DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 349–373, http://dx.doi.org/10.1515/JOS-2016-0020

Interviewer Effects on a Network-Size Filter Question

Michael Josten[1] *and Mark Trappmann*[2]

There is evidence that survey interviewers may be tempted to manipulate answers to filter questions in a way that minimizes the number of follow-up questions. This becomes relevant when ego-centered network data are collected. The reported network size has a huge impact on interview duration if multiple questions on each alter are triggered. We analyze interviewer effects on a network-size question in the mixed-mode survey "Panel Study 'Labour Market and Social Security'" (PASS), where interviewers could skip up to 15 follow-up questions by generating small networks. Applying multilevel models, we find almost no interviewer effects in CATI mode, where interviewers are paid by the hour and frequently supervised. In CAPI, however, where interviewers are paid by case and no close supervision is possible, we find strong interviewer effects on network size. As the area-specific network size is known from telephone mode, where allocation to interviewers is random, interviewer and area effects can be separated. Furthermore, a difference-in-difference analysis reveals the negative effect of introducing the follow-up questions in Wave 3 on CAPI network size. Attempting to explain interviewer effects we neither find significant main effects of experience within a wave, nor significantly different slopes between interviewers.

Key words: Partial falsification; network generator; filter questions; interviewer cheating.

1. Introduction and Research Question

Within the total survey error framework (Groves et al. 2004; Biemer 2010) survey interviewers play a central role as the agents who implement the survey design. Their tasks may comprise selecting the right target person, achieving contact with the target person and eliciting cooperation, explaining the task to the survey respondent, asking the questions, probing and coding answers. Consequently, interviewers can influence almost every error source in a survey.

This article focuses on interviewer effects on measurement. Interviewers may influence survey measurement in a variety of ways: there is ample evidence that the mere presence of an interviewer causes respondents to give more socially desirable answers than in self-administered surveys (Tourangeau and Yan 2007). Furthermore, respondents may be influenced by observable interviewer characteristics like his or her age (Freeman and Butler 1976), gender (Groves and Fultz 1985; Huddy et al. 1997) or ethnic affiliation (Schuman and Converse 1971) – where direction and strength of the effect have often

[1] Opinion Market Research & Consulting GmbH, Rollnerstr. 8, 90408 Nuremberg, Germany. Email: michael.josten@opinion.de
[2] Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany and University of Bamberg, Germany. Email: mark.trappmann@iab.de
Acknowledgments: We would like to thank Stephanie Eckman, three anonymous reviewers, and the associate editor, for their valuable comments on and suggestions for previous versions of this article.

been shown to depend on the interaction between respondent and interviewer characteristics. Another source of interviewer variance is nonstandard behavior during the interview like probing (Freeman and Butler 1976; Mangione et al. 1992; van der Zouwen et al. 2004). All explanations mentioned so far refer to unintended interviewer effects on measurement. A different explanation focuses on deliberate interviewer misbehavior in the sense of partial interview falsifications, emanating from an incentive to cheat.

Following the seminal article by Crespi (1945), there have been regular publications on interviewer cheating (cf. Blasius and Friedrich 2012 for a brief overview), focusing on interviewers' motivation for cheating (Crespi 1945), on methods to detect (Biemer and Stokes 1989) and prevent (AAPOR 2003) cheating and on consequences for estimates from surveys (Schnell 1991; Schraepler and Wagner 2005). Cheating is usually considered to be a problem in CAPI rather than CATI surveys (Guterbrock 2008). While interviewers who fabricate complete interviews run a high risk of detection, other, more subtle techniques are harder to detect (Schnell 2012). These include the selection of the wrong target persons, the fabrication of parts of the interview, but also "[. . .] deliberately miscoding the answer to a question in order to avoid follow-up questions" (Guterbrock 2008, 267). Depending on the payment scheme and the tightness of supervision, interviewers might be tempted to increase their efficiency by editing answers to filter questions in a way that reduces interview duration.

With respect to panel surveys, interviewers face a high risk of detection when fabricating complete interviews. Inconsistencies in answers across waves are easily detected. In addition, in most panel surveys, respondents are routinely contacted by the survey agency in between waves for tracking purposes and a complete fabrication would immediately become apparent in case of interviewer changes. In the German Socio-Economic Panel Survey (GSOEP), such falsifications were thus almost exclusively found in initial wave interviews of new refreshment samples (cf. Schraepler and Wagner 2005). This makes resorting to partial falsifications, like taking shortcuts at filter questions, even more attractive in panel surveys.

Underreporting in filter and screener questions has received increased attention in recent years (Kreuter et al. 2011; Tourangeau et al. 2012). Several projects have been launched in order to investigate the processes leading to such underreporting (Tourangeau et al. 2015; Eckman et al. 2014). Most of the results published so far refer to respondent effects. The investigation of interviewer effects on measurement in filter questions has been identified as an important topic for future research in this literature (Kreuter et al. 2011) as only few studies have been devoted to this.

We identified three studies dealing with interviewer effects on filter questions in general. Schnell and Kreuter (2000) argue that differences in victimization rates between different German victimization surveys are in part due to interviewers taking shortcuts at filter questions (Schnell and Kreuter 2000, 114f.). In a later study, Matschinger et al. (2005) investigated a mental health screening question that triggered a series of follow-up questions if endorsed. They found huge interviewer effects, and a latent class regression clearly revealed classes of dishonest interviewers who exhibited learning effects across fieldwork. Finally, Kosyakova et al. (2015) found interviewers affected the endorsement rate of filter questions in the German PASS panel. Controlling for a large set of interviewer and respondent attributes, about three percent of the variance in responses to filter

questions can be shown to be interviewer variance. For CAPI interviewers the endorsement rate and thus the number of follow-up questions decreased with growing interviewer experience.

One specific area where interviewer effects on filter questions can be assumed to be of particular importance is the collection of ego-centered social network data in surveys. The number of persons reported in response to network generator questions (i.e., questions that are designed to elicit the name of alters in the network of a respondent) can have a huge impact on subsequent interview duration if – as is often the case – multiple follow-up questions on each alter or even questions on the relationship between each pair of alters are asked. Given that this question is also less objective than other common filter questions, for example on current employment status, it may stand out to interviewers as the one question where a lot of time can be saved and detection probability is low. Consequently there is some evidence of strong interviewer effects on the number of persons named in response to network generator questions (van Tilburg 1998; Marsden 2003; Brüderl et al. 2013).

Interviewer effects on filter questions are usually reported for PAPI and CAPI surveys where interviewers are not under constant supervision by their survey organization. We do not know of any CATI study that finds strong effects. (Kreuter et al. (2011) even found no effect of experimentally varied payment schemes (payment by case vs. payment by hour) on answers to filter questions in a CATI survey. They argue that the routine monitoring is what keeps CATI interviewers from cheating). A serious drawback of previous research is that interviewers are usually strongly confounded with primary sampling units (PSUs) and thus interviewer and area effects are difficult to separate. In this article CATI and CAPI interviews within identical PSUs are utilized to disentangle this effect.

Using a mixed-mode survey, we will evaluate four research questions in this article: can we find any interviewer effects on a network-size filter question? Does the size of these effects differ by the mode-specific combination of payment scheme and supervision practice? Can these effects be explained by area differences in network size? And do interviewers learn to cheat as they gain experience with the survey?

2. Previous Research

Social network size measures have been discussed as being prone to interviewer effects for some time. This is usually investigated by analyzing to what extent observed network sizes differ across interviewers and what proportion of these variances can be explained by interviewer characteristics or by differences in respondent composition (van Tilburg 1998; Marsden 2003; Brüderl et al. 2013; Paik and Sanchagrin 2013). Typically the intraclass correlation (ICC) is used as a measure for the size of the interviewer effect. The ICC can be interpreted as the (covariate-adjusted) proportion of variance that is due to differences between interviewers. For a formal definition in the context of random-effects models, see the methods section.

Controlling for a large number of sociodemographic respondent and interviewer characteristics, both van Tilburg (1998) and Marsden (2003) found a strong interviewer effect on the number of social contacts elicited, measured by an ICC of $\rho_{int} = 0.15$ and $\rho_{int} = 0.13$, respectively, after controlling for respondent-level predictors of network size. Compared to the average size of interviewer effects for any kind of questions in personal

interview surveys reported by Groves (1989, 319), these effects are quite large. Nevertheless, recent studies discovered even larger interviewer effects far exceeding this for network size: Paik and Sanchagrin (2013) found an intraclass correlation of up to $\rho_{int} = 0.22$, while Brüderl et al. (2013) found by far the largest effect of $\rho_{int} = 0.40$ after controlling for respondent-level predictors. However, Marsden (2003) and Brüderl et al. (2013) could not identify interviewer characteristics significantly affecting network sizes. Van Tilburg (1998) found experienced interviewers produced reduced network sizes, whereas interviewers who were better educated and had more experience in the particular survey generated larger networks.

One common approach is to classify interviewers according to the (pattern of) network sizes they generate. For example, Brüderl et al. (2013) identified three types of interviewers by analyzing which interviewers affect the intraclass correlation most, thus distinguishing between diligent, normal, and fraudulent interviewers. This approach does not take into account the temporal pattern of responses. An alternative approach applied by Matschinger et al. (2005) for a mental health screening question makes use of this temporal order of the interviews. They used latent class regression techniques to create classes of interviewers with similar patterns of responses to filter questions across their sequence of interviews. Thereby, they were able to identify classes of interviewers who "learn" to cheat while becoming more experienced with the administration of the questionnaire.

The network delineation instrument that previous studies used in their analysis was an interviewer-administered name generator where respondents were instructed to report the names of persons they are regularly in touch with in specific situations. The delineating procedure thus constituted a complex task for both respondents and interviewers. Respondents had to interpret these questions correctly and select appropriate persons from their network, while interviewers had to check answers and apply appropriate probing strategies in case of difficulties of comprehension or lacking plausibility. Due to the questions' complexity, the explanation for the large interviewer variances provided by van Tilburg (1998) and Marsden (2003) resides in different probing strategies among the interviewers. Thus they advise training interviewers more carefully. In contrast, Brüderl et al. (2013) and Paik and Sanchagrin (2013) explain the differences in network size by deliberate interviewer misbehavior in order to shorten an interview, an explanation that has been largely neglected by previous literature. All of these studies investigate interviewer effects on network size in a face-to-face setting where there was a nonrandom allocation of interviewers to respondents. In these designs, there is not sufficient interpenetration to separate interviewer effects from area effects (Groves 1989, 270f.). Thus regional variance and interviewer variance are confounded (O'Muircheartaigh and Campanelli 1998; Schnell and Kreuter 2005). Consequently, the unexplained interviewer variance that these studies interpret as evidence for interviewer effects might instead be an unexplained area effect and reflect local differences in network size instead. To tackle this, van Tilburg (1998) and Brüderl et al. (2013) included area dummies in their models. However, this strategy leaves only those areas which were worked by more than one interviewer for the identification of interviewer effects. Marsden (2003) argued against area effects by showing that the intra-interviewer correlation for a less complex "global" network size measure, almost identical to the one that is analyzed in this article, amounts

to only 0.04–0.05. Paik and Sanchagrin (2013) compared outlier interviewers to the remaining interviewers in the same PSU to prove that there is no significant difference in network size estimates between geographical regions that can account for the existence of outlier interviewers.

In contrast to previous approaches, we are able to exploit the mixed-mode design implemented in PASS as well as the longitudinal character of the data. The mixed-mode design brings with it a direct comparability of a CATI sample, in which interviewer assignment is independent of region, and a CAPI sample, in which interviewer and region are confounded. Different incentive structures and supervision practices between modes can thus be exploited to differentiate between explanations referring to the complexity of the filter question and explanations referring to interviewer cheating.

3. Data and Hypotheses

This article uses data from the third wave of the German panel survey "Labour Market and Social Security" (PASS) (Trappmann et al. 2010; Trappman et al. 2013). PASS is an annual panel survey that focuses on labor market, poverty, and social policy research. While the target population is all households in Germany, low-income households are oversampled. In each household the head of the household answers a household questionnaire. Subsequently, every person aged 15 or older is interviewed with a person questionnaire. PASS is implemented with a sequential mixed-mode design within 300 primary sampling units. Most of the households are interviewed in CATI mode, while households and persons that cannot be contacted by phone or that prefer a personal interview are interviewed in CAPI. In Wave 3, which is analyzed here, 5,663 persons (42.1%) have been interviewed in CAPI and 7,776 (57.9%) in CATI mode (Bethmann and Gebhardt 2011).

The PASS survey offers the rare opportunity of a mixed-mode survey where answers to the same filter question are available from CATI as well as CAPI interviews in the same primary sampling units. This entails a direct comparability of the interviewer influences in the context of different supervision practice and incentive structure as well as the possibility to consider interviewer and area effects separately.

The network-size filter question that is analyzed in this article was asked in the person questionnaire in Wave 3. Respondents were first asked whether they have any close friends or family members outside their household with whom they have a strong relationship. If this first filter question was endorsed, it was followed by the inquiry concerning the number of such contacts outside the household. For each of the three closest relationships, respondents were asked five follow-up questions about the alter's gender, education, employment status, frequency of contact and the kinship relation between respondent and alter. These follow-up questions were limited to the three closest friends in an attempt to constrict the incentive to cheat. (The questionnaire can be found at http://doku.iab.de/fdz/pass/Questionaires_English_W3.zip). While the same two initial questions on personal network size had been asked in the first two waves of the panel study, no follow-up questions had been asked in these waves. Therefore differences in response behavior between Waves 1 and 2 on the one hand and Wave 3 on the other hand can be exploited to analyze the effects of adding additional burden to certain answers to this question.

Depending on the answer to the initial two questions, in Wave 3 there can be between zero and 15 follow-up questions. The easiest way for interviewers to reduce their workload would be to enter "no", "refused" or "don't know" without even asking the initial question, thus skipping the whole set of follow-up questions. An alternative would be to ask and record the first question truthfully and then enter a number smaller than three for the number of close friends (irrespective of whether the second question was actually asked or what answer was given). At the same time, there is hardly any chance that respondents who had never been asked this set of questions before would notice these shortcuts.

In PASS Wave 3 overall, 129 CATI interviewers conducted between two and 238 interviews (with a mean of 60) and 243 CAPI interviewers conducted one to 97 interviews (with a mean of 23). All interviewers work for the same survey organization. Nevertheless, CATI and CAPI interviewers work in completely different environments. On the one hand, modes differ in their payment scheme. While CAPI interviewers are paid the same amount for each successful interview, irrespective of its duration, CATI interviewers are paid by the hour. Consequently, interview duration should not matter for CATI interviewers, while in CAPI a shorter interview implies a higher hourly wage.

On the other hand, modes differ in the tightness of supervision: CAPI interviewers can work fairly autonomously as they organize the workload assigned to them themselves and do not face a recording of their interviews. As a means to detect falsifications, a random sample of respondents and nonrespondents is contacted by the survey organization either by postcard or by phone after the interview. These tests focus mainly on verifying the assigned disposition codes and the time and duration of the interview. It seems very unlikely that cheating with a filter question can be inferred from these tests. In contrast to this, CATI interviewers are frequently monitored by a supervisor in the call centers of the survey organization, who would likely detect any deviations from the protocol leading to a dismissal from the study in case of repeated misbehavior (cf. Büngeler et al. 2010).

All in all – unless they have a preference for creating high-quality data – CAPI interviewers face strong incentives to cheat when collecting data on network size. At the same time it seems very unlikely that there are deliberate falsifications by CATI interviewers. They do not gain anything from cheating, as a shorter interview means that they are paid less or have to contact additional respondents within their shift. Interviewer effects in CATI should thus be limited to unintentional effects. Hence, we derive Hypothesis 1a and 1b with respect to Wave 3 of PASS:

> Hypothesis 1a: The interviewer effect on network size is stronger in CAPI than in CATI.
> Hypothesis 1b: Network sizes in CAPI interviews are smaller compared to CATI interviews.

As the same network size questions had been asked in Waves 1 and 2 of the panel survey without triggering any follow-up questions, an incentive for CAPI interviewers to cheat has been introduced for the first time in Wave 3. This should reduce network size in CAPI in Wave 3 compared to network size in CAPI in Wave 2. This leads to Hypothesis 1c:

> Hypothesis 1c: The introduction of up to 15 follow-up questions in Wave 3 leads to a decrease in CAPI network size compared to Wave 2.

There is some evidence that experienced interviewers produce stronger systematic effects on filter questions than unexperienced interviewers (Hughes et al. 2002). Groves et al. (2004, 273) argue that this might be due to reward systems for CAPI interviewers focusing on productivity instead of measurement quality. Thus we derive our fourth hypothesis:

> Hypothesis 2a: The network size in CAPI mode in Wave 3 decreases with the experience of an interviewer *across* studies.

Furthermore, it is highly plausible to assume that interviewers learn how to cheat effectively during the fieldwork of a specific survey. It might take some time for interviewers to realize which questions have a huge impact on interview duration and can be manipulated easily at the same time (Matschinger et al. 2005). In line with this, we expect the cheating behavior not to be constant across time as interviewers have to exhibit a learning effect first. Thus, our fifth hypothesis is:

> Hypothesis 2b: The network size in CAPI mode in Wave 3 decreases with the experience of an interviewer *within* the study.

4. Methods

To test these hypotheses, we will for the most part use multilevel regression models (Swamy 1971; Goldstein 1986; Longford 1993) to take into account the clustering of respondents within interviewers and to determine the proportion of the total variance that can be attributed to the interviewers as a measure of interviewer influence. In the absence of a gold-standard measurement for network size, interviewer effects can be identified by estimating intra-interviewer correlations ρ_{int} of the survey measure. Within the frame of a random-effects ANOVA model, this is given by $\rho_{int} = \sigma_b^2/\sigma_b^2 + \sigma_w^2$ where σ_b^2 is the between-interviewer variance and σ_w^2 the within-interviewer variance. (We follow the notation in Rabe-Hesketh and Skrondal 2012.) We will refer to the intra-interviewer correlation either as ρ_{int} or as ICC (for intraclass correlation). In the presented form, however, ρ_{int} confounds several effects on the interviewer level: respondents assigned to different interviewers may differ in true network size from the beginning. This may be moderated further by the interviewers' differing ability to contact and convince respondents with different network sizes to participate (West et al. 2013). One approach to fix this is to include variables that explain individual network size in a random-intercept model. The intra-interviewer correlation is then only estimated based on the unexplained variance of the model. Drawing on a rich literature on network size, we decided to include gender, age, education, employment status, health, membership in organizations, and the existence of a partner and the number of children outside the person in question's own household as individual predictors of network size (van Tilburg 1998; Marsden 2003; McPherson et al. 2006; Brashears 2011; Brüderl et al. 2013; Paik and Sanchagrin 2013).

There is an incentive to create networks of sizes smaller than three. For each network person reported less than that, the workload for the interviewer decreases by five questions. However, once the size of three has been reached, it does not matter whether three or more persons are named. While this data structure might be analyzed using double-hurdle models (Cragg 1971) (where step one models whether a network size smaller than three is reported and step two models how much smaller it is) or ordinal

regression models (O'Connell 2006) that use three or more as maximum category, we decided to choose a more straightforward analysis strategy here. In the subsequent analysis, we will only distinguish between networks of size three or more, where the maximum number of follow-up questions was asked, and networks of size smaller than three, where some number of questions were skipped. We will refer to this variable as the "network-size dummy variable" throughout this article.

With this dummy constituting the dependent variable in our model, we suppose a logistic random-intercept model (Snijders and Bosker 2000, 207 ff.; Rabe-Hesketh and Skrondal 2012, 520 ff.). As a starting point we use an empty model

$$logit\{\Pr(y_{ij}=1|x_{ij},z_j,\zeta_j)\}=\beta_1+\zeta_j \tag{1}$$

where Y is the dependent variable that takes on a value of 1 if a respondent i's network is size three or larger. It only includes a mean intercept β_1 and a random intercept $\zeta \sim N(0,\sigma_b)$ that represents the deviation of interviewer j's intercept from β_1. Note that there is no respondent-specific error term in (1). To be able to compute ρ_{int}, a latent-response formulation can be used where an error term ε_i is assumed to have a standard logistic distribution with mean zero and within-respondent variance $\sigma_w=\pi^2/3=3.29$ (Snijders and Bosker 2000, 223 ff.; Rabe-Hesketh and Skrondal 2012, 510 ff.). In a second step, we will extend our model by explanatory variables to a full random-intercept model

$$\begin{aligned} logit\{\Pr(y_{ij}=1|x_{ij},z_j,\zeta_j)\}=&\beta_1+\beta_2x_{ij2}+\ldots+\beta_kx_{ijk}\\ &+\beta_{k+1}z_{jl}+\ldots+\beta_{k+l}z_{jl}+\zeta_j \end{aligned} \tag{2}$$

where X_{ij} are predictors for network size on the respondent level and Z_j are predictors on the level of the interviewer.

The random-intercept model in (2) assumes that the slope of the dependent variable by interviewer experience is constant across interviewers and that each additional interview has the same effect on log odds. It is highly plausible, however, that interviewers differ in their motivation to work as an interviewer and thus in their receptiveness to incentives to reduce workload at the cost of data quality. Accordingly, we can introduce a random slope for interview sequence (here denoted as x_{ijk}) to model its effect heterogeneity across interviewers, resulting in a specific coefficient for interview sequence for each interviewer:

$$\begin{aligned} logit\{\Pr(y_{ij}=1|x_{ij},z_j,\zeta_{j1},\zeta_{j2})\}=&\beta_1+\beta_2x_{ij2}+\ldots+\beta_kx_{ijk}\\ &+\beta_{k+1}z_{jl}+\ldots+\beta_{k+l}z_{jl}+\zeta_{j1}+\zeta_{j2}x_{ijk} \end{aligned} \tag{3}$$

We estimated all random-intercept models using the xtlogit-command for random-effects estimation in Stata 12.1 and estimated the random-slope model using the xtmelogit command. Significance tests for random intercepts and random slopes were performed using likelihood-ratio tests compared to models without random intercept or slope. Stata approximates log likelihoods using adaptive Gaussian quadrature (AGQ)

for the xtlogit and AGQ with the multicoefficient extension from Pinheiro and Bates (1995) and the multilevel extension from Pinheiro and Chao (2006) for xtmelogit (cf. StataCorp 2011).

To test Hypothesis 1c, we will make use of the data's longitudinal nature. Within a difference-in-difference approach (Lechner 2011), we compare differences in CAPI network sizes between Waves 2 and 3 to differences in CATI network sizes between Waves 2 and 3 to assess whether the introduction of follow-up questions for each network person brought about a negative treatment effect on network size. We assume that a treatment is only introduced in CAPI, as CATI interviewers face no changed incentives through the change of the question's character.

5. Results

5.1. Descriptive Results

Table 1 shows average network sizes and proportions of networks smaller than three by mode for Waves 2 and 3. Differences in Wave 3 outcomes between modes are obvious: while only 14.1% of the CATI respondents exhibit small networks, 40.2% of the CAPI respondents do so.

The large mode differences suggest that CAPI interviewers were indeed tempted by the incentive to cheat. However, without more rigorous models these differences might be due to self-selection of isolated respondents to the CAPI part of the study. As all respondents whose phone number could not be identified from the sampling frame or from a telephone directory search were approached in CAPI mode, this alternative hypothesis gains credibility: isolated persons should be more likely to have an unlisted number or no phone number at all. When comparing results for Wave 3 to results for Wave 2 in Table 1, one gains the impression that CAPI networks were indeed smaller even before the incentive to cheat was introduced. In Wave 2, 28.7 percent of CAPI networks, but only 13.6 percent of CATI networks were of a size smaller than three. However, there is a striking growth in small networks in CAPI that cannot be found in CATI.

Further insights can be gained by investigating the proportion of networks of size three or larger by interviewer. For this purpose, all interviewers with less than ten interviews are excluded. Among the remaining 116 CATI interviewers, the standard deviation in the proportion of networks of size three or more is 0.076, while among the remaining 165 CAPI interviewers it is 0.247, that is, CAPI interviewers differ more from each other in the network sizes they produce.

Table 1. Network size by mode and wave

		Proportion of networks smaller than three (in %)	Average network size (std dev)	*n*
CATI	Wave 2	13.6	8.82 (8.89)	7,877
	Wave 3	14.1	8.25 (8.09)	7,771
CAPI	Wave 2	28.7	6.61 (8.28)	4,567
	Wave 3	40.2	4.98 (6.49)	5,633

Figure 1 shows a quantile-quantile plot of the network-size dummy variable by interviewers for CATI and CAPI. It plots quantiles of the distribution of the network-size dummy variable by interviewers in one mode against quantiles of the same distribution in the other mode.

This figure shows that differences between modes are most pronounced when the lower quantiles of the distribution are compared. For ease of interpretation, selected quantiles (10, 25, 50, 75, 90) are highlighted in the plot. For example, the ten-percent quantile is at 0.211 for CAPI, signifying that the ten percent of CAPI interviewers who produce networks of size three or more least frequently do so in at most 21.1 percent of their interviews. In contrast, the ten-percent quantile among CATI interviewers is at 0.778.

This difference of 0.567 declines via 0.218 for the median to 0.085 for the last decile. These descriptive results seem to point to a subgroup of CAPI interviewers who produce very small proportions of networks of size three or more, while CATI interviewers are much more homogenous in their results.

Figure 2 displays by mode how the proportion of networks of size three or more changes within the sequence of interviews conducted by an interviewer. A locally weighted regression using the Stata command "lowess" (Cleveland 1979) was used to smooth the curve. In order not to give the few interviewers with more than 50 interviews too much weight, the figure is cut off at the 50th interview.

As could be expected considering there is no incentive to shorten an interview, in CATI mode there is no visible trend from the first to the last interview. In CAPI, however, we find no decreasing trend as we might have expected. On the contrary, the proportion of respondents with networks of size three or more shows a small decline across the first few interviews and then increases from about the tenth interview.

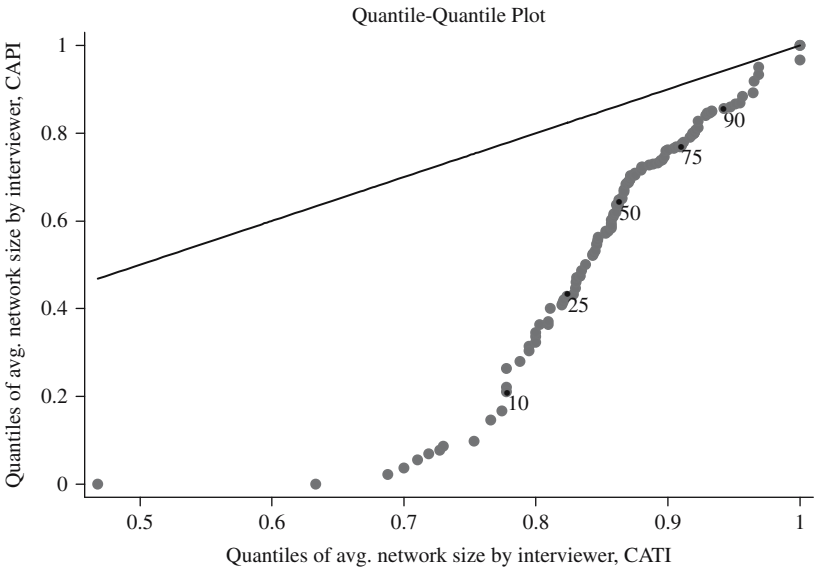

Fig. 1. Quantile-quantile plot of the network-size dummy variable by interviewer for CATI and CAPI (only interviewers with ten or more interviews)

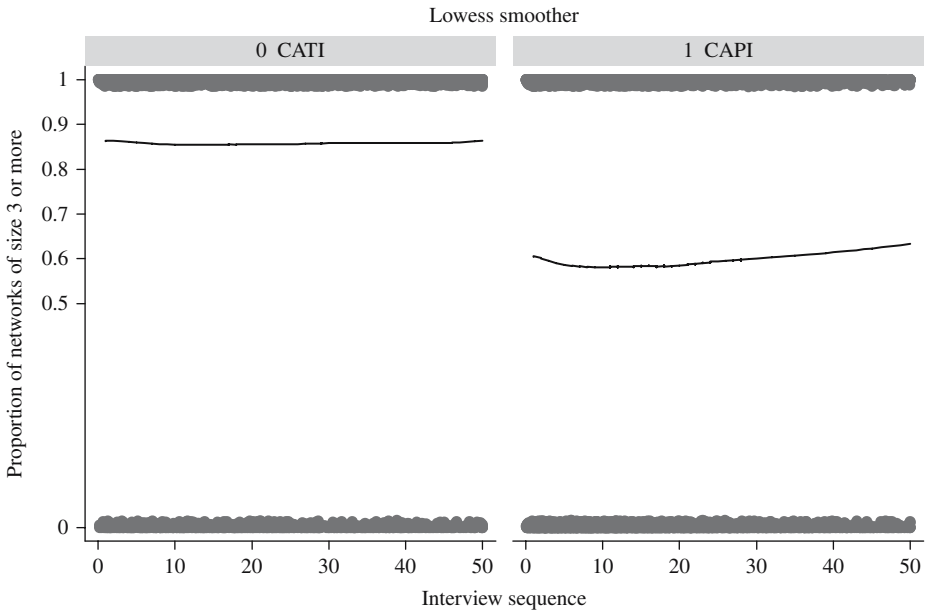

Fig. 2. Proportions of networks of size three or more by interview sequence in CATI and CAPI

5.2. Results from a Difference-in-Difference Approach

In this section the longitudinal character of the study is exploited. As the network-size question was not used as a filter question in Wave 2, there was no incentive to cheat on this question for any interviewer in Wave 2. The introduction of up to 15 follow-up questions in Wave 3 brought about an incentive to cheat for CAPI interviewers only. Thus, a difference-in-difference approach (DiD in the following), comparing the differences between Wave 2 and Wave 3 outcomes for CATI and CAPI interviews, can be used to investigate the effect of introducing follow-up questions.

While the number of extra questions is proportional to the network size for up to three network contacts, it remains constant for all network sizes exceeding three. In this section we will therefore use a truncated network-size variable that is equal to the network size for networks sized three or smaller and three for networks larger than three. This variable can be interpreted as the number of loops over the five-name interpreter questions in Wave 3. Of course this is hypothetical in Wave 2, where there were no actual loops.

In contrast to Table 1, where all cases were included that were interviewed in either Wave 2 or 3, Table 2 focuses only on individual respondents who participated in both of these waves. Altogether, there are 9,113 such respondents with valid network size in both waves. Column (3) contains the difference between (2) and (1). A negative number corresponds to a decrease in the workload for the respective group.

We will distinguish between five different groups of respondents, depending on the interview mode they were interviewed in. Those interviewed in CATI in Waves 2 and 3 can serve as a reference group because cheating incentives have remained unchanged. In this large group ($n = 5{,}330$), the mean number of loops has remained almost constant (2.751 to 2.745). The second group was switched from CAPI in Wave 2 to CATI in Wave 3. Although

Table 2. Network-size differences between Wave 2 and 3 by group

Group	(1) Mean number of loops Wave 2	(2) Mean number of loops Wave 3	(3) Difference (2)-(1)	(4) Diff-in-Diff (compared to Group 1)	(5) *n*
1 CATI -> CATI	2.751	2.745	−0.006	–	5,330
2 CAPI -> CATI	2.264	2.622	0.358	0.365 ($p = 0.015$)	53
3 CATI -> CAPI	2.693	2.391	−0.301	−0.295 ($p < 0.001$)	322
4 CAPI -> CAPI (different ivwer)	2.519	2.359	−0.160	−0.154 ($p = 0.066$)	412
5 CAPI -> CAPI (same ivwer)	2.455	2.190	−0.265	−0.259 ($p < 0.001$)	2,996
Total					9,113

there should be no change in incentives in this group, it is the only one where the number of loops increased from 2.264 to 2.622. Note, however, that this group is rather small ($n = 53$). In addition, in the PASS panel design switches from one mode to another are mainly due to the interviewer's inability to conduct another interview with the respondent in the original mode, which might for example be due to the respondent having moved. Thus, whatever caused those respondents to switch modes might have influenced their network sizes as well.

In all other groups we expect negative changes in the number of loops due to the introduction of an incentive to cheat. For respondents who were switched from CATI in Wave 2 to CAPI in Wave 3 and respondents who were interviewed in CAPI mode in both waves but by different interviewers, CAPI interviewers have not worked these cases in the previous wave and thus cannot have any knowledge of the previous-wave network size. As expected, the number of loops decreases from 2.693 in Wave 2 to 2.391 in Wave 3 for respondents who switched from CATI to CAPI. Again, one should be aware that this is a selective group.

Another 412 respondents (Group 4) were interviewed in CAPI mode in Waves 2 and 3, but the interviewer was switched. Again an incentive to cheat has been introduced and no prior knowledge of network size is available to the interviewers. Network size decreases as expected from 2.519 to 2.359.

The last group consists of 2,996 cases that remained in CAPI in both waves with the same interviewer. This group is special in that an incentive to cheat has been introduced, but interviewers might still be aware of the respondent's network size from the previous year, which might make them more careful as they anticipate the survey organization's ability to check for consistency across waves. Nevertheless, the decline in the number of loops (2.455 to 2.190) is very pronounced in this group.

In a next step, the differences in all other groups will be compared to the difference for the reference group of CATI stayers (Group 1). This DiD is negative for all three groups in which a cheating incentive was introduced and positive only for CAPI to CATI changers where there never was an incentive to cheat. All differences are significant at a ten-percent level (two-sided t-tests taking clustering within interviewers into account), and only the relatively small group of CAPI stayers with different interviewers fails to reach the five-percent significance level.

Only comparing those respondents for whom interviewers never had an incentive to cheat (Groups 1 and 2) and those for whom an incentive to cheat was introduced (Groups 3 to 5), the DiD of -0.254 indicates a highly significant treatment effect which supports hypothesis 1c. The more detailed analysis in Table 2, however, shows that something else is going on in the data. CAPI networks were already smaller in Wave 2, which might be the result of a selection effect (poorly connected persons are less likely to have listed phone numbers). However, the results for the small group of respondents who switched from CAPI to CATI suggests that there might also be a mode effect on networks in the absence of incentives to cheat. We can only speculate at this point whether this might be due to differences in standardization and probing related to differences in fieldwork monitoring. As stated above, an alternative explanation for the increased network size of these mode-switchers is that the mode switch is due to changes in their life circumstances that also affected network size. Thus we should be careful about causal interpretations, as Group (2) is likely to be selective in that it contains a relatively large proportion of respondents whose life circumstances have changed.

Table 3. Empty model for CATI and CAPI

	CATI (Model 1)	CAPI (Model 2)
	β (s.e.)	β (s.e.)
Fixed part		
Person-level predictors	Not included	Not included
Interviewer-level predictors	Not included	Not included
β_1	1.900 (0.051)***	0.479 (0.091)***
Random part		
σ_b	0.394 (0.051)***	1.201 (0.081)***
ρ_{int}	0.045 (0.011)***	0.305 (0.028)***
n	7,402	5,348
Log Likelihood	−2901.773	−3179.452

Dependent variable: network-size dummy (1 = three and more contacts, 0 = less than three contacts)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.3. Results from Multilevel Models

To further investigate the mode differences found in the descriptive analysis, empty random-intercept models without any covariates were estimated for both CATI (Model 1) and CAPI (Model 2) as an initial step. The results can be found in Table 3.

In CATI there is a significant ICC of $\rho_{int} = 0.045$. This is well within the range usually found for interviewer effects (Groves 1989, 319). In contrast, in CAPI the ICC amounts to $\rho_{int} = 0.305$ which constitutes a huge interviewer effect. Interviewer effects of this size are extremely rare. Note, however, that these are ICCs from the empty model. They might at least partly reflect differences in interviewer workload. Therefore, in a next step we controlled for individual-level covariates that explain individual network size. Simultaneously we controlled for observed interviewer-level covariates. In a first step, only interviewer demographics (gender, age, and education) are included. Results can be found in Table 4.

Age in years is centered around its mean and respondents' education is coded in three categories for "low", "medium", and "high" educational background, according to the German educational system. For employment status, we created four dummy variables that comprise unemployed people, those who are still in education or military service, and those not in the labor force, whereas employed persons served as the reference category. The health measure in PASS are the so-called SF-12v2 indicators, which constitute an internationally accepted inventory of health measures (Nübling et al. 2007). Of these, the two superordinate scales "physical health" and "mental health" were generated. To account for the social contacts a person's leisure activities entail, we include a dummy variable that indicates the respondent's active engagement in any kind of club or organization. Furthermore, we generated one dummy variable indicating whether the respondent has a partner outside the household. The network-size question only refers to persons outside the respondent's household and we expect that the partner will usually be among the named persons if this applies to her or him.

The composition with respect to observed person-level variables and observable interviewer characteristics only plays a very minor role in explaining intra-interviewer correlations in network size in both modes. The ICC for CATI falls from 0.045 to 0.040,

Table 4. Models for CATI and CAPI including respondent- and interviewer-level predictors

	CATI (Model 3)	CAPI (Model 4)
	β (s.e.)	β (s.e.)
Fixed Part		
Person-level predictors		
Gender (ref. = female)	−0.383 (0.073)***	−0.291 (0.068)***
Age centered	−0.005 (0.003)	−0.007 (0.003)**
Still in school (ref. = low education level)	−0.051 (0.203)	0.289 (0.201)
Medium education level	0.317 (0.086)***	0.304 (0.083)***
High education level	0.498 (0.095)***	0.704 (0.095)***
Partner outside of household	0.542 (0.115)***	0.498 (0.108)***
Physical health	0.007 (0.004)*	0.008 (0.004)*
Mental health	0.032 (0.003)***	0.027 (0.003)***
Unemployed (ref. = employed)	−0.369 (0.090)***	−0.229 (0.087)**
Student/military service	−0.250 (0.165)	0.190 (0.163)
Retired/homemaker/parental leave	0.045 (0.120)	0.074 (0.107)
Active engagement in any organization	0.430 (0.076)***	0.589 (0.078)***
Interviewer-level predictors		
Gender (ref. = female)	0.044 (0.109)	0.207 (0.193)
Age centered	0.001 (0.005)	−0.029** (0.010)
Medium education level (ref. = low education level)	0.188 (0.335)	−0.238 (0.252)
High education level	−0.004 (0.327)	0.000 (0.274)
β_1	−0.278 (0.413)	−1.204 (0.371)**
Random Part		
σ_b	0.370 (0.051)***	1.201 (0.082)***
ρ_{int}	0.040 (0.011)***	0.305 (0.029)***
n	7,402	5,348
Log Likelihood	−2745.015	−2989.010

Dependent variable: network-size dummy (1 = three and more contacts, 0 = less than three contacts)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

while in CAPI it even remains constant at 0.305. Thus, differences in network size between interviewers can be attributed neither to their demographic characteristics nor to characteristics of the respondents they interviewed.

In spite of the huge difference in interviewer effects between the two modes, the variables explaining the network size are strikingly similar in both modes. Coefficients in Models 3 and 4 point in the same direction and are similar in size with only two exceptions, which both comprise very small numbers of cases and are not significant in any of the models. Models for both of the modes consistently show larger networks for women, medium or highly educated respondents, respondents with a partner outside their household, physically and mentally healthy respondents and smaller networks for unemployed than for employed. Age is significant only in the CAPI model, where younger respondents tend to have larger networks. However, the CATI coefficient at -0.005 is very similar to the CAPI coefficient at -0.007. Furthermore, the age of the interviewer has a negative effect on network size only in CAPI.

Turning to Hypothesis 1a, Models 3 and 4 provide evidence confirming that the interviewer effect is much larger in CAPI than in CATI after controlling for composition effects and observable interviewer characteristics. While separate models for CATI and CAPI are well suited to showing the different magnitude of the intra-interviewer correlation in CATI and CAPI, providing support for Hypothesis 1a, these models do not show that network size is downward biased in CAPI after controlling for respondent composition (Hypothesis 1b). This can be done by employing the same respondent- and interviewer-level variables as in Model 4 but including all cases from both modes and adding a CAPI indicator to measure the effect of CAPI interviews compared to CATI interviews. These results can be found in Model 5. Table 5 shows only the coefficient of the additional CAPI variable and the variance components. As expected, the CAPI effect is strongly negative and highly significant, which implies smaller networks in CAPI mode and thus supports Hypothesis 1b.

So far, we have identified an interviewer effect in CAPI that is much larger than the interviewer effect in CATI and leads to significantly smaller networks after controlling for observable person- and interviewer-level predictors. However, the larger interviewer effects in CAPI might be explained by unobserved differences in the assignment to interviewers. In centralized CATI studios, interviewer assignment to respondents is close to random within a given shift. In contrast, in CAPI interviewer workload is assigned much more selectively, as all cases interviewed by one interviewer usually reside in the same region in one or two sampling points. On average, interviewers have performed 77 percent of their workload in their main sampling point. Thus differences in true network size between regions constitute a serious potential alternative explanation for the larger ICC in CAPI mode. In addition, the strong confounding of interviewers and areas makes the use of cross-classified random-effects models (Rasbash and Goldstein 1994) for the separation of interviewer and sampling point effects unadvisable (Vassallo et al. 2016).

In contrast to previous studies, the PASS data offer a rare opportunity to control for regional differences in network size as an alternative explanation for interviewer variances. Model 6 again uses CAPI cases only. The average network size from all CATI interviews in the same sampling point is included as an additional control for regional differences in expansiveness of networks. Again, Table 5 only shows the coefficients of

Table 5. Models with additional explanatory variables

	CATI + CAPI (Model 5)	CAPI (Model 6)	CAPI (Model 7)	CAPI (Model 8)
	β (s.e.)	β (s.e.)	β (s.e.)	β (s.e.)
Fixed Part				
Person-level predictors	included	included	included	included
Interviewer-level predictors	included	included	included	included
CAPI	−1.057 (0.192)***			
Avg. CATI netw. size in PSU		0.059 (0.038)	0.064 (0.038)	0.066 (0.039)
Sequence			0.004 (0.004)	0.003 (0.003)
Previous wave			−0.340 (0.265)	−0.346 (0.266)
β_1	−0.368 (0.273)	−1.722 (0.498)***	−1.558 (0.538)**	−1.553 (0.545)
Random Part				
σ_b	0.938 (0.054)***	1.202 (0.082)***	1.195 (0.081)***	1.210 (0.102)***
ρ_{int}	0.211 (0.019)***	0.305 (0.029)***	0.303 (0.029)***	
$\sigma_{sequence}$				0.012 (0.005)
Corr(ζ_j, $\sigma_{sequence}$)				−0.139 (0.346)
n	12,750	5,348	5,348	5,348
Log Likelihood	−5780.530	−2987.884	−2985.913	−2985.615

Dependent variable: network-size dummy (1 = three and more contacts, 0 = less than three contacts)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

interest. The coefficient for average CATI network size in the sampling point is in the expected direction.

It is, however, not significant on a five-percent level. More importantly, inclusion of this variable has no influence on the ICC, which is still 0.305 after controlling for regional differences.[3] It should be noted, however, that the average network size per sampling unit constitutes an imperfect measure for regional differences, as a proper disentangling of interviewer and area effects would require an interpenetrated sample design (Campanelli and O'Muircheartaigh 1999). Nevertheless, results are still in favor of Hypothesis 1a after controlling for regional expansiveness of networks.

We now turn to Hypothesis 2a and 2b. Here we test whether interviewers exhibit a learning behavior over time – within the study as well as across studies. While we are able to create an exact measure for experience within the study, the number of previous interviews within the study, the survey organization did not provide an equally good measure for overall experience as an interviewer. The only available indicator is whether an interviewer already conducted interviews in previous PASS waves.[4] This measure is probably correlated only weakly with general interviewer experience. Those who carried out previous PASS interviews have been interviewing for at least one or two years. With primary sampling units being constant, interviewers who did their first PASS interview in Wave 3 will often be newly hired interviewers. Of all interviewers in PASS Wave 3, 19 percent had no experience within the study.

Model 7 in Table 5 includes both indicators for experience in addition to all variables from Model 6. Both indicators have no significant effect. The coefficient for previous wave PASS experience is negative, as expected. As the indicator is quite crude, it remains unclear whether more experience as an interviewer leads to a reduction in network size. The coefficient for the number of previous interviews within the wave (sequence) surprisingly is positive, that is, CAPI interviewers tend to produce larger networks in later interviews. Thus, there is definitely no general support for Hypothesis 2b, which presumed learning effects across time within a wave. However, interviewers might differ with respect to their reaction to incentives to cheat and in their learning behavior. It thus makes sense to relax the assumption that the slope with respect to sequence is identical for all interviewers. In the multilevel framework this can naturally be done by replacing the fixed slope for sequence by a random slope, which was done in Model 8. This random slope fails to reach the five-percent significance level when – as is advised by Rabe-Hesketh and Skrondal (2012) – we use a likelihood-ratio test between Models 8 and 7. The other coefficients are scarcely affected by the introduction of the random slope.

Models 7 and 8 thus suggest that there is neither a general trend towards smaller network sizes within the temporal sequence of interviews of one interviewer, nor a significant variation between interviewers with respect to such learning effects. This seems

[3] We also estimated a CATI model (not displayed) in which we replaced interviewers by PSUs as level 2 units. The ICC after controlling for observable person- and interviewer-level variables is only 0.013 and not statistically significant. We interpret this as further evidence that regional differences cannot be considered the reason for interviewer effects.

[4] We should recall that in previous waves the network size question did not serve as a filter and thus direct learning is not the mechanism identified by this measure.

surprising and is in contrast to earlier findings by Matschinger et al. (2005) and Kosyakova et al. (2015). One reason might be that the potential to abbreviate the interview by reporting small networks is very obvious and interviewers learn it so quickly that it does not show up in a linear trend across the whole fieldwork. We thus re-estimated Model 7 ten times using only the first five to 14 interviews for each interviewer. The slope for experience within a wave indeed changes to a negative sign in nine out of these ten models. However, it only becomes significant when taking the first nine or the first twelve interviews of each interviewer. We conclude from this that there might be some learning effects during the initial interviews of each interviewer, but that the evidence for this is in no way convincing.

Models 1 to 8 exclusively make use of Wave 3 data. While these models cannot exactly be replicated using Waves 1 and 2, as some constructs like physical or mental health have only been collected in Wave 3, estimating similar models for previous waves can provide valuable hints. If the increased interviewer effect in CAPI was exclusively driven by the incentive to cheat, then Wave 1 and 2 interviewer effects for network size in CAPI mode should be close to the size that is observed in CATI, and thus it is sufficient to focus on the ICC. Re-estimating a model similar to our Model 6 for Waves 1 and 2 (excluding health variables) and controlling for possible sample-point differences results in ICCs which are substantially higher for CAPI (0.158 in Wave 1 and 0.203 in Wave 2) than those we found for CATI, even before an incentive to abbreviate interviews was introduced. At the same time they are much smaller than the ICC of 0.303 found in Model 6 for Wave 3.

Like the difference-in-difference approach, this suggests that two mechanisms are at work at the same time. The increase in ICC from Waves 1 and 2 to Wave 3 indicates that cheating might play a role, while the difference between interviewer effects in CATI and interviewer effects in CAPI in Waves 1 and 2 indicates that other differences between CATI and CAPI interviews play a role as well. We will return to this in the discussion section.

6. Summary and Discussion

We have found large interviewer effects for a network-size filter question in a large-scale German panel survey. While there is a significant interviewer effect in CATI, it is far exceeded by the interviewer effect in CAPI. This latter effect remains identical in size when observed respondent- and interviewer-level variables are included and even when average network size in a region (measured independently) is controlled for. In contrast to earlier research on interviewer effects on filter questions (Matschinger et al. 2005; Kosyakova et al. 2015), there is no evidence for a general learning effect of the same interviewer across interviews within a wave of data collection or for interviewer-specific differences in these learning effects.

How can we interpret these findings? Past studies have attributed large interviewer effects for network generators to differences in probing behavior (van Tilburg 1998; Marsden 2003). These studies investigated complex name generators that involved a lot of probing. The network-size question under consideration here is less complex. Nevertheless, interviewer effects are even larger. Furthermore, large interviewer effects can only be found in CAPI, where interviewers have an incentive to cheat, while interviewer effects in CATI are comparably small. In addition, the findings from our DiD

approach suggest that introducing an incentive for CAPI interviewers to produce smaller networks actually results in reduced network size. This leads us to the conclusion that purposeful manipulation of answers on the side of the interviewers – and not only differences in probing – is one likely explanation.

However, this study has some limitations. Interviewer effects in CAPI were much larger than in CATI in Waves 1 and 2, where no additional follow-up questions were triggered and thus manipulation on the side of the CAPI interviewers has no obvious payoff. Furthermore, a difference-in-difference analysis indicates that respondents who were switched from CAPI in Wave 2 to CATI in Wave 3 show significantly increasing network sizes, although in both cases there were no incentives to produce smaller or larger networks. This indicates that other mechanisms are at work as well. Likely candidates are probing or other deviations from standardized interviewing, which are more likely to happen in the less well supervised CAPI field.

Investigating this in detail would require more detailed information on CAPI fieldwork. The IAB, as the institution responsible for PASS, has set up a project for this purpose and recorded CAPI interviews as part of this research. This might help us to gain a better understanding of the origin of large interviewer effects in CAPI even in the absence of incentives to shorten an interview. Other possible explanations include that interviewer characteristics not controlled for in the models could be more relevant in CAPI than in CATI interviews, or that unobserved respondent characteristics lead to smaller networks in CAPI.

We used a rather simple analysis strategy that is adequate in our view as it reflects interviewer incentives. We only distinguished between networks of size three and larger and networks smaller than three. We performed several sensitivity checks that all replicate the main findings, but result in somewhat smaller ICCs. Including network size as a metric variable in a random-effects linear regression results in an ICC of 0.16, while truncating the network variable at three and using a random-effects tobit model results in an ICC of 0.24. Thus, our simple operationalization seems to grasp what creates the largest unexplained differences between interviewers.

Our findings are limited to one study in Germany. The recent publication by Brüderl et al. (2013) finds interviewer effects of a similar magnitude for a second study in Germany. This study used the same fieldwork agency and thus probably even shared interviewers with PASS. Given that the studies by Matschinger et al. (2005) and Kosyakova et al. (2015) also use German data, one might suspect that this problem is particularly prevalent in Germany. Given the results of the recent study by Paik and Sanchagrin based on the US General Social Survey it is obvious, however, that the problem is not limited to Germany. Our evidence suggests that more rigorous supervision techniques should be used in CAPI surveys to counteract incentives to shorten the interview by cheating on questions and thereby impairing question validity. Suitable techniques include re-interviews (Biemer and Stokes 1989) or statistical approaches to detect potential falsifiers (Biemer and Stokes 1989; Bredl et al. 2012). In addition, recordings of at least a considerable proportion of CAPI interviews by each interviewer could be made mandatory.

Alternatively, payment schemes might be changed so that interviewers are paid better for longer interviews. It is not easy, though, to find an increase rate that makes most

interviewers indifferent to the length of the interview. If the premium for long interviews is too high, adverse incentives to the ones observed in this study might arise and interviewers might try to artificially stretch interviews. The good news of this article seems to be that in spite of falsifications by a large proportion of interviewers that lead to a pronounced downward bias of average network size, regression coefficients in models explaining network size seem to be only marginally affected. The size and direction of effects of respondent characteristics on network size is very similar between CAPI interviews and CATI interviews that offer no incentive to cheat, and both are in line with the previous literature.

The question whether this still holds for longitudinal analyses is beyond the scope of this article. The data analyzed here are from the third wave of the PASS panel. The same network-size question had been used in two prior waves without any follow-up questions. The consequence is a marked decrease in average network size from Wave 2 to 3, followed by another increase from Wave 3 to 4 where incentives to cheat diminished again as no follow-up questions were asked. It is obvious that this biases estimates of average network size across time. This is not a new finding – several authors have argued that findings about declining social capital (Putnam 1995; McPherson et al. 2006) might be artifacts (Fischer 2009; Paik and Sanchagrin 2013).

A related question is how coefficients of models for longitudinal data that use either network size as a predictor or as an outcome are affected. This topic is a matter for future research.

7. References

American Association for Public Opinion Research (AAPOR). 2003. *Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection, and Repair of its Effects*. Available at: https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf (accessed April 6th, 2016).

Bethmann, A. and D. Gebhardt. 2011. User Guide "Panel Study Labor Market and Social Security" (PASS) ∗ Wave 3. *FDZ Datenreport, 04/2011 (en)*, Nuremberg. Available at: http://doku.iab.de/fdz/reporte/2011/DR_04-11_EN.pdf (accessed April 6th, 2016).

Biemer, P.P. and S.L. Stokes. 1989. "The Optimal Design of Quality Control Samples to Detect Interviewer Cheating." *Journal of Official Statistics* 5: 23–39.

Biemer, P.P. 2010. "Overview of Design Issues: Total Survey Error." In *Handbook of Survey Research*, 2nd ed., edited by P.V. Marsden and J.D. Wright, 27–58. Bingley: Emerald.

Blasius, J. and J. Friedrichs. 2013. "Faked Interviews." In *Methods, Theories, and Empirical Applications in the Social Sciences*, edited by S. Salzborn, E. Davidov, and J. Reinecke, 49–56. Wiesbaden: VS Verlag für Sozialwissenschaften.

Brashears, M.E. 2011. "Small Networks and High Isolation? A Reexamination of American Discussion Networks." *Social Networks* 33: 331–341. Doi: http://dx.doi.org/10.1016/j.socnet.2011.10.003.

Bredl, S., P. Winker, and K. Koetschau. 2012. "A Statistical Approach to Detect Interviewer Falsification of Survey Data." *Survey Methodology* 38: 1–10.

Brüderl, J., B. Huyer-May, and C. Schmiedeberg. 2013. "Interviewer Behavior and the Quality of Social Network Data." In *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention*, edited by P. Winkler, R. Porst, and N. Menold, 147–160. Frankfurt: Peter Lang.

Büngeler, K., M. Gensicke, J. Hartmann, R. Jäckle, and N. Tschersich. 2010. *IAB-Haushaltspanel im Niedrigeinkommensbereich Welle 3 (2008/09): Methoden- und Feldbericht. FDZ Methodenreport, 10/2010 (de)*, Nuremberg. Available at: http://doku.iab.de/fdz/reporte/2010/MR_10-10.pdf (accessed April 6th, 2016).

Campanelli, P. and C. O'Muircheartaigh. 1999. "Interviewers, Interviewer Continuity, and Panel Survey Nonresponse." *Quality & Quantity* 33: 59–76. Doi: http://dx.doi.org/10.1023/A:1004357711258.

Cleveland, W.S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplot." *Journal of the American Statistical Association* 74: 829–836.

Cragg, J.G. 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica* 39: 829–844. Doi: http://dx.doi.org/10.2307/1909582.

Crespi, L.P. 1945. "The Cheater Problem in Polling." *Public Opinion Quarterly* 9: 431–445.

Eckman, S., F. Kreuter, A. Jäckle, A. Kirchner, S. Presser, and R. Tourangeau. 2014. "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys." *Public Opinion Quarterly* 78: 721–733. Doi: http://dx.doi.org/10.1093/poq/nfu030.

Fischer, C.S. 2009. "The 2004 GSS Finding of Shrunken Social Networks: An Artifact?" *American Sociological Review* 74: 657–669. Doi: http://dx.doi.org/10.1177/000312240907400408.

Freeman, J. and E.W. Butler. 1976. "Some Sources of Interviewer Variance in Surveys." *Public Opinion Quarterly* 40: 79–91. Doi: http://dx.doi.org/10.1086/268-269.

Goldstein, H. 1986. "Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares." *Biometrika* 73: 43–56. Doi: http://dx.doi.org/10.1093/biomet/73.1.43.

Groves, R.M. and N.H. Fultz. 1985. "Gender Effects among Telephone Interviewers in a Survey of Economic Attitudes." *Sociological Methods Research* 14: 31–52. Doi: http://dx.doi.org/10.1177/0049124185014001002.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2004. *Survey Methodology*. Hoboken, NJ: Wiley.

Guterbrock, T.M. 2008. "Falsifications." In *Handbook of Survey Research*, edited by P.J. Lavrakas, 267–270. Los Angeles: Sage.

Huddy, L., J. Billig, J. Bracciodieta, L. Hoeffler, P.J. Moynihan, and P. Pugliani. 1997. "The Effect of Interviewer Gender on the Survey Response." *Political Behavior* 19: 197–220. http://dx.doi.org/10.1023/A:1024882714254.

Hughes, A., J. Chromy, K. Giacoletti, and D. Odom. 2002. "Impact of Interviewer Experience on Respondent Reports of Substance Use." In *Redesigning an Ongoing National Household Survey*, edited by J. Gfroerer, J. Eyerman, and J. Chromy, 161–184. Washington: Substance Abuse and Mental Health Services Administration.

Kosyakova, Y., J. Skopek, and S. Eckman. 2015. "Do Interviewers Juggle Filter Questions? Evidence from a Multilevel Approach". *International Journal of Public Opinion Research* 27: 417–431. Doi: http://dx.doi.org/10.1093/ijpor/edu027.

Kreuter, F., S. McCulloch, S. Presser, and R. Tourangeau. 2011. "The Effects of Asking Filter Questions in Interleafed Versus Grouped Format." *Sociological Methods & Research* 40: 88–104. Doi: http://dx.doi.org/10.1177/0049124110392342.

Lechner, M. 2011. "The Estimation of Causal Effects by Difference-In-Difference Methods." *Foundations and Trends in Econometrics* 4: 165–224.

Longford, N.T. 1993. *Random Coefficient Models*. Oxford: Oxford University Press.

Mangione, T.W., F.J. Fowler, and T.A. Louis. 1992. "Question Characteristics and Interviewer Effects." *Journal of Official Statistics* 8: 293–307.

Marsden, P.V. 2003. "Interviewer Effects in Measuring Network Size Using a Single Name-Generator." *Social Networks* 25: 1–16. Doi: http://dx.doi.org/10.1016/S0378-8733(02)00009-6.

Matschinger, H., S. Bernert, and M.C. Angermeyer. 2005. "An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview." *Journal of Official Statistics* 21: 657–674.

McPherson, M., L. Smith-Lovin, and M.E. Brashears. 2006. "Social Isolation in America: Changes in Core Discussion Networks over Two Decades." *American Sociological Review* 71: 353–375. Doi: http://dx.doi.org/10.1177/000312240607100301.

Nübling, M., H.H. Andersen, A. Mühlbacher, J. Schupp, and G.G. Wagner. 2007. "Computation of Standard Values for Physical and Mental Health Scale Scores Using the SOEP Version of SF12v2." *Schmollers Jahrbuch: Journal of Applied Social Science Studies* 127: 171–182. Available at: https://www.researchgate.net/publication/23645941_Computation_of_Standard_Values_for_Physical_and_Mental_Health_Scale_Scores_Using_the_SOEP_Version_of_SF12v2 (accessed April 6th, 2016).

O'Connell, A.A. 2006. *Logistic Regression Models for Ordinal Response Variables*. Thousand Oaks, CA: Sage.

O'Muircheartaigh, C. and P. Campanelli. 1998. "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 161: 63–77. Doi: http://dx.doi.org/10.1111/1467-985X.00090.

Paik, A. and K. Sanchagrin. 2013. "Social Isolation in America: An Artifact." *American Sociological Review* 78: 339–360. Doi: http://dx.doi.org/10.1177/0003122413482919.

Pinheiro, J.C. and D.M. Bates. 1995. "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." *Journal of Computational and Graphical Statistics* 4: 12–35. Doi: http://dx.doi.org/10.1080/10618600.1995.10474663.

Pinheiro, J.C. and E.C. Chao. 2006. "Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models." *Journal of Computational and Graphical Statistics* 15: 58–81. Doi: http://dx.doi.org/10.1198/106186006X96962.

Putnam, R.D. 1995. "Bowling Alone: America's Declining Social Capital." *Journal of Democracy* 6: 65–78. Doi: http://dx.doi.org/10.1353/jod.1995.0002.

Rabe-Hesketh, S. and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata. Volume II: Categorical Responses, Counts, and Survival*. College Station, TX: Stata Press.

Rasbash, J. and H. Goldstein. 1994. "Efficient Analysis of Mixed Hierarchical and Cross-Classified Random Structures Using a Multilevel Model." *Journal of Educational and Behavioral Statistics* 19: 337–350. Doi: http://dx.doi.org/10.3102/10769986019004337.

Schnell, R. 1991. "Der Einfluß gefälschter Interviews auf Survey-Ergebnisse." *Zeitschrift für Soziologie* 20: 25–35.

Schnell, R. 2012. *Survey-Interviews: Methoden standardisierter Befragungen*. Wiesbaden: VS Verlag.

Schnell, R. and F. Kreuter. 2000. "Untersuchungen zur Ursache unterschiedlicher Ergebnisse sehr ähnlicher Viktimisierungssurveys." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 52: 96–117. Doi: http://dx.doi.org/10.1007/s11577-000-0005-y.

Schnell, R. and F. Kreuter. 2005. "Separating Interviewer and Sampling-Point Effects." *Journal of Official Statistics* 21: 389–410.

Schraepler, J.P. and G.G. Wagner. 2005. "Characteristics and Impact of Faked Interviews in Surveys – An Analysis of Genuine Fakes in the Raw Data of SOEP." *Allgemeines Statistisches Archiv* 89: 7–20. Doi: http://dx.doi.org/10.1007/s101820500188.

Schuman, H. and J. Converse. 1971. "The Effects of Black and White Interviewers on Black Responses in 1968." *Public Opinion Quarterly* 35: 44–68. Doi: http://dx.doi.org/10.1086/267866.

Snijders, T.A.B. and R. Bosker. 2000. *Multilevel Analysis*. London: Sage.

StataCorp. 2011. *Stata. Longitudinal-Data/Panel-Data Reference Manual*. Release 12. College Station, TX: StataCorp.

Swamy, P.A.V.B. 1971. *Statistical Inference in a Random Coefficient Model*. New York: Springer.

Tourangeau, R. and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133: 859–883.

Tourangeau, R., F. Kreuter, and S. Eckman. 2012. "Motivated Underreporting in Screening Interviews." *Public Opinion Quarterly* 76: 453–469. Doi: http://dx.doi.org/10.1093/poq/nfs033.

Tourangeau, R., F. Kreuter, and S. Eckman. 2013. Motivated Misreporting: Shaping Answers to Reduce Survey Burden. In *Survey Measurement: Techniques and Findings from Recent Research*, edited by U. Engel, 24–41, Frankfurt: Campus.

Trappmann, M., S. Gundert, C. Wenzig, and D. Gebhardt. 2010. "PASS: a Household Panel Survey for Research on Unemployment and Poverty." *Schmollers Jahrbuch. Journal of Applied Social Science Studies* 130: 609–622 Doi: http://dx.doi.org/10.3790/schm.130.4.609.

Trappmann, M., J. Beste, A. Bethmann, and G. Müller. 2013. "The PASS Panel Survey After Six Waves." *Journal for Labour Market Research* 46: 275–281. Doi: http://dx.doi.org/10.1007/s12651-013-0150-1.

van der Zouwen, J., W. Dijkstra, and J.H. Smit. 2004. "Studying Respondent-Interviewer Interaction: The Relationship Between Interviewing Style, Interviewer Behavior, and Response Behavior." In *Measurement Errors in Surveys*, edited by P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman, 419–437. New York: Wiley.

van Tilburg, T.G. 1998. "Interviewer Effects in the Measurement of Personal Network Size. A Non-Experimental Study." *Sociological Methods and Research* 26: 300–328. Doi: http://dx.doi.org/10.1177/0049124198026003002.

Vassallo, R., G.B. Durrant, and P.W.F. Smith. 2016. Separating Interviewer and Area Effects Using a Cross-Classified Multilevel Logistic Model: Simulation Findings and Implications for Survey Designs. Submitted manuscript (available from the author on request: g.durrant@southampton.ac.uk).

West, B.T., F. Kreuter, and U. Jaenichen. 2013. "Interviewer Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?" *Journal of Official Statistics* 29: 277–297. Doi: http://dx.doi.org/10.2478/jos-2013-0023.

Received August 28, 2014
Revised June 27, 2015
Accepted September 8, 2015