ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)
181, *Part* 1, *pp.* 181–203

Can conversational interviewing improve survey response quality without increasing interviewer effects?

Brady T. West and Frederick G. Conrad,

University of Michigan, Ann Arbor, USA

Frauke Kreuter

University of Maryland, College Park, USA, University of Mannheim and Institute for Employment Research, Nuremberg, Germany

and Felicitas Mittereder

University of Michigan, Ann Arbor, USA

[Received January 2016. Revised October 2016]

Summary. Several studies have shown that conversational interviewing (CI) reduces response bias for complex survey questions relative to standardized interviewing. However, no studies have addressed concerns about whether CI increases intra-interviewer correlations (IICs) in the responses collected, which could negatively impact the overall quality of survey estimates. The paper reports the results of an experimental investigation addressing this question in a national face-to-face survey. We find that CI improves response quality, as in previous studies, without substantially or frequently increasing IICs. Furthermore, any slight increases in the IICs do not offset the reduced bias in survey estimates engendered by CI.

Keywords: Conversational interviewing; Interviewer effects; Intra-interviewer correlation; Multilevel modelling; Standardized interviewing; Survey paradata

1. Introduction

Interviewer-administered survey data collection is generally performed by using one of two interviewing techniques. One is known as standardized interviewing (SI), where survey interviewers are instructed to read questions exactly as worded and to provide only neutral or non-directive probes in response to questions from respondents (Henson *et al.*, 1976; Groves and Magilavy, 1986; Fowler and Mangione, 1990; Mangione *et al.*, 1992; Belli and Lepkowski, 1996). A second is known as conversational interviewing (CI), where survey interviewers are trained to read questions exactly as worded, initially, and then to respond to respondents' questions or evidence of confusion by providing definitions of key terms (possibly in their own words, assuming that they have demonstrated mastery of the concepts during training) or whatever other information is required to assure that respondents understand the questions as they are intended (e.g. Schober and Conrad (1997)).

Address for correspondence: Brady T. West, Michigan Program in Survey Methodology, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106-1248, USA.
E-mail: bwest@umich.edu

 0964–1998/18/181181

The survey interviewing literature includes several studies demonstrating that CI reduces the bias in survey responses relative to SI (Schober and Conrad, 1997; Conrad and Schober, 2000; Schober *et al*., 2004, 2012; Hubbard *et al*., 2012; Conrad *et al*., 2015; Bruckmeier *et al*., 2015). These studies have shown that, although CI can lead to longer interviews (potentially increasing costs), it can also improve respondents' comprehension of terms in survey questions that may be ambiguous with respect to the situations of particular respondents. For example, although a respondent who works on her family's farm in exchange for meals and a place to live might know what 'work for pay' means in general, she may be uncertain how to answer a question about whether she receives pay in exchange for work. A conversational interviewer can work with the respondent, explaining that the concept requires the worker to be compensated monetarily. More specifically, these studies have demonstrated that respondent answers to factual survey questions elicited via CI are more consistent with definitions of terms that feature *exclusive* concepts (i.e. the respondent should *not* count certain activities or states when determining how to answer; for example income does not include lottery winnings) or *inclusive* concepts (i.e. the respondent *should* consider certain activities or states when determining how to answer; for example income includes overtime payments). We expect CI to operate in a similar fashion in the present study.

Despite the demonstrated ability of CI to increase the accuracy of survey responses to complex factual questions, the increased flexibility that is granted to interviewers using CI introduces the risk of this technique increasing the intra-interviewer correlation (IIC) in the survey responses. Formally, the IIC for a particular survey item is defined as

$$\rho_{\text{int}} = \frac{\tau^2}{\tau^2 + \sigma^2}, \tag{1}$$

where τ^2 is the variability *between* interviewers in the means of a given item (generally, the variance of a random interviewer effect, and usually a random interviewer intercept, included in a multilevel statistical model), and σ^2 is the residual variability *within* interviewers in a particular measure of interest. Although we use general notation in equation (1), one can fit models allowing both variance components (and thus the IIC) to be a function of other factors (e.g. interviewer characteristics) or even the interviewer (in the case of σ^2), depending on the model that is used to estimate these two components of variance (Brunton-Smith *et al*., 2017). Section 4.2 contains additional details on this approach.

This IIC reduces the efficiency of survey estimates in a manner that is similar to cluster sampling (Brick *et al*., 1995; O'Muircheartaigh and Campanelli, 1998; Groves, 2004; Schnell and Kreuter, 2005). One can define a multiplicative 'interviewer effect' on the variance of an estimator of the mean for a particular survey item as $1 + (\bar{m} - 1)\rho_{\text{int}}$, where $\bar{m}$ corresponds to the average number of interviews completed across all interviewers. Thus, given an estimated IIC of 0.02 and an average of 30 interviews completed by each interviewer, we would expect the variance of the estimator of a mean to be inflated by 58%. The literature on interviewer effects has not benefitted from any rigorous studies of the differences in the IICs introduced by CI and SI, and whether, when considering the overall mean-squared error (MSE) of survey estimates, any increase in the IIC due to CI offsets the reductions in response bias that it has been shown to provide.

By analysing the results of a randomized field experiment implemented in a national survey in Germany, this paper seeks to fill these gaps in what is known about the effects of CI and SI on the overall quality of survey estimates. Because interviewer error is a property of a specific question and not an overall survey, we conduct rigorous comparisons of the IICs that are introduced by each technique for several factual survey items (in addition to a small number of attitudinal items) and assess the effects of differential IICs on the overall quality of multiple

survey estimates derived from those items. Overall, we find that CI does not tend to increase the IIC substantially relative to SI and, when it does, increases in the IIC do not tend to offset reductions in response bias.

The remainder of the paper is structured as follows. Section 2 provides a brief review of the existing literature comparing CI and SI in terms of data quality and considers potential sources of increased IICs due to CI. Section 3 describes the design of the field experiment. Section 4 details the statistical analyses that are used to compare the two techniques in terms of IICs and other measures of data quality, and Section 5 presents the results of the analyses. Section 6 provides a summary of the findings, and concludes with implications for practice and directions for future research in this area.

2. Background

In theory, SI requires that all interviewers use the same or very similar wording each time that they ask a survey question, including provision of the response options. This methodology is designed to reduce the variability between interviewers in interviewer-related measurement errors (Cannell and Axelrod, 1956; Kahn and Cannell, 1957; Cannell *et al*., 1975; Fowler and Mangione, 1989). By exposing respondents to the same question wording and response options, irrespectively of which interviewers actually ask the questions, respondents' answers should be comparable across interviewers (Fowler and Mangione, 1989). There is actually little, if any, empirical evidence (one way or the other) about how SI affects IICs compared with alternative interviewing techniques. Interviewers who are trained to use SI do occasionally deviate from strict standardization, either by mistake (Houtkoop-Steenstra, 1995) or to simplify potentially complex interactions with respondents (Peneff, 1988; Haan *et al*., 2013; Ackermann-Piek and Massing, 2014; Bell *et al*., 2016). These deviations, whatever the cause, introduce the possibility of inflated IICs when some interviewers deviate more often than others. Moreover, there is evidence that SI may interfere with accurate question comprehension (and thus accurate responses) because it does not permit *conversational grounding*, or the everyday interactional process that speakers and listeners use to ensure that they understand each other (e.g. Clark and Brennan (1991); for applications of grounding to interaction in interviews, see Schober and Conrad (1997) and Suchman and Jordan (1990)).

Conversational grounding is central to the practice of CI. When speakers and listeners engage in grounding, they talk about what has been said to be sure that they understand each other sufficiently well for the purposes of the conversation. Interviewers are given discretion to choose their words when communicating the intentions behind the question, but the overall philosophy of promoting comparability across responses is the same as it is in SI. What differs is the focus on standardizing question *interpretation*—in particular how respondents' individual circumstances correspond to the concepts in the question—rather than on standardizing *wording*. We note that CI as defined here is distinct from more 'personal' interviewing styles that have been discussed in the literature, which have also been shown to improve response accuracy (e.g. Dijkstra (1987) and van der Zouwen *et al*. (1991)).

What might lead to higher IICs in CI rather than SI, reducing the precision of survey estimates based on the answers collected by using CI? The additional flexibility that is granted to interviewers in CI may lead to uneven implementation of the technique (i.e. some interviewers might provide more clarification than others or go off on tangents, resulting in longer interviews (Schober *et al*., 2004)), or more variation in the actual question wording that is used. If conversational interviewers consistently differ in the meanings of question concepts that they provide to respondents, then they may collect different responses from those which would be expected

with SI. Researchers have suggested that non-zero IICs in SI arise due to interviewer-specific deviations from the types of neutral probes that are generally called for by the SI approach, i.e. IICs will tend to be higher for questions requiring more probing (Mangione *et al.*, 1992). When respondents ask *conversational* interviewers for clarification, the interviewers must decide whether they should provide complete definitions or just the most relevant parts, and whether they should provide the gist of the definition or provide it *verbatim*. They also must exercise substantially more discretion when deciding whether to volunteer clarification if they believe that the respondent has misunderstood the question. Although this kind of judgement is common in everyday conversation, people—including interviewers—vary in the accuracy of their judgements about how well respondents have understood a question (Hubbard *et al.*, 2012). This could lead to inconsistencies across interviewers in how and when clarification is provided, which can, in turn, affect the distributions of answers that are elicited by different interviewers.

The key question is thus how much this potential variation in clarification behaviour increases variability in the answers that are collected. To date, only one partially related study has attempted to evaluate differences in IICs introduced by more- and less-structured interviewing approaches. Sayles *et al.* (2010) found evidence of modest increases in IICs when comparing the event history calendar approach (which is less scripted than SI but is not CI, at least in our sense of promoting grounding through clarification) to SI. No other study has rigorously examined the differences in IICs between CI and SI (and the implications of these differences for the overall quality of survey estimates) by using a well-powered experimental design in a real field setting (see Conrad and Schober (2000) for discussion).

3. Experimental design

3.1. Sampling

For this study, we sampled people from the 'Integrated employment biographies' (IEB) database, managed by the Institute for Employment Research in Nuremberg, Germany. The IEB database contains official government information regarding employment spells at the person level (including current addresses for the people) for all people who have been employed in Germany. People who are eligible for sampling needed to have

(a) been employed on December 31st, 2012, and have held at least two different jobs in the previous 10 years (dating back to December 31st, 2002),
(b) had at least one spell of unemployment, no matter how long,
(c) held at least one part-time job and
(d) been at least age 18 years.

These four eligibility criteria ensured that eligible respondents could potentially benefit from additional interviewer clarification related to questions about their employment history. The presence of official administrative records on this sampling frame allowed us to compare survey estimates based on respondents interviewed by using each technique (SI and CI) to estimates based on information from the IEB database.

Before sampling, we worked with the German data collection organization the Institut für angewandte Sozialwissenschaft (known as 'infas') (`www.infas.de`) to determine 15 areas in Germany where it would be feasible for four professional interviewers per area—who had prior experience working on infas-related projects—to recruit and interview randomly sampled people. These 15 areas, which were specifically (and not randomly) selected because infas knew of several interviewers in each area who would be willing and able to work on this project, included Bad Homburg Oberursel, Berlin, Bremen, Dortmund, Dresden, Essen, Freiburg,

Hamburg, Hannover, Heidelberg, Leipzig, Neumarkt, Neuwied Andernach, Rostock and Tubingen. Within each of these 15 areas, infas staff identified postal codes located close to the homes of the four available interviewers where it would be feasible for the interviewers to access sampled addresses and to conduct face-to-face interviews easily.

From a list of addresses of people living in the identified postal codes within each of the 15 areas and meeting the four eligibility criteria above, we then selected a simple random sample of 480 addresses (i.e. four interviewer workloads of 120 addresses per area; see Section 4.1 for power analysis). This sample of 480 addresses was selected from each *area*, and not each *postal code*, and the sample of size 480 was not allocated in any specific fashion to the identified postal codes within an area. As a result, some postal codes may have contributed more sampled addresses than others given the simple random sampling. Finally, it was not possible for the same person to be included twice in these initial samples; the list of 480 sampled addresses was randomly ordered and divided into four subsamples of 120 addresses each, and these subsamples were then assigned to each of the four interviewers in an area.

The infas staff then checked the addresses of the randomly sampled people who were assigned to each of the four interviewers to make sure that the interviewers would be able to access each address that was assigned to them and to conduct interviews plausibly at those addresses. Given knowledge about the transportation capabilities and experience of the 60 interviewers, infas decided to drop some of the sampled addresses as a result of this process (e.g. the travel time between a single address and another sampled address in adjacent postal codes would not be realistic for a given interviewer), resulting in the interviewers from certain areas having fewer than 120 assigned people in their samples. We therefore selected an additional simple random sample of total size $n = 7200$ across the 15 areas (again, 480 per area), and then randomly ordered the additional 480 sampled cases within a particular area. Each interviewer without 120 assigned people then had the required number of supplemental cases from the beginning of this randomly ordered list assigned to them (e.g. one interviewer needing 10 more cases received the first 10 cases, the next needing five more cases received the next five cases; and so forth). If duplicate cases were identified (i.e. a sampled address had already been assigned to an interviewer), the next case in the randomly ordered list was assigned to that interviewer.

After this process, three of the 15 areas had interviewers who still did not have sufficient sample sizes (120 addresses). We therefore selected a third set of simple random samples in these three particular areas and followed the same process. Although this process did therefore result in unequal probabilities of inclusion for different people in the resulting gross sample of 7200 addresses, our objective was not to make design-based inferences about the entire finite populations of these 15 areas, but rather to have a balanced, well-powered, randomized experimental design across the 15 areas (the details are below). Accordingly, we followed a model-based approach in all our analyses.

3.2. Interviewer randomization and training

We identified four professional interviewers within each of the 15 areas who had past experience working on infas projects and could easily travel to the sampled addresses within a given area for the current project. After these 60 interviewers were identified, two interviewers within an area were assigned at random to the CI condition, and the other two interviewers were assigned to the SI condition. To ensure that interviewers in the two experimental groups did not systematically differ in terms of their experience with the survey topic or any other characteristics that are known to introduce interviewer effects (such as gender, age, race and overall interviewing experience; see Schaeffer *et al.* (2010)), the two groups of interviewers were then compared across areas on prior experience working on a closely related study,

gender and age. No significant differences were found and the 60 interviewers were predominantly white, suggesting that the two experimental groups were balanced in terms of these features.

Most of the interviewers (57) were trained in a 1-day session in March 2014 led by our research team and infas staff. Three interviewers who could not attend the initial training were given the same face-to-face training at a later date. This training session first introduced all the interviewers to the objectives of the study and the concepts that were going to be measured on the questionnaire. Interviewers were then divided into two groups according to the technique to which they had been assigned, and they were rigorously trained in the details of using that technique. These group-specific sessions involved instruction about the survey concepts and their definitions, in addition to role playing exercises, examples and question-and-answer sessions. Each group-specific session concluded with a multiple-choice quiz, presenting hypothetical examples of questions that respondents might raise about the concepts in particular items, and asking the interviewer to choose the best way to clarify the respondent's confusion consistently with how the concepts were defined in the current study. Both groups of interviewers did quite well in general on the 'concepts' test, each averaging 81% (25/31 items) correct. Concepts that interviewers consistently misunderstood were reviewed in a final debriefing session at the end of the day. The training sessions were staggered so that each group was trained by the same set of instructors (while the other group was working on the test or engaged in group discussions).

After being trained in a particular interviewing technique, the 60 interviewers were assigned their sets of 120 sampled addresses, along with all available contact information (including telephone numbers, if available) for these addresses. Before the onset of data collection, an advance letter was mailed to all the sampled people, alerting them to the objective and importance of this study, in addition to making them aware of the incentive for participating (€20). The interviewers then proceeded with initial contact attempts at the sampled addresses and attempted to convince the sampled people to participate in the face-to-face interview. The proportions of sampled people agreeing to participate were ultimately quite similar in the two interviewing conditions (24.9% in SI, and 24.3% in CI). During the conversational interviews, the definitions were displayed with the question text on the screen for the interviewer to see. These interviewers were instructed to use these definitions, either *verbatim* or by paraphrasing the relevant parts, when assisting respondents with any comprehension or clarification issues.

3.3. Questionnaire design and data collection

In line with existing CI research, the questionnaire that was used for this study was composed primarily of factual questions, grouped into four parts: questions on housing and living conditions, questions capturing the employment status history of the respondent (including current employment questions, such as monthly income, and measures of job satisfaction), questions on social networks and demographic questions. The majority of the questions were taken either from previous studies designed to evaluate CI techniques (e.g. Conrad and Schober (2000)), or from the German Panel Arbeitsmarkt und soziale Sichering (PASS) ('Labour market and social security') study, which is an annual household panel survey for labour market, welfare state and poverty research commissioned by the Institute for Employment Research (Trappmann *et al.*, 2013). When selecting questions for the current study, we purposely used as many 'existing' questions as possible to increase relevance of the findings and to build on prior experience related to the performance of these items. We also wanted to make sure that the survey items had been tested in real field studies. The full questionnaire, in German and English formats and including both the response options and

CI definitions, can be downloaded via `http://doku.iab.de/fragebogen/CIIV_Questionnaire_English_08052015.pdf` or `http://doku.iab.de/fragebogen/CIIV_Questionnaire_German_V13_08052015.pdf`. All completed interviews were conducted in German.

The questions that were related to housing and living conditions (borrowed from Conrad and Schober (2000)) were translated into German and adapted to the current study. Each of these questions asks about elements of the house or around the house, e.g. 'In the past five years, since January 1, 2009, have you had moving expenses?'. Interviewers in the CI group were trained to clarify the meaning of terms like 'moving expenses' in the case that respondents asked for clarification regarding these terms, e.g. 'Should I include do-it-yourself moving?'.

For the questions that were related to employment history, social networks and demographics, we identified existing questions that

(a) had validation information available in the IEB database,
(b) have historically resulted in more interviewer–respondent interaction and requests for clarification, according to Institute for Employment Research researchers, or
(c) showed larger interviewer effects in prior PASS waves.

For this study, PASS questions were modified so that definitions that were provided to all PASS interviewers were removed from the question display and available only to the conversational interviewers. Existing questions from the PASS survey that can be compared with variables that are available in the IEB administrative records are generally items related to employment spells (beginning and end), gross income, unemployment spells (beginning and end), welfare benefits, age, nationality and education.

For the measures of most recent gross monthly income (which dated back to January 1st, 2013, and excluded any marginal or 'short-term contract' jobs) and gross annual income in 2013 (which included marginal jobs but were only generated from dependent employment), we needed to account for the fact that unemployed individuals may have had an income of 0 (recall that people had to be employed on December 31st, 2012, but the survey data were collected from March to October in 2014), and that the values of income are *top coded* in the IEB administrative data: €4000 maximum for gross monthly income, and €50 000 maximum for 2013 annual income. As a result of both unemployment and the administrative top coding, distributional assumptions underlying any models for the continuous income values (e.g. normality of residuals) may be difficult to satisfy, and means in the survey responses could be artificially larger than means in the administrative data. We therefore generated categorical versions of the income measures in both the administrative data and the survey data. For monthly income, the four categories were €0, €1–1787.60 (the median of the administrative values that were non-zero and not top coded), €1787.61–3999.99 and €4000.00 or more. For annual income, the four categories were €0–1000.00, €1000.01–17 680.52 (the median of the administrative values that were non-zero and not top coded), €17 680.53–49 999.99, and €50 000 or more. Any missing values in the administrative or survey data were coded into the respective first categories.

Measures on all of these variables (aside from education) have been found to be quite accurate in the IEB records (Jacobebbinghaus and Seth, 2007). Most of the retrospective work history questions were asked with a time frame since January 1st, 2013, because administrative data are usually readily available for the prior calendar year. In addition, respondents were asked about the industry and the size of the company by which they are currently employed (or have been employed previously, for those currently not employed).

Even though no validation is possible for attitudinal survey items, we included job satisfaction questions in the employment module, to estimate IICs for these types of question for each of the

two experimental groups. These questions were combined with questions about task difficulties. Here interviewers were again given definitions during the training session and on the screen in the CI condition. Interviewers were also asked to record four post-survey observations (with four ordered categories each) describing the quality of the respondent's response behaviour.

Data collection continued from the last week of March until the first week of October in 2014. All interviews were conducted by using computer-assisted personal interviewing software, and took about 35 min to complete on average. Respondents were promised a €20 token of appreciation delivered via postal mail within 2 weeks after completing the interview. By the end of data collection, a total of 1850 interviews had been completed by the 60 interviewers (American Association for Public Opinion Research RR1 = 25.7%). Although individual interviewers using each technique did vary in terms of their response rates, the two groups of interviewers did not vary significantly in the aggregate. Additional investigation of the factors affecting this interviewer variability in response rates and the implications of that variance for the overall IICs in the responses assessed here is certainly a direction for future research but is outside the scope of this study; we revisit this point in Section 6. We also note that this study was distinct from other national labour market surveys in Europe, in that the target population included only people who were currently employed on December 31st, 2012; employed individuals are generally more difficult to reach. In addition, the incentive (€20) is not extremely large for employed people. Furthermore, given our objective of estimating IICs, we did not allow multiple interviewers to work on the same case, so case transferring was not possible. Taken together, these design features may have lowered response rates overall compared with other national surveys on similar topics.

3.4. Audio recording and interviewer evaluation

To enable analyses of the implementation and effects of SI and CI in more detail, the interviewers asked people agreeing to participate in the survey for permission to audio-record the entire interview. All the recordings that we analysed contained the respondent's affirmative consent to be recorded and did not include any identifying information. The infas supervisors provided the interviewers with biweekly feedback on their performance based on these recordings and reported that the performance of selected interviewers improved (i.e. the interviewers became better at using their assigned technique) after receiving this feedback. Overall, when analysing a subsample of these recordings in detail for interviewers in each of the two groups, we found that the CI and SI techniques had been implemented correctly and consistently (Mittereder *et al.*, 2017).

4. Statistical analysis

4.1. Power analysis

Given the specific objectives of this study, we wanted to ensure that we would have enough power to detect *differences in the variance components that were used to compute IICs* between standardized and conversational interviewers as being significant. We therefore performed a customized Monte Carlo simulation study given these objectives (see the SAS code that is available from `http://wileyonlinelibrary.com/journal/rss-datasets`). A review of the literature on interviewer effects (see West and Olson (2010) or Schnell and Kreuter (2005)) suggests that most IICs will range from 0.01 to 0.12 in face-to-face surveys, with many falling below 0.02. Furthermore, in our recent work analysing data on survey items from the face-to-face PASS study in Germany (which are similar to those that we included in our questionnaire;

see West *et al*. (2013)), we found that these IICs ranged from below 0.01 to approximately 0.09 (where an IIC of 0.09 would *quadruple* the variance of an estimate, reducing the effective sample size by 75%).

We therefore used these earlier results to perform power calculations for this study, ensuring that we would have enough power to detect differences of this magnitude for both a continuous survey measure (e.g. the longest period of uninterrupted employment in the past 20 years) and a binary survey measure (e.g. receipt of unemployment benefits). In the simulation studies, we found that having 30 interviewers measuring a continuous item for each of the two techniques (60 interviewers total) and 30 respondents for each interviewer (or 1800 respondents in total) would yield approximately 80% power to detect a 6.6-fold difference in the between-interviewer variance components that were used to compute the IIC in equation (1) as significant, based on a likelihood ratio test (West and Elliott, 2014) with a 5% level of significance. A difference of this size in the variance components falls within the aforementioned range of IICs (from 0.01 to 0.09) that we have seen in related studies with similar subject matter. Furthermore, we found that having 30 interviewers in each of the two experimental groups measuring a binary item and 30 respondents for each interviewer would yield approximately 82% power to detect a similar difference in the between-interviewer variance components, again by using a likelihood ratio test with a 5% level of significance. We therefore based our data collection protocol on meeting these targets.

4.2. Statistical modelling

We fit multilevel statistical models to the data that were collected for each of 55 survey items, specifically focusing on estimates of the following parameters:

(a) the IIC for each technique (1), describing the within-interviewer correlation in the survey reports on a given item; and
(b) the fixed effect of the CI technique (relative to SI), controlling for fixed area effects, which captures the shift in the response distribution for a given variable that is associated with the use of CI and allows us to assess the consistency of our results with the prior literature that has focused on response bias only.

Estimating the IICs for each item–technique combination required estimation of the between- and within-interviewer variance in the measures for a particular survey item introduced by a given technique (1). We used the likelihood ratio testing approach of West and Elliott (2014) to test formally whether these variance components were significantly different from each other, and we also tested whether the fixed CI effect was different from 0.

For continuous survey items (including measures of interview duration), we estimated these parameters by fitting the following model using restricted maximum likelihood estimation (to ensure unbiased estimates of the variance components), as implemented in the `mixed` command of Stata (version 14.1):

$$y_{ij} = \beta_0 + \beta_1 I(\mathrm{CI}_i = 1) + \sum_{p=2}^{15} \beta_p I(\mathrm{AREA}_i = p) + u_{1i} I(\mathrm{CI}_i = 1) + u_{2i} I(\mathrm{SI}_i = 1) + \varepsilon_{ij},$$
$$u_{1i} \sim N(0, \tau^2_{\mathrm{CI}}), \qquad u_{2i} \sim N(0, \tau^2_{\mathrm{SI}}), \qquad \varepsilon_{ij} \sim N(0, \sigma^2_{\mathrm{CI}}) \text{ if } \mathrm{CI}_i = 1,$$
$$\varepsilon_{ij} \sim N(0, \sigma^2_{\mathrm{SI}}) \text{ if } \mathrm{SI}_i = 1. \tag{2}$$

In expression (2), i is an index for interviewers and j is an index for respondents nested within interviewers. This model includes a fixed effect associated with the CI technique (β_1), represent-

ing the difference in expected means on the survey measure between the two techniques when adjusting for the fixed area effects, and 14 fixed effects of the various non-reference areas in Germany where the samples were selected, which capture any unexplained variance in measures between interviewers due to the areas where they were assigned. We note that the components of variance for the residual errors that are associated with observations within interviewers (σ^2_{CI} and σ^2_{SI}) and the random interviewer effects (τ^2_{CI} and τ^2_{SI}) are allowed to *vary* depending on the interviewing technique that is used (SI or CI); models of this form have also been discussed and evaluated by Brunton-Smith *et al*. (2017). As mentioned above, we formally test the null hypothesis that $\tau^2_{\mathrm{CI}} = \tau^2_{\mathrm{SI}}$ by using the likelihood ratio test that was outlined and evaluated by West and Elliott (2014), which involves fitting a nested model allowing the interviewer variance components to be equal. We followed a similar procedure to test the null hypothesis that $\sigma^2_{\mathrm{CI}} = \sigma^2_{\mathrm{SI}}$. This model specification also allows us to compute estimates of IICs that are specific to each technique (e.g. $\rho_{\mathrm{int,CI}} = \tau^2_{\mathrm{CI}}/(\tau^2_{\mathrm{CI}} + \sigma^2_{\mathrm{CI}})$).

For binary survey items (e.g. any moving expenses in the past 5 years), we fit the following multilevel logistic regression model:

$$\ln\left\{\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right\} = \beta_0 + \beta_1 I(\mathrm{CI}_i = 1) + \sum_{p=2}^{15} \beta_p I(\mathrm{AREA}_i = p) + u_{1i} I(\mathrm{CI}_i = 1) + u_{2i} I(\mathrm{SI}_i = 1)$$
$$u_{1i} \sim N(0, \tau^2_{\mathrm{CI}}), \qquad u_{2i} \sim N(0, \tau^2_{\mathrm{SI}}). \tag{3}$$

These models were fitted using the adaptive Gaussian–Hermite quadrature procedure that is implemented in the `melogit` command of Stata (version 14.1), which has been shown to work well in studies with smaller sample sizes (Kim *et al*., 2013). IICs were then computed for each technique based on the underlying logistic distribution; for example,

$$\rho_{\mathrm{int,CI}} = \frac{\tau^2_{\mathrm{CI}}}{\tau^2_{\mathrm{CI}} + \pi^2/3}.$$

We followed a similar approach for any ordinal items (e.g. most recent gross monthly income), only using the `meologit` command of Stata to fit multilevel *ordinal* logistic regression models of the same form as in expression (3).

For all fitted models, we carefully examined model diagnostics by using the approaches that are outlined in the literature (Gelman and Hill, 2007; West *et al*., 2014). We examined the distributions of standardized model-based residuals, symmetry in the standardized residuals, potential outliers and robustness of results to those outliers, and distributions of empirical best linear unbiased predictors for the random interviewer effects arising from each technique for a given survey item. For binary survey items that were not found to have significant between-interviewer variance components arising from either technique (i.e. random interviewer effects were not necessary in the models), we performed Hosmer–Lemeshow goodness-of-fit tests and assessed distributions of binned residuals. For any ordinal survey items and binary items presenting evidence of significant between-interviewer variance components, we applied the simulation-based approach for checking model fit that was outlined by Gelman and Hill (2007), chapters 8 and 24. Examples of the Stata and SAS code that were used for the modelling, testing and diagnostic analyses, along with examples of the output produced, are available from `http://wileyonlinelibrary.com/journal/rss-datasets`.

4.3. Mean-squared error comparisons

For privacy and data protection reasons we were not granted access to the entire IEB database

and could work with only selected administrative data for the people who were sampled for this study. As a result, we developed *estimates* of the MSE that is associated with the respondent-based estimates of means and proportions for each technique by using a model-based approach. First, we assumed that the model of interest was defined by the gross sample of $n = 7200$. We estimated the parameters in this hypothetical model by fitting the same types of models as we used for our analyses of the survey data (as described in expressions (2) and (3)) to the administrative data that are available from the IEB database for the gross sample. We then used the `margins` command in Stata to compute model-based estimates of the overall marginal means (or proportions) for the variable of interest for each of the interviewing techniques (see `http://www.stata.com/manuals13/rmargins.pdf` for details). We defined these marginal means as our target parameters of interest for the MSE calculations.

We then fitted the same models to the 'true' administrative values from the IEB database for *respondents only* and estimated the marginal means for each interviewing technique, allowing us to assess the amount of bias that is introduced in the overall estimates of the marginal means for a given technique due to survey non-response. Finally, we fitted the same models as we used in our analyses of the survey responses and computed estimates of the overall marginal means for each interviewing technique. The model-based standard errors of these estimates of the overall marginal means (for the MSE calculations) were computed by using the delta method, as described in `http://www.stata.com/manuals13/rmargins.pdf` (pages 48–52), where the variances of the estimated fixed effects accounted for the variances of the random interviewer effects and the variances of the residuals (when applicable). We computed an estimate of the MSE for a given respondent-based estimate arising from a given technique as the squared difference between the respondent-based estimate and the target parameter (the squared bias), plus the estimated model-based variance of the respondent-based estimate (the variance).

4.4. Criteria for evaluating data quality

For evaluation of data quality, we determined which technique

(a) yielded a significantly lower IIC, which would result in smaller interviewer effects on the variance of descriptive estimates, along with the source(s) of the reduced IIC (i.e. significant differences in between- and/or within-interviewer variance), and
(b) yielded a significantly different mean that was consistent with higher response accuracy.

We organize our results for the entire set of 55 survey items based on

(a) items with differences in IICs due to significant differences in between-interviewer variance components;
(b) items with potentially different IICs due only to significant differences in the within-interviewer variance components and
(c) items with no differences in IICs but significant differences in response distributions between the two interviewing techniques.

5. Results

In total, we found evidence of significant differences in either IICs or response distributions for 19 of the 55 survey items (34.5%), i.e., for 36 (65.5%) of the items, the interviewing technique that was used did not significantly affect the distributions of the survey responses, and this was not due to a lack of statistical power to detect differences in either the variance components that

were used to compute the IICs or means and proportions. We first consider items with different IICs due to significant differences in between-interviewer variance components.

5.1. Items with differences in intra-interviewer correlations due to significantly different between-interviewer variance

We found only five items with different IICs due to significant or marginally significant differences in between-interviewer variance components (Table 1). For all five items, the CI technique was found to have a larger IIC (and larger between-interviewer variance) and, for four of the five items, the difference in between-interviewer variance was found to be only marginally significant ($p < 0.20$). For the one item with available validation data and a significant difference in between-interviewer variance (the longest uninterrupted period of employment in the previous 20 years, $p < 0.01$; see the on-line supporting information for estimates of the variance components in the final model for this variable), a comparison of the estimated MSE values between the two techniques suggested that CI was still producing an estimate with lower MSE (i.e. better quality) *despite the increased IIC* (Table 2). This slight improvement in the quality of the estimate appeared to be arising from less underreporting (on average) in the CI group, where recall error in each group was probably leading to the respondent-based estimates with

Table 1. Items with different IICs due to significant differences in between-interviewer variance components†

Variable (item)	*IICs*			*Estimated variance components*		*Estimated fixed effect of CI*		
	CI	*SI*	*Better?*	*Between-interviewer variance*	*Within-interviewer variance*	$\hat{\beta}_1$	$\hat{\beta}_1/SE(\hat{\beta}_1)$	*Better?*
Number of rooms in housing unit (question 2)	0.087	0.014	SI	CI > SI‡	SI > CI§	0.83§	5.49	CI
Hours worked per week (question 17)	0.073	0.028	SI	CI > SI‡	CI > SI§	3.09§§	1.94	CI
Longest uninterrupted period of gainful employment since April 1st, 1994, in months (question 29)	0.066	<0.01	SI	CI > SI§	—	2.09	0.58	—*
Count of close friends outside the house (question 33)	0.026	0.001	SI	CI > SI‡	SI > CI**	−0.37	−0.75	—*
Interview duration in minutes (not applicable; from time stamp paradata)	0.401	0.364	SI	CI > SI‡	CI > SI§	0.97	0.42	—*

†For all criteria, the 'better' technique is determined on the basis of the technique with the lower IIC, a significantly lower variance component or a significant fixed effect of CI that suggests higher response accuracy for one technique. For variable names, actual questionnaire items are indicated in parentheses (see Section 3.3 for Web links to questionnaires). Total sample sizes for the analyses ranged from 1423 (hours worked per week) to 1850 (interview duration in minutes), and all analyses included 60 interviewers.
‡$p < 0.20$.
§$p < 0.01$.
§§$p < 0.10$.
*Not applicable (no notable difference).
**$p < 0.05$.

Table 2. Comparisons of estimated MSE values for respondent-based estimates of selected descriptive parameters†

Variable	*Estimated MSE: CI group*				*Estimated MSE: SI group*				*CI/SI MSE ratio*
	Benchmark value (full sample)	*Benchmark value (respondents only)*	*Respondent estimate (standard error)*	*Estimated MSE*	*Benchmark value (full sample)*	*Benchmark value (respondents only)*	*Respondent estimate (standard error)*	*Estimated MSE*	
Most recent gross monthly income since January 1st, 2013 (€)									
0	15.3%	14.1%	24.6% (1.3%)	88.18	15.2%	13.5%	28.2% (1.3%)	170.69	0.52
1–1787.60	38.2%	37.5%	28.0% (1.1%)	105.25	38.1%	36.9%	29.0% (1.1%)	84.02	1.25
1787.60–3999.00	38.4%	40.1%	39.5% (1.4%)	3.17	38.5%	40.9%	36.2% (1.3%)	6.98	0.45
⩾4000	8.2%	8.3%	7.9% (0.7%)	0.58	8.2%	8.7%	6.6% (0.6%)	2.92	0.20
Mean longest uninterrupted period of gainful employment since April 1st, 1994 (months)	98.76	101.24	82.22 (3.16)	283.56	97.83	101.12	80.13 (1.72)	316.25	0.90

†The benchmark values of the marginal means and percentages are based on models for the 'true' values, available for the full sample of $n = 7200$ in the administrative (IEB) data, and are treated as the population parameters of interest. Benchmark values for respondents only are similar estimates based on models for the true values, fitted to respondents only, and provide an indication of possible non-response bias. See Section 4.3 for details regarding these MSE estimates (where $\widehat{\text{MSE}} = \{\text{respondent estimate} - \text{benchmark value (full)}\}^2 + \text{SE(respondent estimate)}^2$).

negative bias. We note that the estimated MSE values in Table 2 appeared to be dominated by the bias in respondent reports in each of the two groups, and not non-response bias.

For two of the five items (hours worked per week and interview duration), these differences in IICs arose despite significantly larger within-interviewer variance ($p < 0.01$) introduced by CI, suggesting that the increase in between-interviewer variance relative to SI is outpacing the increase in within-interviewer variance relative to SI. This finding for interview duration was consistent with the literature in this area (Loosveldt and Beullens, 2013), suggesting that conversational interviews may take longer because of the clarification that is provided by the interviewers. We view this finding, suggesting more variability in interview duration within conversational interviewers, as additional confirmation of correct implementation. On average, CI interviews also took 1 min longer overall, but this difference was not significant.

Notably, for two of these five items (the number of rooms in the household and count of close friends outside the house), the within-interviewer variance was significantly larger ($p < 0.05$) when the SI technique was used. This would serve to *reduce* the IICs that are associated with SI, and could also be a source of the different IICs that were noted here. Determining whether this reduction in within-interviewer variance due to CI was reflective of higher data quality overall (despite the resulting increase in the IIC) would require validation data that were not available for these two items. We did find evidence of a significantly higher mean for the number of rooms in the housing unit due to CI despite the increased IIC, which would suggest improved data quality (given the *inclusive* definition of this concept). We also found evidence of a marginally higher mean for hours worked per week, once again suggesting higher data quality (given the *inclusive* definition in this case) despite the increased IIC. Collectively, these results suggest that increases in IICs arising from increased between-interviewer variability in CI are rare, and that the quality of estimates based on CI is not generally hampered by the (rarely) increased IICs.

Considering the diagnostic assessments for these models, we found that the distributions of the standardized residuals for the measures of interview duration and number of rooms in the household had a slight right skew. We therefore refitted the models after performing a natural logarithmic transformation of these measures (adding 1 to each measure before applying the transformation). Residuals arising from these new models appeared to be normally distributed, no outliers were evident and our overall conclusions and hypothesis test results for these two measures did not change. Regarding the measure of longest uninterrupted period of gainful employment, model diagnostics revealed that eight cases reporting a value of 0 were emerging as outliers in terms of the standardized residuals, and the residuals once again had a slight right skew. After dropping these eight cases and considering the same natural logarithmic transformation (after adding 1 to each measure), we arrived at the same conclusions, only with slightly weaker evidence of a significant difference in between-interviewer variance ($p = 0.05$).

Finally, for hours worked per week and count of close friends outside the house, we found similar evidence of a right skew in the standardized residuals and applied similar transformations. Although this once again improved the distributions of the standardized residuals, the differences in between- and within-interviewer variance for these two measures were no longer approaching significance, and the fixed effect of CI on the hours worked per week measure was now significant at the 0.05-level (suggesting a positive effect of CI on reporting accuracy). These diagnostic assessments indicate that whereas some of our findings were robust to possible violations of distributional assumptions, some of the marginal differences in between-interviewer variance (and thus the IICs) may have been driven by a small number of extreme reports. This lends even more support to the conclusion that larger IICs due to increased between-interviewer variance introduced by CI tend to be rare.

Table 3. Items with significant differences in within-interviewer variance components†

Variable (item)	*Within-interviewer variance*	*Estimated fixed effect of CI*		
		$\hat{\beta}_1$	$\hat{\beta}_1/SE(\hat{\beta}_1)$	*Better?*
Number of employees at primary employer (question 16)	CI > SI‡	145.41§	2.02	CI
Total number of employees (all employers) (question 16a)	CI > SI‡	146.67§	2.03	CI
Hours per week in marginal employment (question 23)	SI > CI‡	−1.78‡	−2.70	CI
Number of times registered as unemployed since April 1st, 1994 (question 30)	SI > CI§	0.01	0.09	—§§

†For variable names, actual questionnaire items are indicated in parentheses (see Section 3.3 for Web links to questionnaires). Total sample sizes for the analyses ranged from 760 (hours per week worked in marginal employment) to 1813 (number of times registered as unemployed), and all analyses included 60 interviewers.
‡$p < 0.01$.
§$p < 0.05$.
§§Not applicable (no notable difference).

5.2. Items with significantly different within-interviewer variance only

We found four items producing no evidence of differences in the between-interviewer components of variance, but significant differences in the within-interviewer variance components (Table 3). For each of these four items, the estimated between-interviewer variance components defining the numerators of the IICs (1) were quite close to 0 for both techniques, meaning that the increased within-interviewer variance that is associated with a particular technique was not producing noticeable changes in the IICs.

For three of these four items, the fixed effects of CI were significant and suggested improvements in quality of response. For the two measures of total counts of employees, the CI technique produced significantly ($p < 0.05$) higher means, suggesting reports that are more consistent with the *inclusive* concepts underlying these questions. The CI technique also produced significantly higher within-interviewer variance components for these two measures but, given the negligible IICs overall, the significant and positive fixed effects suggest higher overall quality of the CI estimates due to improved reporting that would not be significantly offset by the IICs. Unfortunately, validation data were not available for these two measures to verify this suggestion.

For the measure of hours per week in marginal employment (where analyses were restricted to only those 760 cases reporting some form of marginal employment), the CI technique produced significantly lower within-interviewer variance but also a significantly *lower* mean, which is not consistent with the results in Table 1 for overall hours worked per week or the *inclusive* concept underlying this question. Conversational interviewers may have found it difficult to communicate the definition of hours worked per week when discussing marginal jobs. Once again, validation data were not available for this measure of hours worked per week, and we note below that this difference in response distributions was not robust to possible violations of assumptions regarding the residual distribution.

Finally, for the number of times registered as unemployed in the previous 20 years, CI produced significantly lower within-interviewer variance but did not affect the response distribution. These results suggest that the interviewing technique did not play a critical role in affecting the

quality of estimates based on this item. Reports of how many times a respondent was registered as unemployed in the previous 20 years could be validated by using the IEB data. We found that the estimated MSE values for the two estimates based on this variable were quite similar between the two techniques, with most of the bias arising from underreporting regardless of the interviewing technique that was used (and not non-response bias). This may reflect socially desirable responding that could not be overcome by CI, given the sensitive nature of this topic.

Considering the diagnostic assessments for these four models, we once again found evidence of right skew in the model-based residual distributions for the two measures of employee counts, and natural logarithmic transformations corrected this problem without substantially changing the results. The differences in within-interviewer variance between the techniques for these two items were now only marginally significant ($p < 0.10$), suggesting that these differences may have been affected by particularly large reports. The positive fixed effects of CI on these response distributions remained robust and significant. For hours per week in marginal employment, a similar transformation once again helped with the residual distribution, and the difference in within-interviewer variance remained significant, but the fixed effect of CI was reduced to marginal significance ($p = 0.12$). Finally, for times registered as unemployed, both natural logarithmic and square-root transformations of the outcome were not found to improve the residual distribution, and we instead considered multilevel ordinal logistic regression models for a recoded version of this measure with five categories (0, 1, 2, 3, and 4 or more). We once again did not find any evidence of effects of CI on this response distribution, and differences in within-interviewer variance may have once again been driven by a small number of extreme reports. In sum, differences in within-interviewer variance between the techniques for those items with negligible between-interviewer variance components did not seem to be as robust to extreme reports by individual respondents.

5.3. Items with differences in response distributions and no differences in intra-interviewer correlations

Table 4 presents estimated fixed effects of CI for the 10 items that were found to have significant fixed effects but no differences in the within- or between-interviewer variance components (and hence similar IICs). We first consider most recent gross monthly income, given that administrative data were available for this item. The positive and significant fixed effect in the multilevel ordinal logistic regression model suggests that CI tended to increase the odds of reporting higher income categories, but assessment of whether this is consistent with higher data quality (per the *inclusive* definition underlying this concept) requires assessment of the administrative data. Two results related to the validation of monthly income reports in Table 2 are critical. First, the significant fixed effect of CI on the distribution of these reports is not arising from differential non-response error in the CI group, as the model-based marginal percentages across the four income categories based on true values for the respondents were quite similar to the same percentages for the full sample in both groups. Second, when considering the model-based marginal percentages based on respondent *reports*, we see evidence of CI resulting in reports that were closer to the true values. Both groups tended to underreport monthly income, but significantly less so in the CI group. Indeed, the estimated MSE values for three of the four gross monthly income percentages in Table 2 estimated by using CI are either half or less than half of the estimated MSE values for SI. We found evidence of similar underreporting for gross annual income in 2013, but no differences between the techniques were found.

Considering the other nine items in Table 4 without validation data in the IEB database, we see consistent evidence of CI shifting response distributions in the direction of higher accuracy. For

Table 4. Items with no differences in IICs but significant fixed effects of CI†

Variable (item)	*Estimated fixed effect of CI*		
	$\hat{\beta}_1$	$\hat{\beta}_1/SE(\hat{\beta}_1)$	*Better?*
Moving expenses in past 5 years (question 5)	−0.45‡	−4.67	CI
Home improvement expenses in past 2 years (question 7)	−0.24§	−2.29	CI
Any short-term contracts since January 1st, 2013? (question 10A)	−1.06‡	−3.76	CI
Most recent gross monthly income (question 19 and question 20)	0.19§	2.24	CI
Annual income exact? (question 21A)	0.78‡	5.23	CI
Indicator of belonging to a church (question 35C)	−0.40§	−2.42	CI
Indicator of using a social networking site (question 36)	0.19§§	1.90	CI
Consent to audio recording (INTRO4)	−0.80‡	−3.23	SI
Quality of respondent understanding/comprehension (interviewer observation; POST3)	−0.88‡	−4.37	CI
Quality of information provided by respondents (interviewer observation; POST4)	−0.67‡	−2.96	CI

†For variable names, actual questionnaire items are indicated in parentheses (see Section 3.3 for Web links to questionnaires). Total sample sizes for the analyses ranged from 1370 ('Annual income exact?') to 1850 ('Most recent gross monthly income'), and all analyses included 60 interviewers.
‡$p < 0.01$.
§$p < 0.05$.
§§$p < 0.10$.

both moving expenses in the previous 5 years and home improvement expenses in the previous 2 years, we find significant *negative* fixed effects of CI, which suggest reporting that is more consistent with the *exclusive* concepts underlying these questions (e.g. buying food for friends who helped you to move does not count, and simple household maintenance, such as fixing light bulbs, does not count). CI also resulted in a significantly lower probability of reporting any short-term (or marginal job) contracts since January 1st, 2013, which suggests higher response accuracy (given an *exclusive* definition of what exactly is considered short-term employment), and significantly increased the probability that a respondent states that they provided an exact measure of annual income (see the on-line supporting information for the estimated parameters in this model). For indicators of being a church member and using a social networking Web site (e.g. Facebook), CI was found to decrease the probability of reporting that you belong to a church (suggesting less socially desirable reporting) and to increase the probability of reporting that you use a social networking site (perhaps because conversational interviewers were able to provide more examples and to clarify the meaning of a 'social networking site').

Considering next the measures of paradata in Table 4, we found evidence of high estimated IICs in both groups for the indicator of consent to audio recording (0.201 for CI and 0.123 for SI), but we also found puzzling evidence of CI significantly *decreasing* the probability of consenting to audio recording overall. We were unable to identify reasons for this finding, given that consent was requested before the interview began, and no audio recordings of this question were available. This is certainly an area that is worthy of future research. For example, we could speculate that there might be a 'carry-over' effect among conversational interviewers, where the more conversational approach ends up being used when asking for consent to audio recording. This lack of formality may ultimately lead to respondents refusing to have the interview recorded at a higher rate. For the ordinal post-survey interviewer observations of data quality (in terms of respondent understanding or comprehension and overall data quality), where higher values

indicate poorer data quality, respondents assigned to CI interviewers were perceived to have significantly lower odds of providing responses of poor quality. Both groups were once again found to have substantial IICs on these post-survey observations as well, which is consistent with the existing literature in this area (O'Muircheartaigh and Campanelli, 1998; Kaminska *et al*., 2010; Eckman *et al*., 2013; West and Kreuter, 2013; West and Peytcheva, 2014). These findings suggest that, despite the high IICs in the distributions of these post-survey observations, interviewers who were trained in CI tended to perceive that their respondents were providing data of higher quality as a whole. In sum, Table 4 presents general evidence of CI shifting response distributions in the direction of higher data quality without increasing IICs substantially relatively to SI.

Considering next diagnostics for the models that were fitted to these 10 variables, five were binary items that did not present evidence of significant between-interviewer variance for either technique (any moving expenses, any home improvement expenses, any short-term contracts, going to church and using a social network). Hosmer–Lemeshow goodness-of-fit tests for these five items suggested that the fits of the final logistic regression models (excluding the random interviewer effects) were adequate in each case, with the range of the Hosmer–Lemeshow p-values being 0.05–0.38. We applied the simulation-based model checking approach that was outlined above to the three ordinal variables in Table 4 (two of which were interviewer observations about reporting quality that presented evidence of substantial between-interviewer variance overall) and the two binary indicators of reporting exact income and consenting to audio recording, each of which presented evidence of significant overall between-interviewer variance. The simulation-based model checking approaches revealed that the 2.5- and 97.5-percentiles for the 1000 simulated probabilities of interest (generated from the final model that was estimated for each variable, e.g. the probability of consenting to audio recording) covered the observed probabilities from the respondent sample in nearly all cases, once again suggesting adequate model fit. The SAS code that was used to implement these simulation-based approaches for assessing model fit is available from `http://wileyonlinelibrary.com/journal/rss-datasets`.

Similar diagnostic assessments for the 36 items that did not produce evidence of significant differences in response distributions or IICs due to the interviewing techniques that were used did not reveal any substantial differences in the conclusions that one would draw. In general, we found consistent evidence (for non-attitudinal survey items) of CI shifting response distributions slightly in the direction of higher accuracy of response based on the underlying concepts for these items (the results are not shown). We also found that CI tended to produce higher estimates of IICs for the majority of these 36 items, but differences in the between-interviewer variance components (and thus the IICs) tended to be slight as well. For example, for total length of time working in the previous 20 years, the estimated IIC for CI was 0.006, and the estimated IIC for SI was 0.004.

6. Discussion

6.1. Summary of results and main conclusions

In this study, we found that CI does not generally increase IICs. CI was found to produce marginally or significantly higher between-interviewer variance (and thus higher IICs) for only five of the 55 survey items that were analysed (see Table 1), and for the one item among these with official administrative data available for the full sample (the longest uninterrupted period of employment in the previous 20 years), the increase in the IIC was not sufficiently substantial to produce an estimated mean (based on respondent reports) with higher estimated MSE.

Importantly, this small number of differences was not due to a lack of statistical power to detect meaningful differences between variance components. CI was also found to produce shifts in response distributions for 14 of the 55 items that were significant and suggestive of higher quality of data (see Tables 1, 3 and 4), and these findings were confirmed for selected items (e.g. most recent gross monthly income) by analyses of administrative data that were available on the sampling frame. Finally, no differences in IICs or response distributions were found for 65.5% of the items that were analysed.

Overall, these findings suggest that

(a) significant increases in IICs due to the use of CI are rare,
(b) CI does have the ability to improve reporting relative to SI for complex survey items (as shown in the literature),
(c) any increases in IICs due to CI do not offset reductions in response bias and
(d) for many items, there are no significant differences in quality of response between SI and CI.

These findings provide an important contribution to the existing literature in this area and indicate that the use of CI in practice does not typically increase IICs to the point of harming survey estimates.

6.2. Suggestions for practice

CI was found to produce marginally or significantly higher IICs for five items: reports of the number of rooms in the housing unit, hours worked per week, the longest uninterrupted period of employment in the previous 20 years, counts of close friends outside your house and interview duration. CI was also found to shift the means of response distributions in a manner that suggested significantly higher accuracy of response for 14 items, including variables measuring housing unit characteristics (the number of rooms, recent maintenance or moving expenses), recent short-term employment, hours worked per week (in marginal or regular employment), income (most recent monthly income or reporting an exact income amount), counts of employees at places of work (primary employer or all employers), belonging to a church, use of a social networking site and two interviewer observations of quality of interview. Importantly, for the one item among these with available administrative data (most recent gross monthly income), this significant difference in response distributions ultimately produced estimated proportions with reduced bias and improved MSE relative to SI. Furthermore, interviewers using CI perceived the quality of the data that were reported to be substantially higher compared with interviewers using SI.

So what can survey researchers draw from these specific findings? Questions about specific household expenses or characteristics, where response difficulty and ambiguity might arise depending on the context that is associated with the question (e.g. what would a respondent count as a 'room' in their housing unit?) seem to benefit from CI, and this is consistent with the literature. CI also seemed to result in better reports on items that were related to recent employment, income, working hours and employer characteristics, each of which could also be cognitively challenging. However, SI was found to produce lower IICs for some items, as noted above. This did not ultimately seem to affect the overall quality of the estimates; in general, we found that shifts in response distributions tended to have a larger effect on the estimated MSE of estimates than did differences in IICs between the two techniques.

In terms of interview duration, the CI technique resulted in interviews that were roughly 1 min longer, on average, but this difference was not significant. This may have been due to the fact that many questions did not end up requiring extensive clarification from the interviewers

(Mittereder *et al*., 2017). The CI technique did result in higher within-interviewer variance in interview length, which was expected, and marginally higher between-interviewer variance in duration of interview (which also was expected). We view these results as evidence of successful implementation, but we also note that increased duration for some interviewers using this technique (and certain respondents who were interviewed with this technique) has the potential to increase costs, as noted in the literature. Survey managers therefore need to monitor carefully the average interview durations of conversational interviewers, and possibly intervene if selected interviewers are consistently producing higher-than-expected interview times.

Regardless of the important differences that we found favouring CI, strategies for reducing the potential for interviewer effects when using CI are still needed. Careful monitoring of audio recordings of the interviews by interviewer managers may ultimately improve 'on-the-fly' training of interviewers using CI; for example, this monitoring may reveal that a particular interviewer tends to go off on too many tangents. Finally, CI also introduced more variability in response times, which has potential cost implications. Analyses of audio recordings would also help to ensure that this increased variability is not arising from specific conversational interviewers going off on unrelated tangents during the interview.

Overall, we see the results of this study as supporting an approach where interviewers use scripted, standardized language as much as possible but are not restricted from providing additional clarification (possibly on top of providing a simple definition) if needed to ensure respondent comprehension for more challenging questions that may introduce ambiguities. Survey organizations in some European countries already train interviewers in this manner for selected projects (e.g. Germany and Finland). Committing exclusively to CI for all items would probably increase interview duration and costs (in part due to the increased amount of time that is required for training on concepts and definitions); committing exclusively to SI may save costs but could also lead to reduced accuracy of response for more challenging items.

6.3. Directions for future research

Future research needs to focus on the *sources* of the IICs that are introduced by these techniques (and especially CI) in the reports on particular survey measures. For example, for reports regarding the longest uninterrupted period of employment in the past 20 years, at what point in the survey process was the most variability introduced between conversational interviewers? Given that the sample was interpenetrated by design (controlling for fixed effects of the areas where the interviewers were working), did the variability in response rates noted between interviewers using CI lead to the recruitment of individuals with different types of employment histories? Or, given that variability in the response rates was also noted for the SI interviewers, did all of the variance between the interviewers in the reports come from measurement error variance? Future work needs to apply the methods that were discussed in West and Olson (2010) and West *et al*. (2013) to decompose the total between-interviewer variance that is found for a given survey item and a given technique (which could increase the IIC) into measurement error variance and non-response error variance between interviewers, assess the correlations of these two interviewer-specific error sources and examine whether they tend to offset each other (e.g. Kreuter *et al*. (2010)). Doing so will require the types of administrative data that were considered in this study.

Future research also needs to focus on understanding what exactly transpires during conversational or standardized interviews, possibly by using conversation analytic techniques (e.g. Maynard *et al*. (2010)). Although we coded the behaviours that were evident in a small sample of these interviews mainly as a means of technique verification (Mittereder *et al*., 2017), this

did uncover some unexpected findings: for example, the conversational interviewers tended to use neutral probes at a higher rate than expected when faced with a question or confusion. Was this because specific clarification was not needed? How much of this decision to use a neutral probe instead of providing additional clarification is based on the judgement of the interviewer or the current context of the conversation? Applications of standardized conversation analytic methods to random subsamples of interviews using each technique would be helpful for understanding the mechanics of each technique that may lead to higher or lower quality of data. Furthermore, interviewers using CI may benefit from additional training and practice; would higher amounts of training in CI ultimately lead to even better responses? This is another potential direction for future research.

Finally, we urge survey methodologists to consider replications of this study. Although this may be difficult, given the need for

(a) training of a large number of interviewers,
(b) interpenetrated samples and
(c) validation data for selected variables,

additional replications would enable the field to accumulate a body of evidence that fully examines the bias and variance properties that are associated with these two interviewing techniques. Clearly the results of this study are based on a specific German population, the interviewers employed by a specific data collection organization and a survey including a variety of content but primarily focusing on employment histories. Differences in cultural norms regarding interviewer–respondent interactions between Germany and other countries may lead to different effects of CI and SI, meaning that replications of this study in different cultures will be important. Additional replications of this study could be used by survey managers to make informed decisions about the benefits or costs of using these techniques depending on the context of their survey.

Acknowledgements

Financial support for this study was provided by grants from the US National Science Foundation (SES-1324689 and SES-1323636), the Institute for Employment Research in Nuremberg, Germany, and infas in Bonn, Germany. We sincerely thank the staff of infas (especially Birgit Jesske and Anne Kersting) for their dedicated assistance with interviewer training and data collection operations, and the staff of the Institute for Employment Research (especially Malte Schierholz, Daniela Hochfellner, Ulrich Thomsen and Arne Bethmann) for their assistance with data management and access. We also acknowledge the valuable comments and critiques provided by Mark Trappmann.

References

Ackermann-Piek, D. and Massing, N. (2014) Interviewer behavior and interviewer characteristics in PIAAC Germany. *Meth. Data Anal.*, **8**, 199–222.

Bell, K., Fahmy, E. and Gordon, D. (2016) Quantitative conversations: the importance of developing rapport in standardised interviewing. *Qual. Quant.*, **50**, 193–212.

Belli, R. F. and Lepkowski, J. M. (1996) Behavior of survey actors and the accuracy of response. *5th Health Survey Research Methods Conf., Breckenridge, June 1995*.

Brick, J. M., McGuinness, R., Lapham, S. J., Cahalan, M., Owens, D. and Gray, L. (1995) Interviewer variance in two telephone surveys. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 447–452.

Bruckmeier, K., Müller, G. and Riphahn, R. T. (2015) Survey misreporting of welfare receipt—respondent, interviewer, and interview characteristics. *Econ. Lett.*, **129**, 103–107.

Brunton-Smith, I., Sturgis, P. and Leckie, G. (2017) Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *J. R. Statist. Soc.* A, **180**, 551–568.
Cannell, C. F. and Axelrod, M. (1956) The respondent reports on the interview. *Am. J. Sociol.*, **62**, 177–181.
Cannell, C. F., Lawson, S. A. and Hausser, D. L. (1975) A technique for evaluating interviewer performance: a manual for coding and analyzing interviewer behavior from tape recordings of household interviews. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.
Clark, H. H. and Brennan, S. A. (1991) Grounding in communication. In *Perspectives on Socially Shared Cognition*. (eds L. B. Resnick, J. M. Levine and S. D. Teasley). Washington: American Psychological Association Books.
Conrad, F. G. and Schober, M. F. (2000) Clarifying question meaning in a household telephone survey. *Publ. Opin. Q.*, **64**, 1–28.
Conrad, F. G., Schober, M. F., Jans, M., Orlowski, R. A., Nielsen, D. and Levenstein, R. (2015) Comprehension and engagement in survey interviews with virtual agents. *Front. Psychol.*, **6**, article 1578.
Dijkstra, W. (1987) Interviewing style and respondent behavior: an experimental study of the survey-interview. *Sociol. Meth. Res.*, **16**, 309–334.
Eckman, S., Sinibaldi, J. and Möntmann-Hertz, A. (2013) Can interviewers effectively rate the likelihood of cases to cooperate? *Publ. Opin. Q.*, **77**, 561–573.
Fowler, F. J. and Mangione, T. W. (1990) *Standardized Survey Interviewing: Minimizing Interviewer-related Error*. New York: Sage.
Gelman, A. and Hill, J. (2007) *Data Analysis using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
Groves, R. M. (2004) The interviewer as a source of survey measurement error. In *Survey Errors and Survey Costs*, 2nd edn, ch. 8. New York: Wiley-Interscience.
Groves, R. M. and Magilavy, L. J. (1986) Measuring and explaining interviewer effects in centralized telephone surveys. *Publ. Opin. Q.*, **50**, 251–266.
Haan, M., Ongena, Y. and Huiskes, M. (2013) Interviewers' questions: rewording not always a bad thing. In *Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention* (eds P. Winker, N. Menold and R. Porst), pp. 173–193. Frankfurt: Peter Lang Academic Research.
Henson, R., Cannell, C. F. and Lawson, S. (1976) Effects of interviewer style on quality of reporting in a survey interview. *J. Psychol.*, **93**, 221–227.
Houtkoop-Steenstra, H. (1995) Meeting both ends: between standardization and recipient design in telephone survey interviews. In *Situated Order: Studies in the Social Organization of Talk and Embodied Activities* (eds P. ten Have and G. Psathas), pp. 91–107. Washington DC: University Press of America.
Hubbard, F., Antoun, C. and Conrad, F. G. (2012) Conversational interviewing, the comprehension of opinion questions and nonverbal sensitivity. *A. Conf. American Association for Public Opinion Research, Orlando*.
Jacobebbinghaus, P. and Seth, S. (2007) The German integrated employment biographies sample (IEBS). *Zeits. Wirts. Sozlwissensch.*, **127**, 335–342.
Kahn, R. L. and Cannell, C. F. (1957) *The Dynamics of Interviewing: Theory, Technique, and Cases*. New York: Wiley.
Kaminska, O., McCutcheon, A. L. and Billiet, J. (2010) Satisficing among reluctant respondents in a cross-national context. *Publ. Opin. Q.*, **74**, 956–984.
Kim, Y., Choi, Y.-K. and Emery, S. (2013) Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages. *Am. Statistn*, **67**, 171–182.
Kreuter, F., Müller, G. and Trappmann, M. (2010) Nonresponse and measurement error in employment research: making use of administrative data. *Publ. Opin. Q.*, **74**, 880–906.
Loosveldt, G. and Beullens, K. (2013) The impact of respondents and interviewers on interview speed in face-to-face interviews. *Socl Sci. Res.*, **42**, 1422–1430.
Mangione, T. W., Fowler, F. J. and Louis, T. A. (1992) Question characteristics and interviewer effects. *J. Off. Statist.*, **8**, 293–307.
Maynard, D. W., Freese, J. and Schaeffer, N. C. (2010) Calling for participation: requests, blocking moves, and rational (inter)action in survey introductions. *Am. Sociol. Rev.*, **75**, 791–814.
Mittereder, F., Durow, J., West, B. T., Kreuter, F. and Conrad, F. G. (2017) Interviewer-respondent interactions in conversational and standardized interviewing. *Fld Meth.*, to be published.
O'Muircheartaigh, C. and Campanelli, P. (1998) The relative impact of interviewer effects and sample design effects on survey precision. *J. R. Statist. Soc.* A, **161**, 63–77.
Peneff, J. (1988) The observers observed: French survey researchers at work. *Socl Prob.*, **35**, 520–535.
Sayles, H., Belli, R. F. and Serrano, E. (2010) Interviewer variance between event history calendar and conventional questionnaire interviews. *Publ. Opin. Q.*, **74**, 140–153.
Schaeffer, N. C., Dykema, J. and Maynard, D. W. (2010) Interviewers and interviewing. In *Handbook of Survey Research*, 2nd edn (eds J. D. Wright and P. V. Marsden). Bingley: Emerald.
Schnell, R. and Kreuter, F. (2005) Separating interviewer and sampling-point effects. *J. Off. Statist.*, **21**, 389–410.
Schober, M. F. and Conrad, F. G. (1997) Does conversational interviewing reduce survey measurement error? *Publ. Opin. Q.*, **61**, 576–602.

Schober, M. F., Conrad, F. G., Dijkstra, W. and Ongena, Y. P. (2012) Disfluencies and gaze aversion in unreliable responses to survey questions. *J. Off. Statist.*, **28**, 555–582.
Schober, M. F., Conrad, F. G. and Fricker, S. S. (2004) Misunderstanding standardized language in research interviews. *Appl. Cogn. Psychol.*, **18**, 169–188.
Suchman, L. and Jordan, B. (1990) Interactional troubles in face-to-face survey interviews. *J. Am. Statist. Ass.*, **85**, 232–241.
Trappmann, M., Beste, J., Bethmann, A. and Müller, G. (2013) The PASS panel survey after six waves. *J. Lab. Markt Res.*, **46**, 275–281.
Trappmann, M., Krumpal, I., Kirchner, A. and Jann, B. (2014) Item sum: a new technique for asking quantitative sensitive questions. *J. Surv. Statist. Methodol.*, **2**, 58–77.
West, B. T. and Elliott, M. R. (2014) Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers. *Surv. Methodol.*, **40**, 163–188.
West, B. T. and Kreuter, F. (2013) Factors affecting the accuracy of interviewer observations: evidence from the National Survey of Family Growth (NSFG). *Publ. Opin. Q.*, **77**, 522–548.
West, B. T., Kreuter, F. and Jaenichen, U. (2013) Interviewer effects in face-to-face surveys: a function of sampling, measurement error or nonresponse? *J. Off. Statist.*, **29**, 277–297.
West, B. T. and Olson, K. (2010) How much of interviewer variance is really nonresponse error variance? *Publ. Opin. Q.*, **74**, 1004–1026.
West, B. T. and Peytcheva, E. (2014) Can interviewer behaviors during ACASI affect data quality? *Surv. Pract.*, **7**, no. 5.
West, B. T., Welch, K. B. and Galecki, A. T. (2014) *Linear Mixed Models: a Practical Guide using Statistical Software*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
van der Zouwen, J., Dijkstra, W. and Smit, J. H. (1991) Studying respondent-interviewer interaction: the relationship between interviewing style, interviewer behavior, and response behavior. In *Measurement Errors in Surveys* (eds P. P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz and S. Sudman), pp. 419–438. New York: Wiley.