the intra-class correlation due to clustering to be attributable to interviewers, and the remainder to areas. If our results are biased due to non-random allocation of sample units to interviewers, we should expect to see substantially larger interviewer variances than would be the case for an interpenetrating design. Second, and also in line with existing research based on random allocation of households to interviewers, we find systematic differences in the magnitude of interviewer variances as a function of question characteristics, with questions which require higher levels of input from interviewers being the most susceptible and factual questions with no interviewer involvement (beyond reading the question aloud) showing little or no interviewer variance component (Schnell and Kreuter 2005). Neither do we find any evidence of distributional differences across a range of demographic characteristics on either the contact and cooperation measure of interviewer success.

It is difficult to see how this pattern of effects can easily be reconciled with an account based on compositional bias across achieved interviewer workloads. In contrast, it fits parsimoniously with the idea that, within the total

pool of interviewers working on a survey, there is a minority who achieve a poor level of performance in both of these key aspects of their work. The research reported here provides a first insight into the joint impact of interviewers on nonresponse and measurement error and suggests a number of ways in which the approach we have set out can be used to both further our understanding of the how interviewers contribute to survey error and to improve survey practice.

References

Bates, Douglas, and Martin Maechler. 2010. Package 'lme4'. Version 0.999375-35 (http://lme4.r-forge.r-project.org/).

Biember, Paul. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly*, 74(5) 817-848.

Biemer, Paul, and Lyberg, Lars. 2003. *Introduction to Survey Quality*. Wiley Series in Survey Methodology. Hoboken, NJ: John Wiley & Sons, Inc.

Blom, Annelies G., Edith D. de Leeuw and Joop J. Hox. 2010. "Interviewer Effects on Nonresponse in the European Social Survey." ISER working paper series.

Bollen, Kenneth. 1989. *Structural Equations with Latent Variables.* Wiley Series in Probability and Mathematical Statistics. New York: Wiley & Sons, Inc.

Campanelli, Pamella, Patrick Sturgis, and Susan Purdon. 1997. *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates.* London: S.C.P.R.

Cannell, Charles. F., Peter V. Miller, and Lois Oksenberg. 1981. "Research on Interviewing Techniques*."* In (ed.) Samuel Leinhardt, *Sociological Methodology*. San Francisco: Jossy-Bass: 389-437

Collins, Martin, and Bob Butcher. 1982. "Interviewer and Clustering Effects in an Attitude Survey." *Journal of the Market Research Society*, 25(1): 39-58.

Couper, Mick P., and Robert M. Groves. 1992. "The role of the interviewer in survey participation." *Survey Methodology*, 18(2): 263-278.

de Leeuw, Edith, Joop J. Hox, Snijkers, G., and de Heer, W. 1998, *Interviewer Opinions, Attitudes and Strategies Regarding Survey Participation and Their Effect on Response.* ZUMA Nachrichten Spezial, 4: 239-248.

Digman, John. 1990. “Personality structure: Emergence of the five-factor model.” *Annual Review of Psychology*, 41, 417–440.

Durrant, Gabriele, and Fiona Steele. 2009. “Multilevel Modelling of Refusal and Noncontact Nonresponse in Household Surveys: Evidence from Six UK Government Surveys.” *Journal of the Royal Statistical Society*. Series A. 172(2): 1-21.

Durrant, Gabriele, Robert M. Groves, Laura Staetsky, and Fiona Steele. 2010. “Effects of interviewer Attitudes and Behaviors on Refusal in Household Surveys.” *Public Opinion Quarterly*, 74(1): 1-36.

Fricker, Scott and Tourangeau, Roger. 2010 “Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys.” *Public Opinion Quarterly*, 74(5), 934-955.

Fowler, Floyd J., and Thomas W. Mangione. .1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* Newbury Park: Sage.

Freeman, John, and Edgar W. Butler. 1976. “Some Sources of Interviewer Variance in Surveys.” *Public Opinion Quarterly.* 40(1): 79-91.

Goldstein, Harvey. 2003. *Multilevel Statistical Models*. (3 ed.) London: Arnold.

Grant, Catherine, Keith Bolling, and Matthew Sexton. 2006. *2005-6 British Crime Survey (England and Wales): Technical Report Volume 1.* Research Development and Statistics. Home Office.

Groves, Robert M. 1989. *Survey Errors and Survey Costs.* New York: John Wiley & Sons.

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

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley.

Groves, Robert M., Mick P. Couper, Stanley Presser, Eleanor Singer, Roger Tourangeau, Giorgina Piani Acosta, and Lindsay Nelson. 2006. “Experiments in Producing Nonresponse Bias.” *Public Opinion Quarterly*. 70(5): 720-736.

Groves, Robert, M. and Lyberg, Lars. 2010. “Total Survey Error: Past, Present, and Future.” *Public Opinion Quarterly* 74(5), 849-879.

Groves, Robert M. and McGonagale, Kate. 2001. “A Theory-Guided Interviewer Training Protocol Regarding Survey Participation.” *Journal of Official Statistics*, 17: 249-266.

Groves, Robert M., and Emilia Peytcheva. 2008. “The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis.” *Public Opinion Quarterly*. 72(2): 167-189.

Hirschman, Albert O. 1964. “The Paternity of an Index.” *The American Economic Review* 54(5): 761.

Hox, Joop J. 1994. “Hierarchical Regression Models for Interviewer and Respondent Effects.” *Sociological Methods & Research* 22(3): 300-318.

Hox, Joop J. 2010. *Multilevel Analysis: Techniques and Applications*. Second edition. Routledge.

Hox, Joop J., and Edith D. de Leeuw. 2002. “The Influence of Interviewers’ Attitude and Behavior on Household Survey Nonresponse: An International Comparison.” In (Eds.) Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J.A. Little. *Survey Nonresponse*. New York: John Wiley and Sons, Inc. 103-119.

Kaiser, Henry F. 1970. “A Second Generation Little Jiffy.” *Psychometrika* 35: 401-417.

Kass, Gordon V. 1980. “An Exploratory Technique for Investigating Large Quantities of Categorical Data.” *Applied Statistics*. 29(2): 119–127.

Kish, Leslie. 1962. “Studies of Interviewer Variance for Attitudinal Variables.” *Journal of the American Statistical Association,* 57(297): 92–115.

Kominska, Olena, McCutcheon, Alan, and Billiet Jaak. 2010. “Satisficing Among Reluctant Respondents in a Cross-National Context.” *Public Opinion Quarterly*, 74(5), 956-984.

Kreuter, Frauke., Muller, Gerrit., and Thompson, Mark. 2010. “Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data.” *Public Opinion Quarterly*, 74(5), 880-906.

Krosnick, Jon. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology*. 5: 213-36.

Mahalanobis, Prasanta Chandra. 1946. “Recent Experiments in Statistical Sampling in the Indian Statistical Institute.” *Journal of the Royal Statistics Society*, 109: 325-378.

Mangione, Thomas, Floyd J. Fowler and Thomas A. Louis. 1992. “Question Characteristics and Interviewer Effects.” *Journal of Official Statistics*. 8(3): 293-307.

Marquis, Kent H., and Charles F. Cannell. 1969. *A Study of Interviewer-Respondent Interaction in the Urban Employment Surveys*. Michigan: Ann Arbor, Institute for Research.

Martin, David. 2001. *Geography for the 2001 Census in England and Wales*. ONS.

Morton-Williams, Jean. 1993. *Interviewer Approaches*. Aldershot, Dartmouth: Ashgate

O'Muircheartaigh, Colm. 1976. "Response Errors in an Attitudinal Sample Survey." *Quality and Quantity*, 26: 115.

O'Muircheartaigh, Colm, and Pamela Campanelli. 1998. "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision." *Journal of the Royal Statistics Society A*, 161(1): 63-77.

Olson, Kristen. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly*, 70(5): 737-758.

Olson, Kristen, and Andy Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly*. 71(2): 273-286.

Pickery, Jan, Geert Loosveldt, and Ann Carton. 2001. "The Effects of Interviewer and Respondent Characteristics on Response Behaviour in Panel Surveys: A Multilevel Approach." *Sociological Methods and Research*, 29(4): 509-523.

Rasbash, Jon, and Harvey Goldstein. 1994. "Efficient Analysis of Mixed Hierarchical and Cross-Classified Random Structures Using a Multilevel Model." *Journal of Educational and Behavioral Statistics*. 19(4): 337-350.

Sakshaug, Joseph W, Ting Yang, and Roger Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-mode Survey of Sensitive and Non-sensitive Items." *Public Opinion Quarterly*. 74(5): 907-933.

Schnell, Rainer, and Frauke Kreuter. 2005. "Separating Interviewer and Sampling Point Effects." *Journal of Official Statistics*. 21(3): 389-410.

Simmons, Eleanor, and Amanda Wilmot. 2004. "Incentive Payments on Social Surveys: A Literature Review." *Survey Methodology Bulletin.* 53. 1/04. London. ONS.

Singer, Eleanor, Martin R. Frankel, and Marc B. Glassman. 1983. "The Effect of Interviewer Characteristics and Expectations on Response." *Public Opinion Quarterly*. 47(1): 68-83.

Snijkers, Ger, Joop J. Hox, and Edith D. de Leeuw. 1999. "Interviewers' Tactics for Fighting Survey Nonresponse." *Journal of Official Statistics*, 15(2): 185-198.

Sturgis, Patrick, and Pamela Campanelli. 1998. "The Scope for Reducing Refusals in Household Surveys: An Investigation Based on Transcripts of Tape-recorded Doorstep Interactions." *Journal of the Market Research Society* 40(2): 121–39.

Tourangeau, Roger, Robert M. Groves, and Cleo D. Redline. 2010. “Sensitive Topics and Reluctant Respondents. Demonstrating a Link between Nonresponse Bias and Measurement Error.” *Public Opinion Quarterly*. 74(3): 413-432.

West, Brady T. and Olson, Kirsten. 2010. “How Much of Interviewer Variance is Really Nonresponse Error Variance?” *Public Opinion Quarterly*, 74(5), 1004-1026.

Numbered List of Figure Captions

FIGURE 1: A Path Model Showing how a Negative Correlation between Interviewer-induced Nonresponse Bias and Measurement Error Might Arise

FIGURE 2: Items with a Downward Trend Association between Contact Success and Interviewer Variance

FIGURE 3: Items with a U-Shaped Association between Cooperation Success and Interviewer Variance

Table 1. Coefficient estimates for 'Overall rating of health' (higher score = less healthy)

		Estimate (S.E)
FIXED EFFECTS		
	Constant	2.08**
		(0.01)
Respondent level	Male	0.04**
		(0.01)
	Age	0.27**
		(0.005)
	Nonwhite	-0.003
		(0.02)
	Education (contrast: *No qualifications*): GCSE	-0.20**
		(0.01)

	A level	-0.26**
		(0.01)
	Degree	-0.36**
		(0.01)
	Non-traditional/ foreign qualification	-0.16**
		(0.02)
Neighborhood level	Socio-economic disadvantage	0.10**
		(0.005)
	Urbanicity	0.03
		(0.01)
	Population mobility	0.02**
		(0.01)
	Age profile	0.01*
		(0.005)
	Housing structure	0.02**
		(0.01)
	Ethnic diversity	0.05
		(0.05)
Interviewer level	Male	0.04*
		(0.02)
	Age	-0.01
		(0.01)
	Nonwhite	-0.06
		(0.05)
	Experience (months working)	0.003
		(0.01)

RANDOM EFFECTS		**Contact ratio**	**Cooperation ratio**
Top success quintile	$\sigma^2_{u1(j_1)}$	0.014	0.015
	$\sigma^2_{u2(j_1)}$	0.020	0.020
	$\sigma^2_{u3(j_1)}$	0.017	0.018
	$\sigma^2_{u4(j_1)}$	0.023	0.017
Bottom success quintile	$\sigma^2_{u5(j_1)}$	0.033	0.035
Area	$\sigma^2_{u0(j_2)}$	0.003	0.003
Individual	σ^2_e	0.690	0.690

** P<(.01)

* P<(.05)

TABLE 2 Mean Intra-interviewer correlations across 36 items by contact and cooperation success quintiles

	Least successful quintile				Most successful quintile
Contact	0.059	0.048	0.048	0.041	0.034
Cooperation	0.059	0.041	0.039	0.046	0.044

TABLE 3 Mean Intra-Interviewer correlations by contact and cooperation success quintiles for items requiring and not requiring interviewer effort

	Least successful quintile				Most successful quintile
Contact					
No interviewer effort items (n=7)	0.007	0.007	0.005	0.007	0.005
Interviewer effort items (n=36)	0.059	0.048	0.048	0.041	0.034
Cooperation					
No interviewer effort items (n=7)	0.005	0.007	0.005	0.006	0.008
Interviewer effort items (n=36)	0.059	0.041	0.039	0.047	0.045