Published in final edited form as:

J R Stat Soc Ser A Stat Soc. 2019 January ; 182(1): 3–35. doi:10.1111/rssa.12391.

The contributions of paradata and features of respondents, interviewers and survey agencies to panel co-operation in the Survey of Health, Ageing and Retirement in Europe

Johanna Bristle,

Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Munich, Germany

Martina Celidoni,

University of Padua – Department of Economics and Management, Padua, Italy

Chiara Dal Bianco,

University of Padua – Department of Economics and Management, Padua, Italy

Guglielmo Weber

University of Padua – Department of Economics and Management, Padua, Italy and Institute for Fiscal Studies, London, United Kingdom.

Summary

This paper deals with panel cooperation in a cross-national, fully harmonized face-to-face survey. Our outcome of interest is panel cooperation in the fourth wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). Following a multilevel approach, we focus on the contribution of *paradata* at three different levels: fieldwork strategies at the survey agency level, features of the (current) interviewer and paradata describing respondent interview experience from the previous wave. Our results highlight the importance of respondent's prior interview experience, and of interviewer's quality of work and experience. We also find that survey agency practice matters: daily communication between fieldwork coordinators and interviewers is positively associated with panel cooperation.

Keywords

attrition; field practices; interviewer effects; panel data; paradata

1. Introduction

The issue of retention in panel surveys is of paramount importance, particularly when the focus is on slow, long-term processes such as ageing. Lack of retention of subjects in longitudinal surveys, also known as attrition, accumulates over waves and particularly harms the panel dimension of the data.

* Address for correspondence: Johanna Bristle, Max Planck Institute for Social Law and Social Policy, Amalienstr.33, 80799 Munich, Germany, bristle@mea.mpg.de.

Survey participation depends on location, contact and cooperation of the sample unit (Lepkowski and Couper, 2002). In this paper, we investigate the determinants of panel cooperation – interview completion given location and contact - in the fourth wave of SHARE (the Survey on Health, Ageing and Retirement in Europe) given participation in the third wave. We focus on panel cooperation since location and contact are less problematic in a later panel wave.

As recommended by the literature on the determinants of nonresponse behaviour, we exploit information at different levels: individual and household characteristics, interviewer traits and survey design features. A contribution of this paper is its use of information gathered at the interviewer level in a harmonized, multi-country survey. A further, novel contribution lies in our investigation of the role of survey agency practices and variability.

We use a three-level logit model to estimate the determinants of retention and the variance attributable to each level: respondent, interviewer and survey agency. This model accounts for correlation in cooperation probabilities for respondents interviewed by the same interviewer and interviewers working for the same survey agency. Given the limited number of survey agencies at the third level, we provide also a simulation exercise to document how estimates behave in finite samples similar to the one we use.

The multilevel model we estimate uses survey data as well as additional paradata that are obtained as a ‘by-product of the data collection process capturing information *about that process*’ (Durrant and Kreuter, 2013). In SHARE, paradata are available on all three levels. While paradata at the individual or interviewer level have been used in this strand of the literature, information at the survey agency level has not been taken into account to explain participation. One possible reason for this gap could be that, in cross-national research, information at the survey agency level may not be available or harmonized across countries (Blom, Lynn and Jäckle, 2008) so that comparability is limited. SHARE, which provides harmonized information on elderly individuals at the European level, collected such data in wave 4. This additional source of information gives us the opportunity to investigate the nature of nonresponse also at the survey agency level.

Our approach is theoretically based on the framework of survey participation by Groves and Couper (1998), in which the factors that are expected to influence survey participation are divided into two major areas: ‘out of researcher control’ and ‘under researcher control’. In this paper, we are particularly interested in the factors that can be influenced by the researcher, namely survey agency fieldwork strategies, the features of the interviewer and the respondent–interviewer interaction.

We find that variables at all three levels affect the probability of retention. Respondent and interviewer characteristics play an important role. Respondent cooperation decisions are affected by their previous interview experience: for instance, item nonresponse in a previous wave reduces the likelihood of cooperation in a later wave. As far as interviewer characteristics are concerned, we find that previous experience with working as a SHARE interviewer matters more than socio-demographic characteristics, such as age, gender or education. Further, interviewers who perform well on survey tasks that require diligence are

more successful in gaining cooperation. Regarding survey agency-related controls, we find that having contact with interviewers on a daily basis increases the chances of gaining respondents' cooperation. This result may highlight the importance of communication between survey agency coordinators and interviewers, but may also point to other factors at the survey agency level that affect respondent cooperation (such as the relative importance the survey agency attaches to SHARE).

The structure of the paper is as follows. Section 2 reviews the literature and Section 3 presents the features of the available data with a special focus on paradata and the outcome variable. Section 4 presents the empirical strategy, Section 5 comments on the empirical results and Section 6 concludes.

2. Previous findings

Panel studies are affected by attrition of subjects, which can bias parameter estimates due to potential differences between those who stay in the panel and those who drop out. It is by now standard in the literature to conduct exploratory analyses to understand how to prevent unit nonresponse during fieldwork. Literature on the determinants of survey participation recently proposed the use of paradata to gain a better understanding of response behaviour (e.g. Kreuter, 2013; Kreuter, Couper, Lyberg, 2010). However, even though paradata represent a rich source of new information, little attention has been paid for instance to indicators such as keystroke data (Couper and Kreuter, 2013) as well as additional information at higher levels, for example at the country or survey agency level.

High levels heterogeneity might be explained by differences in survey characteristics, in population characteristics or in data collection practices (Blom, 2012).¹ Most studies using cross-national data refrain from investigating the country level due to a small number of countries or due to the unavailability of harmonized information at this level. An exception is Lipps and Benson (2005), who analysed contact strategies in the first wave of SHARE using a multilevel model also taking into account the country level, but did not find significant between-country differences. However, the response process in later waves of a panel might differ from the response process in the baseline wave due to survey agencies accrued organizational experience or respondents' self-selection into later waves (Lepkowski and Couper, 2002). An advantage of using the fourth wave of SHARE, as we do, is that we can exploit additional, harmonized information collected at the survey agency level to understand better whether different fieldwork practices can explain heterogeneity in panel cooperation at the survey agency level, given a common survey topic. In SHARE, the countries and survey agencies mostly overlap; however, since in two countries (Belgium and France) more than one survey agency collected the data, we will use the term survey agency, instead of country, for the third (highest) level.

¹This was highlighted by Blom (2012) by examining country differences in contact rates in the European Social Survey (ESS) – a survey similar to SHARE in its attempt to achieve ex-ante harmonization across several European countries, but different from SHARE since it lacks the longitudinal dimension. By conducting counterfactual analysis, the author attributed the differences in contact rates to differential survey characteristics (mostly related to interviewers' contact strategies), population characteristics and coefficients. Like Blom (2012), we investigate the drivers of variability at the country level, but we are interested in panel cooperation – rather than contact – and use multilevel analysis as our empirical strategy.

Taking the role of the interviewer into account is vital for attrition analyses in face-to-face surveys. In the literature, results regarding interviewer continuity across waves are mixed. For example, Hill and Willis (2001) find a positive strong significant association between response rate and interviewer continuity, Lynn, Kaminska and Goldstein (2014) find that continuity positively affects cooperation in some situations, whereas other studies (Campanelli and O'Muircheartaigh, 1999, Nicoletti and Peracchi, 2005, Pickery, Loosveldt and Carton, 2001) find insignificant effects. These findings have been questioned as not only respondents attrit, but interviewers might attrit non-randomly from surveys as well (Campanelli and O'Muircheartaigh 2002). In the multi-country setting of SHARE, the selection and assignment of interviewers is subject to supervision by the survey agencies. While survey guidelines recommend interviewer continuity, we are not able to link interviewers across waves. Based on Vassallo, Durrant, Smith and Goldstein (2015), we decided to focus on the current (wave 4) interviewer.²

The literature has highlighted that isolating interviewer effects from area effects might be problematic when there is no fully interpenetrated design, i.e. random assignment of sample units to interviewers (Campanelli and O'Muircheartaigh, 1999; Durrant, Groves, Staetsky and Steele, 2010; Vassallo, Durrant, Smith and Goldstein, 2015). The lack of interpenetration is likely in face-to-face surveys, such as SHARE, in which the interviewer generally operates in limited geographical areas. Therefore, if there are geographical patterns in cooperation, these could appear as interviewer effects. It should be noticed that Vassallo, Durrant, Smith and Goldstein (2015) did not find significant area effects, after controlling for interviewer and household level effects in a cross-classified model, which is in line with findings by Campanelli and O'Muircheartaigh (1999) and Durrant, Groves, Staetsky and Steele (2010). Given the lack of interpenetrated assignment in SHARE, following standard practice, we include among our controls some area indicators (living in an urban or rural area) to capture area effects. Unfortunately, more detailed information about the area where respondents live is not available in waves 3 and 4.³

In our analysis, we consider interviewer attributes such as age, gender and experience with the survey that were collected by the agencies, and interviewer work quality indicators (interviewer average number of contacts and rounding indicators) that we compute.⁴ In fact, interviewer socio-demographic characteristics and experience (overall or within a specific survey) are typically included when explaining interviewer-level variance (West and Blom, 2017). The literature has also documented that interviewers with higher contact rates achieve higher cooperation rates (O'Muircheartaigh and Campanelli, 1999; Pickery and Loosveldt, 2002; Blom, de Leeuw and Hox, 2011; Durrant and D'Arrigo, 2014). Based on the literature on 'satisficing behaviour' in surveys (Krosnick, 1991), the underlying hypothesis concerning interviewer effects is that those who are diligent in specific tasks during the interview are more engaged and more successful in gaining cooperation than interviewers who show less diligent interviewing behaviour. Diligent interviewers are those who fulfil their task

²While Pickery, Loosveldt and Carton (2001) stated that the previous interviewer is more relevant, a more recent study (Vassallo, Durrant, Smith and Goldstein, 2015) shows that taking into account both previous and current wave interviewer within a multiple-membership model does not improve on the simpler two-level model that controls only for the current wave interviewer random effect.

³Additional area characteristics have been collected in wave 5 for all respondents, in wave 6 only for the refreshment sample.

⁴We construct indicators of rounding behaviour in measurements following Korbacher and Schröder (2013).

thoroughly to optimize the quality of their interviews, whereas less diligent interviewers use ‘satisficing strategies’, such as skipping introductions or rounding measurements, to minimize effort.

Lugtig (2014) highlighted four mechanisms of attrition at the respondent level, namely shocks (e.g. moving, health decline), habit (consistent participation pattern), absence of commitment and panel fatigue. Paradata can especially help in capturing commitment and panel fatigue to single out respondents at risk of future attrition due to non-cooperation. This can be based on interviewer assessments, for example, willingness to answer or whether the respondent asked for clarification, or directly derived from the interview data, for example item nonresponse. The latter in particular is a good predictor of participation in later waves. According to the theory of a *latent cooperation continuum* (Burton, Laurie and Moon, 1999), in fact, item nonresponse – not providing valid answers to some questions – is a precursor of unit nonresponse – not providing any answers – in the following wave. This theory finds empirical support in Loosveldt, Pickery and Billiet (2002).

The interview length also contributes to shaping the past interview experience. In longitudinal surveys, interview length in an earlier wave might affect the decision to participate in later waves. On the one hand, a longer interview can be seen as a burden and affect cooperation negatively; on the other hand, the length might also measure the respondent’s motivation and commitment to the survey and therefore can have a positive influence on cooperation. Findings in the literature concerning impacts of interview length on panel attrition in interviewer-administered settings are mixed, with some showing a positive association (Fricker et al, 2012; Hill and Willis 2001) with cooperation and some not finding any effect (Lynn, 2013; Sharp and Frankel, 1983). Branden, Gritz and Pergamit (1995) disentangle the wave-specific influence of interview length by taking the longitudinal perspective into account. They found that long interviews are positively correlated with cooperation during the first waves of a panel, but the association vanishes in later waves.

3. Data

3.1 SHARE and sample selection

SHARE is a multidisciplinary harmonized European survey, targeting individuals aged over 50 and their partners, and represents a principal source of data to describe and investigate the causes and consequences of the aging process for the European population (see Börsch-Supan et al, 2013). SHARE was conducted for the first time in 2004/2005 (wave 1) in 11 European countries (Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden and Switzerland) and Israel. In the second wave Poland, the Czech Republic and Ireland joined SHARE and additional refreshment samples were added to ensure representativeness of the targeted population. Wave 3, called SHARELIFE, conducted between 2008 and 2009, differed from the standard waves, since it collected the life histories of individuals who participated in wave 1 or wave 2. The fourth wave of SHARE, which started in 2011, is a regular wave (see Malter and Börsch-Supan, 2013).

The regular wave main questionnaire is composed of about twenty modules, each focusing on a specific topic, for example demographics, mental and physical health, cognitive

functions, employment and pensions. The questionnaire of SHARELIFE differed from the standard waves, since it had very few questions on the current condition⁵ but focused on gathering information regarding the life histories of individuals who participated in wave 1 or wave 2 (Schröder, 2011). We exploit mainly the third and the fourth wave of SHARE by investigating cooperation in wave 4 given participation in SHARELIFE and given contact in W4. The two waves are not completely comparable given the rather special content of the third wave, but the choice was driven mainly by the availability of paradata. The particular sample definition that we refer to implies that we have to be cautious when extending our results.

Both standard and retrospective SHARE interviews were conducted via face-to-face, computer-assisted personal interviews (CAPI). Not every eligible household member was asked to answer every module of the standard CAPI questionnaire: selected household members served as family, financial or household respondents. These individuals answered questions about children and social support, financial issues or household features on behalf of the couple or the household, respectively. This means that the length of the questionnaire varied among respondents by design, which has to be taken into account when analysing participation. An advantage of using SHARELIFE is that the differences among the types of respondents are limited since there is a distinction only between first and second respondent on the basis of very few questions on the household's current economic situation (e.g. household income). In all SHARE waves there is also the possibility to conduct a shorter proxy interview for cognitively impaired respondents. A proxy can answer on behalf of the eligible individual for most of the modules.

We describe our sample definition more precisely in Table 1. The number of individuals interviewed in SHARELIFE is 20,106;⁶ we then deleted 144 cases that were not part of the assigned, longitudinal sample for fieldwork wave 4, for example due to legal restrictions or changes in eligibility. We do not consider individuals from the longitudinal sample whose households were not contacted in wave 4 (522 cases) given that our focus is on cooperation, and excluded individuals who died between waves (599 cases). When linking the different data sources, interviewer information was not linkable for 5.3% of the total sample (994 cases). The proportion of non-linked observations exceeds 10% in Austria and Sweden, but some unresolvable cases remained in all countries.⁷ Furthermore, we do not have complete information on interviewers in wave 4 for 15 cases; wave 3 missing data concern 887 individuals, distributed among all the countries.⁸

⁵The variables related to the current condition are household income, health status, economic status and current income from employment, self-employment and pensions.

⁶We do not consider Greece and Ireland as these countries did not participate in wave 4. We also excluded France as interviewer information was unavailable and Poland due to lack of survey agency practices information.

⁷The sample of non-linked observations presents higher proportions of singles and women – and a higher average (but identical median) respondent age. Given the high prevalence of such observations in Austria and Sweden, we checked that dropping either country from the estimation sample does not affect parameter estimates in a significant way.

⁸Missing information is especially related to questions of the IV module regarding the area and type of building. In the module interviewers have to answer a few questions about the interview situations without the respondent.

3.2 Collection and preparation of paradata in SHARE

The collection of paradata is greatly facilitated by computer-assisted sample management tools and interview instruments. In the following section we describe the data sources in SHARE and the preparation of the variables that we derive from them.

For sample management SHARE uses a tailor-made sample management system (SMS). This program is installed on each interviewer's laptop and enables the interviewers to manage their assigned subsample. The success of a cross-national study such as SHARE depends heavily on the way in which the data are collected in the various countries. Therefore, using a harmonized tool for collecting interview data as well as contact data is crucial to ensure the comparability of the results. The SMS tool enables interviewers to register every contact with a household or individual respondent and enter result codes for every contact attempt (e.g. no contact, contact – try again, or refusal). These data were also used by Lipps and Benson (2005) to analyse contact strategies in the first wave of SHARE. Among the information collected through the SMS tool, we use the average number of contacts that interviewers registered before obtaining household cooperation or the final refusal. Furthermore, the sample definition is partly constructed based on contact information (see Table 1).

While the interview is conducted, additional paradata are collected by tracking keystroke data. Here, every time a key is pressed on the keyboard of the laptop, this is registered and stored by the software in a text file. From these text files, time stamps at the item level can be computed. Additionally, the keystrokes record the number of times an item was accessed, back-ups, whether a remark was made and the remark itself. We compute the interview length of wave 3 based on those files. In contrast to commonly used time stamps at the beginning and the end of the whole interview, this approach provides a precise and adequate length measure that is net of longer interruptions of the interview. To control for the potential effect of the interview length on cooperation propensity, we include it and its square term to account for possible non-linear effects as well. Controlling for the interview length helps to take into account the fact that SHARE interviews vary by design due to the complex structure of the questionnaire. Additionally, we use keystroke information to construct a variable for interviewer quality that is used in the robustness section. We first compute the median reading time, by interviewer, for section introductions that are relatively long, such as social network, activities, financial transfers and income from work and pension. If this value is lower than the country (and language) 25th percentile in at least one case, then we define a 'short introduction' dummy variable. This variable should capture interviewers who are likely to skip section introductions.

Furthermore, as paradata at the respondent level, we include information derived from the CAPI interviews in wave 3, in particular the percentage of item nonresponse to monetary items. The questions considered to construct this variable are: household income (HH017); value of the property (AC019); first monthly wage for employed individuals (RE021) or first monthly work income (RE023) for self-employed individuals; current wage if the respondent is still in employment (RE027), current income if the respondent is still self-employed (RE029); the pension benefit (RE036), wage at the end of the main job if retired (RE041) and income at the end of the main job if retired and worked as self-employed

(RE043). Such questions on monetary values can be both sensitive and difficult (Loosveldt, Pickery and Billiet, 2002; Moore, Stinson and Welniak, 2000). The respondent might perceive them as burdensome or uncomfortable to answer. Previous empirical research showed that the item nonresponse to income questions can predict participation (Loosveldt, Pickery and Billiet, 2002; Nicoletti and Peracchi, 2005). The public release of SHARE also contains a section in which interviewers are asked to evaluate the reluctance of respondents (IV module). Related to this, we include a dummy variable indicating whether the interviewer reported a high level of willingness to answer and whether she asked for clarification. Furthermore, information on the area (urban vs. rural) is derived from the IV module.

Additionally, interviewer information and survey agency fieldwork strategies were gathered and delivered by the survey agencies for wave 4. The interviewer information includes demographics (year of birth, education, gender) and previous experience in conducting SHARE interviews (a dummy that takes value one if the interviewer has already participated in at least one previous wave of SHARE). Interviewers' education level is not available for all countries. For those survey agencies that provided this information, we apply the International Standard Classification of Education (ISCED-97) to harmonize the country-specific answers.⁹

Among interviewer controls, we also add a measure of *work quality*, following Korbmacher and Schröder (2013). We try to capture interviewers' quality based on the grip strength test that SHARE proposes in every wave. The test consists of measuring respondents' grip strength twice for each hand using a dynamometer. In the CAPI, interviewers are explicitly told to record a value between 0 and 100, without rounding numbers to multiples of 5 and 10. 'Previous waves showed that multiples of 5 and 10 were recorded more than statistically expected' (Korbmacher and Schröder, 2013); in Appendix A (Figure A1) we report the wave 4 pattern of grip strength measurement. If interviewers have percentages of multiples of 5 and 10 that lie outside the 90% confidence interval centred on the statistically expected value of 20.8%, then the interviewer is not measuring grip strength properly. We identify interviewers who round too often, by defining a dummy that takes value one if the percentage exceeds the upper bound of the confidence interval and zero otherwise. We also generate another dummy variable for those interviewers who do not report enough multiples of 5 and 10 (the percentage falls short of the lower bound), as they may be strategically concealing inaccurate measurements.

Additional information is gathered at the survey agency level about fieldwork strategies. Topics covered are: recruitment, training, contacting respondents, translation, technical support, interview content, sampling process, management of interviewers and duration of fieldwork.¹⁰ Those data are collected mostly by means of open-ended questions, but some questions have a dropdown list. Open questions are difficult to handle within a multi-country framework. For this reason we focus on questions with standard answering options that show some variability. We consider especially the following questions: 'Who decides which

⁹We exploit this information to run robustness analysis with the subsample of agencies that provided the education information.

¹⁰Unfortunately information on interviewers' pay is not available in wave 4 of SHARE.

project is prioritized, assuming that interviewers work on several projects simultaneously?’ with ‘interviewer, agency or both’ as possible answers and ‘How often are you in contact with your interviewers about the SHARE study?’ with the following answering options: ‘less than once a month, once a month, several times a month, once a week, several times a week or every day’. We define two variables: *priority_agency*, that takes value one if the survey agency decides the priority of projects (four out of eleven survey agencies do) and *daily_contact*, which equals one if the survey agency has contact with the interviewers on a daily basis¹¹ (two out of eleven survey agencies do). An overview table of all the variables used for the analysis with descriptive statistics can be found in Appendix A (Tables A1–A3).

3.3 Attrition and cooperation in SHARE wave 4

After describing the features of SHARE and the paradata used, we present in greater detail the response behaviour patterns in wave 4 for those who participated in SHARELIFE¹², the sample in which we are interested.

The standard distinction in the survey participation process is in terms of location, contact and cooperation (Lepkowski and Couper, 2002):

- a. location of the sample unit means finding geographically eligible individuals at a given address,
- b. contact means reaching an eligible sample unit by telephone or face-to-face visits and
- c. cooperation is the completion of the interview.

Given that step a) is usually less problematic in a panel (Lepkowski and Couper, 2002) and we cannot test it, the final response rate will be the product of the contact and cooperation rates, at least in simplified terms.

Kneip (2013) reports household contact rates for the panel sample of SHARE wave 4 that are consistently above 90% with an average of about 95% across all countries, while household cooperation, which varies between about 60% and about 90%, shows greater variation across countries. Hence, the retention rates, which combine contact and cooperation, vary between 56% and about 90%.¹³

This highlights that establishing contact was not an issue in the panel sample for most countries and non-contact seems to be a very limited phenomenon compared with other surveys, such as the European Community Household Panel (ECHP), for which Nicoletti and Peracchi (2005) analysed participation, modelling contact and cooperation as sequential events. In our case, the very limited number of individuals in non-contacted households (2.6%) leads us to ignore the contact phase and focus exclusively on cooperation, instead.

¹¹We are not able to differentiate the direction of the communication between agency and interviewer – whether agencies check on interviewers frequently or whether interviewers contact the agency on a regular basis (with questions or for reporting) cannot be distinguished.

¹²Our sample of analysis differs slightly from the panel sample because we do not consider those interviewed in wave 1 or 2 who did not participate in wave 3 (SHARELIFE).

¹³All the rates calculated by Kneip (2013) are constructed according to AAPOR standards.

Figure 1 presents the percentage of contacted individuals who cooperate in wave 4. The figure highlights some heterogeneity among survey agencies with rather high cooperation rates (85% or more) in Switzerland and Italy and lower cooperation rates, below 80%, in Czech Republic, Germany and Sweden.¹⁴

4. Empirical strategy

We estimate a multilevel logit model¹⁵ to investigate correlates of subject cooperation while accounting for correlations in probabilities between respondents. This estimation strategy specifies the hierarchical structure of the data and allows us to avoid underestimation of standard errors and therefore incorrect inference (Couper and Kreuter, 2013, Goldstein, 2011). Given that we are interested in understanding how different levels contribute to explaining cooperation, we start by estimating a random-intercept model (null model). We then enrich this baseline model specification by stepwise inclusion of covariates at the individual, interviewer and survey agency level, respectively. This bottom-up procedure has the advantage of keeping the model simple (Hox, 2010). Our outcome of interest is cooperation, denoted with y_{ijk} , which takes the value one if respondent i interviewed by interviewer j of survey agency k participates in wave 4 conditional on having participated in wave 3.

The null model can be specified as follows:

$$\text{logit}\{p_{ijk}|\beta_0, u_{jk}, v_k\} = \beta_0 + u_{jk} + v_k \quad (4.1)$$

and the values of y , conditional on the random components, are independent draws from a Bernoulli random variable with probabilities p_{ijk} , i.e. $y_{ijk}|u_{jk}, v_k \sim \text{Bernoulli}(p_{ijk})$.

In Equation 4.1 the two random terms, u_{jk} and v_k , are interviewer-specific and survey agency-specific random effects, with $u_{jk} \sim N(0; \sigma_u^2)$ and $v_k \sim N(0; \sigma_v^2)$ respectively (Skrondal and Rabe-Hesketh, 2004). In a logit model the error variance at the first level, L^2_{ϵ} , is fixed at $\pi^2/3$, in order to fix the scale (Rabe-Hesketh and Skrondal, 2005). Thus, in the multilevel extension no level 1 variance will be estimated. We then compare Models 1–4, in which the covariates on the three different levels are introduced in a stepwise procedure to the null model in order to understand the role of each group of variables in reducing heterogeneity at different levels.

The first model specification (Model 1) includes a set of controls for individual-level socio-demographic characteristics, x_{ijk} . Among these variables we include SHARELIFE information on demographics, such as age, gender, years of education (and its square), marital status, employment status, health status (including a control for proxy interview),

¹⁴These numbers are our own calculations based on our sample restrictions. For the official rates, please refer to Kneip (2013).

¹⁵We estimate the multilevel logit model with the command of Stata *melogit* with mode-curvature adaptive Gauss–Hermite quadrature integration methods. The estimation results are stable, increasing the number of integration points.

controls for household income (dummy variables for the top three equivalent household income quartiles), a dummy taking value one if the respondent lives in a detached or semi-detached house to control for the type of residential building, and a binary indicator for living in an urban or rural area to capture area effects. Although additional area-related controls would be desirable in the absence of interpenetrated assignment, further information about the area where respondents live is available only from wave 5 onwards.

In the second model specification (Model 2) we add a set of paradata indicators (z_{ijk}) at the individual level. We include a dummy variable controlling for interrupted participation in previous waves, in particular whether the individual was interviewed only in wave 1 (but not in wave 2). To account for the influence of previous interview duration, the wave 3 interview length in hours and its square are added. At this stage we also include the percentage of item nonresponse to monetary questions, the willingness to answer and whether the respondent asked for clarification.

In the third model specification (Model 3) we include controls at the interviewer level (s_{jk}), specifically interviewer age and gender, interviewer experience and the average number of contacts per household registered by the interviewer (before the interview or the final refusal). We also include interviewer quality indicators: a dummy that identifies the interviewers who round least and another dummy for the interviewers who round most on grip strength measurement. Finally, Model 4 controls for a survey agency-level covariate (t_k), indicating daily communication between the interviewers and the survey agency.

The complete model, Model 4, is specified as follows:

$$\text{logit}\{p_{ijk} | \beta, u_{jk}, v_k\} = \beta_0 + \beta'_1 x_{ijk} + \beta'_2 z_{ijk} + \beta'_3 s_{jk} + \beta_4 t_k + u_{jk} + v_k \quad (4.2)$$

where x_{ijk} and z_{ijk} vectors are individual-level socio-demographic and paradata controls, s_{jk} is a vector of interviewers' covariates and t_k is a survey agency control.

As already pointed out, in the logistic model the variance of the lowest-level residuals is fixed at a constant. The main consequence is that in each of the models the underlying scale is standardized to the same standard distribution, meaning that the residual variance cannot decrease when adding controls to the model. Moreover, the value of the regression coefficients associated with the included controls and the value of the higher-level variances are rescaled. As a consequence, it is not possible to compare the null model parameters with the following enriched model specifications or to investigate how variance components change. Hox (2010) extended the rescaling procedure of Fielding (2004) to the multilevel setting and suggested the construction of scaling factors to be applied to parameters of the fixed part and random effects to make the changes in these variables directly interpretable. In the case of a multilevel logistic regression model, the scale correction factor is given by $\sqrt{\sigma_0^2/\sigma_m^2}$ for the parameters of the fixed part and by σ_0^2/σ_m^2 for variance components. The numerator is the total variance of the null model ($\sigma_0^2 = \sigma_e^2 + \sigma_u^2 + \sigma_v^2$) and the denominator is

the total variance of Model m ($m=1, \dots, 4$) including the first-level predictor variables, $\sigma_m^2 = \sigma_F^2 + \sigma_e^2 + \sigma_u^2 + \sigma_v^2 = \sigma_F^2 + \sigma_0^2$, with σ_F^2 variance of the linear predictor of Model m obtained using the coefficients of the predictors of the fixed part of the equation.

One important issue when dealing with multilevel models is to assess the accuracy of model parameters estimates, which is influenced both by the number of observations within groups and by the number of groups. Given our model formulation, the former is not a relevant issue at the third level but it could be at the second level: for some interviewers the number of interviews is particularly low. We address this issue in the robustness section by restricting our analysis only to interviewers with at least six interviews. Regarding the number of groups, the second level has a sufficiently high number of interviewers to ensure accuracy of parameter estimates. However, we might have inaccurate results due to the low number of survey agencies (our third level units). We address this problem with a simulation study to understand the finite sample behaviour of estimates from a three-level logit model when the hierarchical structure of the data is similar to the structure of our sample of analysis. Results and discussion are presented in Appendix B.

5. Results

5.1 Predictors from multilevel analysis

We report in Table 2 the estimated coefficients for the stepwise model specifications, in which we add respondent, interviewer and survey agency controls. The effects for each set of variables are described in the following subsection. We comment mainly on our preferred model specification, that is, the complete model specification reported in Model 4.

As in Durrant and Steele (2009), we comment on our results while referring to some socio-psychological concepts and theories that have been proposed in the literature, bearing in mind that there is an imperfect match between theoretical constructs and variables used.

Table 2 shows that the respondent characteristics are highly predictive of cooperation in wave 4. Both gender and age influence cooperation in wave 4. According to our estimates, age has a non-linear effect on the probability of cooperation. Both regressors, *Age* and *Age2*, are statistically significant: up to about 68 years of age there is a positive association after which it becomes negative - this is controlling for health conditions. Previous research found lower rates of participation among the elderly and interpreted this result as support of the social isolation theory (Krause, 1993). Individuals might decide to underuse their social support network because they are embarrassed or stigmatized or they may reject aid from others because they feel uncomfortable when assistance is provided. Isolation might translate also into lack of survey participation behaviour and may explain the negative age effect we find for older respondents.

If the respondent reported *being in poor health* in wave 3, this has a negative and statistically significant effect on the probability of cooperation. This is not surprising but at the same time inconvenient for a survey on health and aging. In case of very bad health conditions, SHARE allows proxy interviews: the indicator *anyproxy* highlights a negative association

with cooperation suggesting again that health is an important determinant of attrition. We will investigate later whether the health effect changes with interviewer attributes.

The literature finds that single-person households are less likely to cooperate and explain this result referring to the social isolation theory (Goyder, 1987; Groves and Couper, 1998). According to this theory alienation or isolation from the society are predictors of nonresponse. We find the opposite in our analysis of retention: compared with couples, singles who already cooperated in past waves are more likely to participate in the next wave.

In the survey research literature, according to the theory of social exchange (Goyder, 1987; Groves, Cialdini and Couper, 1992), socio-economic status has a non-linear effect on cooperation: low and high SES groups are less likely to cooperate than average. We include four indicators of socio-economic status: years of education (and its square), household income quartile dummies, living in a (semi-)detached house as proxy for wealth, and employment status. Education might be positively correlated with retention as those with higher education might appreciate the value of research more (Groves and Couper, 1998). Years of education is statistically significant and has a non-linear effect on retention. Income quartiles are significant as well. Compared to individuals having high household income (fourth quartile), wave 3 respondents with lower household income are more likely to participate in wave 4. Also in this case we find a non-linear effect (the second quartile dummy has the largest estimated coefficient). Living in a detached or semi-detached house increases the chances of cooperation in wave 4.¹⁶ This is in line with previous research that found lower cooperation among people living in flats (Goyder, 1987; Groves and Cooper, 1998) and may suggest the presence of a wealth effect on retention. Socio-economic conditions seem to be relevant for cooperation in a later panel wave.

Compared with individuals in a non-working condition (retired, unemployed, sick or disabled and homemakers), workers do not have a statistically different probability of cooperating in the next wave. It seems that work-related time constraints do not matter once individuals have enrolled in the panel¹⁷, differently from what has been found by Durrant and Steele (2009) using cross-sectional data. The time constraints theory considers the fact that a rather long and detailed questionnaire - that has the advantage of collecting a rich set of information - requires quite some time to answer all the questions. This might create problems when respondents are still in employment and has to be kept in mind when examining statistics such as employment rates later in life, for which survey participation or even attrition could be an issue. Other factors, such as the characteristics of the area where the respondent lives, might play a role in predicting (continued) cooperation. Living in an urban area in our case is not significant.

In addition to this standard set of respondent characteristics, we use respondent-level paradata. Compared with continuous participation, individuals with interrupted response patterns are less likely to participate again. As interrupted participation might signal a

¹⁶As missing information is especially related to questions of the IV module regarding the area and type of building, we run our analysis including those observations by adding binary indicators for missing information. The results, available upon request, do not change.

¹⁷We obtained similar results when including additional non-working condition dummies (retired, unemployed and disabled).

subgroup of respondents who are difficult to retain, we report in Section 5.3 how the effect of such indicator changes when interacted with interviewer attributes (such as experience with SHARE fieldwork). We can also observe that a very good or good level of willingness to answer and not having asked for clarification during the interview in wave 3 are highly significant predictors of higher probability of cooperation in wave 4. As already explained earlier, we show that the percentage of missing information in monetary amount questions is a significant predictor of cooperation failure in wave 4. This result is consistent with the theory of a *latent cooperation continuum* (Burton, Laurie and Moon, 1999).

As paradata at the respondent level, we also use the length of the whole interview in wave 3. Both the length of the interview in hours and its square are highly statistically significant, showing a inverse-u-shaped effect; therefore, the interview length has a positive association with cooperation up to a certain point, roughly 1.6 hours, when the probability of cooperating starts to decrease.¹⁸ This is in line with previous findings and supports the argument that longer interviews are – at least up to a certain point - a proxy for pleasant, talkative interviews instead of respondent burden. We should note that interview length measures the combined interviewer-respondent interaction and is therefore not exogenous to the interview process (Watson and Wooden 2009). To identify the causal impact of interview length one would probably require an experimental setting, which is out of scope for this paper.

To understand the variation at the interviewer level, we add some socio-demographic controls, age and gender, a variable indicating experience in previous SHARE waves, the average number of contacts per interviewer and two dummies capturing interviewers' quality based on grip strength rounding behaviour. Age and gender do not significantly affect cooperation in wave 4, whereas experience does play a role; more precisely, having experience with previous SHARE waves increases the likelihood of retaining respondents in the survey. Results concerning interviewer experience are consistent over different studies, leading to the conclusion that experience is positively associated with gaining cooperation (West and Blom, 2017). However, it is still unclear what drives the effect, that is, whether this is a selection effect (bad interviewers quit - Jäckle, Lynn, Sinibaldi and Tipping (2013)) or a learning effect (interviewers improve their skills in approaching resistance over time - Lemay and Durand (2002)). Durrant, Groves, Staetsky and Steele (2010) showed that experience in terms of the skill level acquired matters more than the time spent on the job.¹⁹ Our results are partly in line with the previous findings by Jäckle, Lynn, Sinibaldi and Tipping (2013) on the effect of experience, measured in years working for the survey agency. Regarding the average number of contacts, we see that an interviewer who on average registers many contacts is less likely to gain cooperation. A high average number of contacts can be an indicator of interviewer quality, i.e. such interviewers are less persuasive,

¹⁸Using interview pace rather than length does not change our results (estimates are available upon request). Pace is an alternative way of capturing the potential burden experienced by respondents in the previous wave. Here, we define pace as the ratio of length to the number of items asked and thus accounts for differences in instrument length by respondent type (for applications see Korbmacher and Schröder, 2013, and Loosveldt and Beullens, 2013). In the case of SHARELIFE the number of items asked is similar across respondent types, and this may explain why results do not substantially change when we replace length with pace.

¹⁹See Groves and Couper, 1998; Hox and de Leeuw, 2002; Jäckle, Lynn, Sinibaldi and Tipping, 2013; Lipps and Pollien, 2011.

or it can be seen as a measure of a workload complexity, as interviewers with difficult case loads end up trying more times.

It can also be noticed that the two variables measuring interviewer quality in terms of diligent interviewing behaviour are significant, with signs as predicted previously. If interviewers rounded grip strength scores more or less than average in wave 3, then gaining cooperation in wave 4 is less likely than in cases in which the rounding percentage is as expected. This finding is in accordance with Korbmacher and Schröder (2013) on consent to record linkage. While rounding too often is a clear indication of poor compliance to quality standards, rounding too little is probably due to interviewers strategically avoiding multiples of 5 to prevent being accused of cheating.

The final set of covariates in Table 2 is related to harmonized information collected at the survey agency level to gain knowledge on the correlation between survey agency strategies and cooperation. In this model specification we consider the variable *daily_contact* that captures the frequency of communication between the survey agencies and their interviewers. We find that having daily contact with interviewers increases the chances of obtaining the cooperation of respondents. This result hints at the importance of communication between survey agency coordinators and interviewers to conduct surveys successfully. We report in Table A4 in the Appendix A a model specification (column 3) in which both three-level variables (*priority_agency*²⁰ and *daily_contact*) are included among controls together with the two model specifications (column 1 and 2) in which the three-level predictors are instead included one at a time. *Priority_agency* is never significant. Degrees of freedom considerations lead us to be parsimonious in level-3 specification and therefore we decide not to include *priority_agency* in the main specification.²¹

5.2 Variance component analysis

Table 3 reports the results of various specifications of random-intercept models, without and with covariates, in terms of estimated variance components, intraclass correlations and model fit statistics.²² The definitions of level-2 (ICC_j) and level-3 (ICC_k) intraclass correlations in a three-level logit model are provided in Appendix B.

Looking at the intraclass correlations, we notice that survey agencies contribute about 1.6% of the variation, whereas interviewers account for about 25.9% (model 0 on Table 3). Based on the adjusted likelihood ratio test, we reject the null that the third-level variance component is zero. The test statistic takes value 10.90, and it is asymptotically distributed as a mixture of chi-squared with zero and chi-squared with one degree of freedom (Self and

²⁰We consider whether the priority of the projects is decided by the survey agency compared with situations in which interviewers can totally or partly choose how to organize their work. This can be seen as a variable capturing the extent to which interviewers are autonomous and free to choose among several projects on which they are currently working (e.g. working on SHARE or working on another survey on a specific day).

²¹Our simulations show that parsimony is a key issue to reduce bias in the estimation of level-3 variance, see Appendix B for additional details.

²²A similar approach can be found in a paper about interviewer effects on nonresponse in the European Social Survey (Blom, De Leeuw and Hox, 2011). Although the approach is similar, we refrain from comparing the findings across SHARE and ESS here. Nonresponse processes can differ substantially between cross-sectional cooperation and cooperation in a later wave of a panel.

Liang, 1987). The intraclass correlations in Table 3 suggest that most (72.5%, that is $(1 - ICC_j - ICC_k) \times 100$ in model 0) of the variation in cooperation is at the individual level.

Table 3 also reports the AIC values as a measure of goodness of fit for each successive model specification. Reductions in the AIC show improvements in the model fit. An examination of the log-likelihoods yields similar conclusions, whereby the full model is to be preferred. The likelihood ratio test shows that adding respondent-level paradata improves the model significantly and reduces the scaled variance at the respondent level by 3%.²³ If we compare Model 2 and Model 3, in which we introduce interviewer characteristics, it can be seen that this set of interviewer-level fixed effects accounts for a modest proportion of the variation at that level. Comparing the variance, σ_u^2 , between model 2 and model 3, we see that about 19% of the variation is captured by interviewer age, gender, experience, average number of contacts and rounding behaviour. The likelihood ratio test reveals that adding interviewer characteristics as predictors of cooperation results in a statistically significant improvement in model fit ($p < 0.0001$).

Finally, in Model 4 we add survey agency fieldwork strategies. The inclusion of survey agency-related variables captures a large part of the variation at the third level; comparing σ_v^2 between Model 3 and Model 4, we notice that we are able to explain about 90% of the variation. However, we need to take into account the fact that the variation at the survey agency level in total is rather small in comparison with the variance at the interviewer level. We recall that the survey agencies contribute about 1.6% of the variation, whereas interviewers account for about 24.9%. According to the likelihood ratio test, adding the survey agency characteristic as predictor of cooperation improves the model fit ($p < 0.0001$).

Our results should be interpreted cautiously because accuracy of higher level parameter estimates might be problematic in the context of multilevel models, particularly when the number of groups is small. In Appendix B we present simulation analyses along this line.

5.3 Cross-level interactions and robustness analysis

In this subsection we show that our results are robust to the inclusion of cross-level interactions and to various changes in model specification.

Considering cross-level interactions allows us to investigate non-cooperation for certain subgroups of respondents that are difficult to interview for several reasons – e.g. individuals in bad health, employed, living alone and with ‘unpleasant’ previous interview experience. We focus on how the effect of individual characteristics differs according to interviewer attributes.

As Groves and Couper (1998) suggest, interviewers with more experience are more able to gain cooperation in problematic situations (e.g. resistance). Therefore, we first investigate whether interviewer experience is able to mitigate the negative association of respondent bad health, marital status and previous interview indicators (item nonresponse and interrupted

²³The percentage change in the scaled variance at the respondent level is defined as $(3.125 - 3.226) / 3.125$, following the approach of Couper and Kreuter (2013). Percentage changes in the other scaled variance components are computed accordingly.

response patterns) with cooperation. We find statistically significant interaction effects only for interrupted response patterns. To clarify: in Table 2, Column 4, we see that experience has a 0.642 positive coefficient and interrupted response pattern has a negative coefficient of 0.977. In Column 1 of Table A5 we see that experience has a 0.716 positive coefficient, interrupted response pattern a 0.601 negative coefficient and their interaction a 0.675 negative coefficient. Thus, experience *per se* is predictive of retention, but experienced interviewers are less likely to gain cooperation when the respondent has an interrupted participation history than inexperienced interviewers. A possible explanation is that experienced interviewers put more effort where they expect higher rewards - and do not work as hard at regaining cooperation where they know respondents are harder to keep in the sample.

Although the gender of the interviewer is generally not significant in explaining cooperation (West and Blom, 2017), we investigate if for at least some respondents it plays a role. We include in the model cross-level interactions between interviewer gender (female) and respondent characteristics (such as bad health, marital status and previous interview indicators) to see whether being interviewed by a female changes the propensity to participate. We find statistically significant effects for marital status: the positive correlation between being single and cooperation in the baseline model specification (Table 2, Model 4) seems to be mainly driven by singles interviewed by female interviewers (see Column 2 of Table A5).

Finally, based on the evidence that sociodemographic similarities between respondent and interviewer increase the propensity to cooperate (West and Blom, 2017), we test whether matching based on age and gender affects cooperation. We find that the nearness of age between interviewer and respondent, measured as the distance between interviewer and respondent age, has insignificant effect on cooperation. We find a similar, insignificant result for gender concordance.

We further run robustness analyses by redefining the estimation sample, the list of covariates and the number of levels considered.

We redefine our estimation sample along three dimensions. First, we look at the effect of carrying out the analysis at the household, rather than the individual, level (Column 3, Table A5). Although Durrant and Steele (2009) highlight that cooperation is a complex social phenomenon that is explained by individual, rather than household characteristics, we show that household-level estimates are in line with individual-level ones. Next, in Column 4 of Table A5, we drop interviewers with fewer than six interviews. This second model specification addresses the potential inaccuracy of the estimates when the group sizes are small (Hox, 2010). Results do not change. We use this model specification to perform the goodness of fit test proposed by Perera, Sooriyarachchi and Wickramasuriya (2016), and fail to reject the null hypothesis that the specified model fits the data well.²⁴ Lastly, in Column 5

²⁴Perera, Sooriyarachchi and Wickramasuriya (2016) develop the goodness of fit test for a two level model. Therefore, in performing the test we treat our model specification as if it is a two-level model. The code to perform the test is available in the supplementary material.

of Table A5, we drop proxy interviews to check whether this rather particular subsample of SHARE respondents drives our baseline results, but find that this is not the case.

Our results are quite robust to the inclusion of further interviewer-level controls. For a subgroup of interviewers, we have information on education²⁵ and in Column 6 of Table A5 we show that adding this variable does not change our results. We do not report here estimation results for a model specification that includes a ‘short introduction’ variable (results are available upon request). This variable should capture interviewers who are likely to skip section introductions and is an additional quality indicator. To ensure harmonization, interviewers are instructed to read the whole CAPI question carefully. However, some interviewers do not follow this instruction: when we compare keystroke data about section introductions, we find that there are interviewers who read them quickly. This variable is insignificant and its inclusion leaves other parameter estimates unchanged.

The results regarding the effects of survey agency practices on the conditional mean of the dependent variable are also robust to the way we treat level three variability. Although the simulation exercise confirms the robustness of our baseline result regarding the positive effect of the third level control, we present in Table A5 Column 7 a two-level model with controls for groups of countries.²⁶ The third level variable, *daily_contact*, remains highly significant.

6. Conclusions

Panel cooperation has been a longstanding issue in survey research, with several studies seeking to identify the factors that affect subject attrition in panel surveys. Our analysis, based on observational data, focuses especially on the role of paradata in providing additional information to predict cooperation in a later wave of a panel. We are especially interested in the factors affecting cooperation propensity that are ‘under the researcher’s control’: survey agency fieldwork strategies, the features of interviewers and the respondent–interviewer interaction. We investigate which paradata from SHARE waves 3 and 4 help to predict cooperation in wave 4 regarding (1) the way the previous interview was conducted, (2) the characteristics of the wave 4 interviewer and (3) agency-level fieldwork indicators. Using multilevel models, we find that factors at all three levels (respondent, interviewer and survey agency) influence cooperation.

Panel respondents may base their cooperation decision on the way their previous interview was conducted. We find corroborating evidence for this: for instance, item nonresponse to monetary questions predicts cooperation in the next wave - respondents who answered most of the monetary items are more likely to participate in wave 4 than those who refused to answer a considerable number of questions. Interview length is another factor that is associated with cooperation in wave 4. We find that very long interviews are associated with lower participation in later waves. However, as long as the total interview length is less than

²⁵Interview education (*ISCED 5-6*) is a dummy that takes the value one if the interviewer has tertiary education.

²⁶We group countries as follows: the dummy *Southern* countries takes the value 1 for Italy and Spain and 0 otherwise and the dummy *Central* countries takes the value 1 for Belgium, Switzerland, Germany, the Czech Republic and Austria, whereas *Northern* countries, the reference group, equals 1 for Denmark, Sweden and the Netherlands.

1.6 hours, which holds for the vast majority of our cases, longer interviews are associated positively with future cooperation, possibly reflecting the respondent's interest in the survey or the quality of the interviewer-respondent interaction. This finding shows the difficulty with deriving implications for questionnaire development from interview length.

As far as interviewer characteristics are concerned, we find that previous experience with working as a SHARE interviewer matters more than socio-demographic characteristics, such as age, gender, or education. This is in line with the literature: interviewer gender and age have been generally found to be weak or insignificant determinants of cooperation, whereas experience does play a role, although the mechanisms behind it are still not well understood (West and Blom, 2017). Interviewers who perform well on survey tasks that require diligence are also more successful in gaining cooperation. This again reflects the importance of high-quality training and selecting diligent individuals as interviewers. Although the interviewer work-quality indicators we have are statistically significant, they account, together with socio-demographic characteristics, only for a modest percentage of the variance at the interviewer level. Important determinants, not considered here due to lack of information, are for instance interviewer continuity (Watson and Wooden, 2009, Lynn, Kaminska and Goldstein, 2014), socio-economic status, general attitudes, own behaviour, expectations and more comprehensive measures of job experience.

Finally, regarding survey agency-related controls, we find that having contact with interviewers on a daily basis increases the chances of gaining respondents' cooperation. This result may highlight the importance of communication between survey agency coordinators and interviewers to conduct surveys successfully, but may also point to other factors at the survey agency level that affect respondent cooperation (such as the relative importance the survey agency attaches to SHARE compared to other surveys they are managing at the same time). The limited number of survey agencies in our sample and the paucity of agency indicators prevent us from using more agency-level covariates and limits our ability to ascertain which explanation is the correct one. To investigate further the role of survey-agency controls one should probably use the most recent SHARE waves (7 and 8) that cover a much larger number of countries. Ideally, more detailed quantitative paradata at the agency level should also be collected.

We have also investigated cross-level interactions: the most interesting finding is that interviewer experience is generally predictive of retention, except when the respondent has an interrupted participation history. A possible explanation is that experienced interviewers put less effort when they expect lower chances of success. We also find significant interaction effects between interviewer gender (female) and respondent marital status (being single), and this may be used to devise a profitable assignment strategy.

Our analysis provides a description of response behaviour in SHARE for a specific, relatively early wave. Even in this setting, we have shown that an interrupted participation pattern makes retention less likely. The response process in later waves might depend on previous participation in more complex ways. To investigate this one should consider the whole longitudinal gross sample, that is, all the individuals who have been interviewed at least once, as this would allow the separation of retention and recovery. The underlying

mechanisms for subsequent participation on the one hand (retention) and interrupted participation on the other hand (recovery) might differ. We leave this to future research.

Acknowledgments

We are grateful for comments and suggestions made by participants at the conference of the European Survey Research Association, the Panel Survey Methods Workshop and the seminar of the Munich Center for the Economics of Aging, as well as by referees and the editor. We thankfully acknowledge discussions with Thorsten Kneip, Julie Korbmacher, Omar Paccagnella, and Annette Scherpenzeel. This paper uses data from SHARE wave 1, wave 2, wave 3 (SHARELIFE) and wave 4 release 6.0.0, as of March 31st 2017 (DOI: 10.6103/SHARE.w1.600; DOI: 10.6103/SHARE.w2.600; DOI: 10.6103/SHARE.w3.600; DOI: 10.6103/SHARE.w4.600). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N° 211909, SHARE-LEAP, N° 227822 and SHARE M4, N° 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions)

Appendix A. Additional Tables and Figures

Figure A1.
Frequency of grip strength values

Table A1

Descriptive statistics of the variables at the respondent level (N=16945)

Variable	Mean	SD	Min.	Max.	Description
Cooperation	0.84	0.37	0	1	Cooperation in w4 (outcome)
Female	0.56	0.50	0	1	Gender (reference: male)
Age	66.75	9.60	34	100	Age of respondent in years
Being in poor health	0.38	0.48	0	1	Self-reported poor health
Any proxy	0.06	0.24	0	1	A proxy helped answering the questionnaire
Single	0.23	0.42	0	1	Marital status
Years of education	10.74	4.47	0	25	
HH income – 1st quartile	0.32	0.47	0	1	HH income, 1st quartile by country
HH income – 2nd quartile	0.24	0.42	0	1	HH income, 2nd quartile by country
HH income – 3rd quartile	0.27	0.44	0	1	HH income, 3rd quartile by country
Working	0.30	0.46	0	1	If R declares to be employed or self-employed
Living in an urban area	0.23	0.42	0	1	Small town or rural area (ref.: urban)
Living in a (semi-) detached house	0.70	0.46	0	1	Living in a (semi-) detached house (ref.: flat)
Interrupted response pattern	0.06	0.24	0	1	Interviewed in w1 and w3 but not in w2
Item nonresponse to monetary questions	0.20	0.28	0	1	Proportion of item non-response to monetary items in w3
Length of interview	0.90	0.37	0.26	2.54	Length of interview in w3 (in hours)
Willingness to answer	0.93	0.26	0	1	Willingness to answer in w3

Variable	Mean	SD	Min.	Max.	Description
Did not ask for clarification	0.84	0.37	0	1	Did not ask for clarification in w3

Data: SHARELIFE release 6.0.0, SHARE wave 4 release 6.0.0 and SHARE paradata wave 3 and 4.

Table A2

Descriptive statistics of the variables at the interviewer level (N=643)

Variable	Mean	SD	Min.	Max.	Description
Interviewer's age	55.07	11.52	19	79	
Interviewer is female	0.63	0.48	0	1	
Interviewer's experience with working on previous SHARE waves	0.68	0.47	0	1	
Interviewer's average # of contacts in wave 4	2.41	0.73	0.20	7.11	Average # of contacts in wave 4 by interviewer
Rounding to a multiple of 5 for grip strength measure (too many)	0.35	0.48	0	1	If the interviewer's percentage of rounding is below (above) the lower (upper) cut-off of the 90% confidence interval centered around the statistically expected value of 20.8%
Rounding to a multiple of 5 for grip strength measure (too few)	0.03	0.16	0	1	
Short introductions	0.52	0.50	0	1	Interviewer has at least one short introduction (i.e. time recoded lower than a country-specific median)
Interviewer education (ISCED 5-6)	0.37	0.48	0	1	Interviewer has tertiary education (restricted sample)

Data: SHARE wave 4 release 6.0.0 and SHARE interviewer information wave 4.

Table A3

Descriptive statistics of the variables at the agency level (N=11)

Variable	Mean	SD	Min.	Max.	Description
Priority_by agency	0.36	0.48	0	1	Agency decides the priority of projects
daily_contact	0.18	0.38	0	1	Agency monitors and has contact with the interviewers on a daily basis

Data: SHARE agency information wave 4.

Table A4

Estimated multilevel models including alternative sets of agency characteristics (dependent variable: cooperation)

Model	1	2	3
	Only <i>daily_contact</i>	Only <i>prioritySA</i>	Both controls
<i>Respondent characteristics</i>			
Female	0.100 ** (0.049)	0.099 ** (0.049)	0.100 ** (0.049)
Age	0.171 *** (0.029)	0.171 *** (0.029)	0.171 *** (0.029)
Age ²	-0.001 *** (0.000)	-0.001 *** (0.000)	-0.001 *** (0.000)
Being in poor health	-0.194 *** (0.051)	-0.195 *** (0.051)	-0.192 *** (0.051)

Model	1	2	3
	Only <i>daily_contact</i>	Only <i>prioritySA</i>	Both controls
Single	0.211 *** (0.069)	0.211 *** (0.069)	0.210 *** (0.069)
Any proxy	-0.194 ** (0.097)	-0.195 ** (0.097)	-0.192 ** (0.097)
Years of education	0.045 ** (0.021)	0.041 * (0.021)	0.047 ** (0.021)
Years of education squared	-0.003 *** (0.001)	-0.002 *** (0.001)	-0.003 *** (0.001)
HH income – 1st quartile	0.234 *** (0.078)	0.236 *** (0.078)	0.233 *** (0.078)
HH income – 2nd quartile	0.376 *** (0.078)	0.378 *** (0.078)	0.375 *** (0.078)
HH income – 3rd quartile	0.153 ** (0.078)	0.157 ** (0.078)	0.150 * (0.078)
Living in a (semi-) detached house	0.301 *** (0.058)	0.297 *** (0.058)	0.303 *** (0.058)
Working	-0.066 (0.067)	-0.065 (0.067)	-0.065 (0.067)
Living in an urban area	0.042 (0.073)	0.048 (0.073)	0.044 (0.073)
<i>Paradata at the respondent level</i>			
Interrupted response pattern (<i>int. in wave 1 but not in wave2</i>)	-0.977 *** (0.083)	-0.972 *** (0.083)	-0.975 *** (0.083)
Item nonresponse to monetary questions	-0.527 *** (0.088)	-0.526 *** (0.089)	-0.527 *** (0.088)
Length of interview (hours)	1.147 *** (0.272)	1.152 *** (0.273)	1.151 *** (0.272)
Length of interview ² (hours)	-0.362 *** (0.114)	-0.362 *** (0.114)	-0.363 *** (0.114)
Willingness to answer	0.451 *** (0.090)	0.441 *** (0.090)	0.450 *** (0.090)
Did not ask for clarification	0.264 *** (0.068)	0.260 *** (0.068)	0.264 *** (0.068)
<i>Interviewer characteristics (w4)</i>			
Age	-0.004 (0.005)	-0.007 (0.005)	-0.004 (0.005)
Female	0.060 (0.104)	0.075 (0.105)	0.064 (0.104)
Experience (previous SHARE waves)	0.642 *** (0.109)	0.611 *** (0.114)	0.608 *** (0.113)
Interviewer-specific mean of contacts with HH until cooperation/refusal	-0.133 * (0.069)	-0.161 ** (0.068)	-0.128 * (0.070)
Rounding to a multiple of 5 for grip strength measure (too many)	-0.239 ** (0.105)	-0.219 ** (0.105)	-0.247 ** (0.105)
Rounding to a multiple of 5 for grip strength measure (too few)	-0.769 *** (0.228)	-0.785 *** (0.230)	-0.750 *** (0.229)
<i>Agency control variables</i>			
daily contact	0.714 *** (0.153)		0.689 *** (0.141)
Priority decided by Survey Agency		0.180 (0.211)	0.136 (0.116)

Model	1	2	3
	Only <i>daily_contact</i>	Only <i>prioritySA</i>	Both controls
Constant	-5.389*** (1.075)	-5.062*** (1.081)	-5.433*** (1.074)
σ_u^2 (interviewer level)	1.006	1.000	1.008
σ_v^2 (agency level)	0.007	0.081	0.001
N	16945	16945	16945

Standard errors in parentheses;

*

$p < 0.05$,

**

$p < 0.01$,

$p < 0.001$.

P-values for fixed effect covariates significance refer to Wald-type tests.

Table A5

Robustness analysis – Multilevel model estimates (dependent variable: cooperation)

Model	1	2	3	4	5	6	7
	Cross-level interactions			Additional robustness analysis			
	Response pattern × Interviewer experience	Single × Female Interviewer	Household level	Number of Interviews >5	No proxy interviews	Interviewer education	Two-level model grouped countries
<i>Respondent characteristics</i>							
Female	0.102** (0.049)	0.098** (0.049)	0.167*** (0.059)	0.108** (0.049)	0.120** (0.051)	0.069 (0.054)	0.100** (0.049)
Age	0.173*** (0.029)	0.171*** (0.029)	0.207*** (0.036)	0.172*** (0.029)	0.166*** (0.031)	0.171*** (0.033)	0.171*** (0.029)
Age ²	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)
Being in poor health	-0.195*** (0.051)	-0.194*** (0.051)	-0.184*** (0.060)	-0.200*** (0.051)	-0.181*** (0.053)	-0.182*** (0.056)	-0.195*** (0.051)
Single	0.206*** (0.069)	0.066 (0.098)	0.341*** (0.079)	0.200*** (0.070)	0.229*** (0.072)	0.211*** (0.077)	0.212*** (0.069)
Proxy	-0.192** (0.097)	-0.193** (0.097)	-0.355*** (0.123)	-0.183* (0.098)		-0.157 (0.104)	-0.194** (0.097)
Years of education	0.045** (0.021)	0.045** (0.021)	0.004 (0.026)	0.046** (0.021)	0.042* (0.022)	0.046** (0.023)	0.047** (0.021)
Yredu2	-0.003*** (0.001)	-0.003*** (0.001)	-0.000 (0.001)	-0.003*** (0.001)	-0.002*** (0.001)	-0.003*** (0.001)	-0.003*** (0.001)
HH income – 1st quartile	0.236*** (0.078)	0.235*** (0.078)	0.114 (0.096)	0.238*** (0.078)	0.213*** (0.081)	0.268*** (0.085)	0.235*** (0.078)
HH income – 2nd quartile	0.384*** (0.078)	0.378*** (0.078)	0.268*** (0.084)	0.393*** (0.078)	0.380*** (0.080)	0.400*** (0.084)	0.376*** (0.078)
HH income – 3rd quartile	0.154** (0.078)	0.154** (0.078)	0.130 (0.084)	0.162** (0.078)	0.153* (0.080)	0.195** (0.086)	0.154** (0.078)
Living in a detached or semi-detached house	0.299*** (0.058)	0.301*** (0.058)	0.274*** (0.067)	0.313*** (0.058)	0.303*** (0.060)	0.327*** (0.063)	0.303*** (0.058)
Working	-0.065 (0.067)	-0.065 (0.067)	0.001 (0.082)	-0.050 (0.068)	-0.055 (0.069)	-0.146** (0.075)	-0.063 (0.067)
Living in an urban area	0.046 (0.073)	0.040 (0.073)	0.024 (0.082)	0.033 (0.074)	0.051 (0.076)	0.068 (0.081)	0.041 (0.073)
<i>Paradata at the respondent level</i>							

Model	1	2	3	4	5	6	7
	Cross-level interactions			Additional robustness analysis			
	Response pattern × Interviewer experience	Single × Female Interviewer	Household level	Number of Interviews >5	No proxy interviews	Interviewer education	Two-level model grouped countries
Interrupted response pattern (<i>Int. in w1 but not in w2</i>)	-0.601 *** (0.124)	-0.977 *** (0.083)	-0.924 *** (0.090)	-0.974 *** (0.083)	-0.946 *** (0.087)	-0.979 *** (0.089)	-0.978 *** (0.083)
Item nonresponse to monetary questions	-0.523 *** (0.088)	-0.529 *** (0.088)	-0.570 *** (0.107)	-0.537 *** (0.089)	-0.498 *** (0.094)	-0.546 *** (0.097)	-0.526 *** (0.088)
Length of interview (hours)	1.122 *** (0.272)	1.152 *** (0.272)	0.651 *** (0.154)	1.088 *** (0.275)	1.179 *** (0.286)	1.267 *** (0.297)	1.164 *** (0.274)
Length of interview ² (hours)	-0.350 *** (0.114)	-0.364 *** (0.114)	-0.110 *** (0.039)	-0.336 *** (0.115)	-0.356 *** (0.120)	-0.435 *** (0.125)	-0.367 *** (0.114)
Willingness to answer	0.451 *** (0.090)	0.452 *** (0.090)	0.383 *** (0.111)	0.457 *** (0.091)	0.506 *** (0.097)	0.370 *** (0.096)	0.454 *** (0.090)
Did not ask for clarification	0.264 *** (0.068)	0.262 *** (0.068)	0.285 *** (0.080)	0.260 *** (0.069)	0.279 *** (0.072)	0.302 *** (0.074)	0.265 *** (0.068)
Interviewers' characteristics (w4)							
Age	-0.004 (0.005)	-0.004 (0.005)	-0.002 (0.004)	-0.004 (0.005)	-0.004 (0.005)	-0.005 (0.005)	-0.003 (0.005)
Female	0.055 (0.104)	0.005 (0.107)	0.100 (0.100)	0.048 (0.107)	0.052 (0.105)	0.050 (0.120)	0.056 (0.104)
Interviewer education (ISCED 5-6)						0.037 (0.127)	
Experience with working on previous SHARE waves	0.716 *** (0.110)	0.644 *** (0.109)	0.688 *** (0.104)	0.690 *** (0.112)	0.623 *** (0.111)	0.663 *** (0.125)	0.634 *** (0.109)
Interviewer-specific mean of contacts with HH until cooperation/refusal	-0.133 * (0.069)	-0.134 * (0.069)	-0.101 (0.063)	-0.142 * (0.074)	-0.130 * (0.071)	-0.145 * (0.075)	-0.123 * (0.065)
Rounding to a multiple of 5 for grip strength measure (too many)	-0.241 ** (0.105)	-0.237 ** (0.105)	-0.255 ** (0.101)	-0.277 *** (0.107)	-0.239 ** (0.107)	-0.199 * (0.119)	-0.249 ** (0.105)
Rounding to a multiple of 5 for grip strength measure (too few)	-0.785 *** (0.228)	-0.772 *** (0.228)	-0.763 *** (0.224)	-0.901 *** (0.259)	-0.809 *** (0.232)	-0.779 *** (0.248)	-0.763 *** (0.228)
Interactions							
Single × Female Iwer		0.236 ** (0.115)					
Interrupted response pattern × Iwer Experience	-0.675 *** (0.165)						
Agency control variables							

Model	1	2	3	4	5	6	7
	Cross-level interactions			Additional robustness analysis			
	Response pattern × Interviewer experience	Single × Female Interviewer	Household level	Number of Interviews >5	No proxy interviews	Interviewer education	Two-level model grouped countries
daily contact	0.712 *** (0.151)	0.711 *** (0.152)	0.664 *** (0.134)	0.707 *** (0.149)	0.695 *** (0.154)	0.591 *** (0.216)	0.665 *** (0.150)
Southern countries							0.116 (0.172)
Central countries							0.053 (0.122)
Constant	-5.477 *** (1.076)	-5.344 *** (1.076)	-6.787 *** (1.312)	-5.484 *** (1.082)	-5.333 *** (1.131)	-5.343 *** (1.221)	-5.543 *** (1.085)
σ_u^2 (interviewer level)	1.006	1.005	0.763	1.003	1.012	1.072	1.011
σ_v^2 (agency level)	0.006	0.007	<0.001	0.004	0.007	0.012	
N	16945	16945	11890	16713	15913	13574	16945

Standard errors in parentheses;

* $p < 0.05$,

** $p < 0.01$,

*** $p < 0.001$.

P-values for fixed effect covariates significance refer to Wald-type tests. Household level model specification: The interview length is defined as the sum of the single interview lengths; participation to the previous waves is defined at the household level.

Appendix B. Simulation study

Multilevel model estimation is generally based on a maximum likelihood approach and standard errors are derived under the assumption of asymptotic Normal distribution of the estimator. There are several simulations studies assessing finite sample performance of multilevel models when the outcome is continuous (see Maas and Hox, 2005, for a recent review), but fewer analyses exist for discrete response multilevel models. Moreover, these results are mainly for two-level binary models (see Paccagnella, 2011, for a literature review), with the exception of a recent study by Kim, Choi and Emery (2013).

The main conclusions of the simulation analyses for binary multilevel models are that parameter estimates are downward biased whenever there are few observations per group (Rodriguez and Goldman, 1995), and when the number of groups is small, in particular when considering higher level covariates and variance components (Bryan and Jenkins, 2016; Paccagnella, 2011).²⁷ Results for standard error bias exhibit the same pattern. Paccagnella (2011) investigates the accuracy of model estimates in the case of a two-level logit model and concludes that the bias in the fixed part of the model is negligible even with 10 clusters, but the number of clusters should increase significantly to ensure accuracy of the variance components estimate. Moreover, his simulation results show that the bias in the variance estimate is higher when the second level ICC is lower. Kim, Choi and Emery (2013) focus on the comparison of estimation performance in both two- and three-level models when using different methods and statistical packages, but do not investigate the role of group size and number on estimation accuracy.²⁸

If the two-level logit model results extend to a three-level framework, two features of our model specification are likely to imply inaccurate estimates: on the one hand, the small number of level-3 groups, i.e. the number of survey agencies, on the other hand the small ICC at the third level. Given the lack of simulation results on binary response three-level models, we study the finite sample properties of a multilevel logit model with a simulation exercise in which the hierarchical structure of the generated datasets replicates the one of our survey dataset.

Following Goldstein and Rasbash (1996), we specify our baseline model as follows:

$$\begin{aligned} \text{logit}\{p_{ijk} | (Z_{ijk}; \beta, u_{jk}, v_k)\} &= \beta_0 + \beta_1 X_{1ijk} + \beta_2 D_{1ijk} + \beta_3 X_{2jk} + \beta_4 D_{2jk} + \beta_5 D_{3k} + u_{jk} + v_k \\ Y_{ijk} | u_{jk}, v_k &\sim \text{Bernoulli}(p_{ijk}) \end{aligned} \quad (5.1)$$

²⁷Less than 30 groups lead to unacceptable downward biases in the parameter estimates of a two-level logit model according to Bryan and Jenkins (2016). Similar results are obtained in Paccagnella (2011).

²⁸Simulation results for the three-level specification are based on datasets in which there are 50 level-1 units, 10 level-2 units and 30 level-3 units throughout.

where the controls Z_{ijk} are continuous (X_*) and binary (D_*) variables respectively, and the random effects are independent and normally distributed, $u_{jk} \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$. We consider two other model specifications: a null model and a model without level-3 controls.

The baseline model specification presented in Equation 5.1 replicates the full model specification estimated in Column 4 of Table 2. However for simplicity we include only two controls at level-1 and level-2, a continuous (X_{1ijk}, X_{2jk}) and a binary (D_{1ijk}, D_{2jk}) control respectively, and only one binary control at the third level, having the same distribution of *daily_contact* binary variable in our model.

According to Davis and Scott (1995), the intraclass correlations at level-2 and level-3 in a multilevel logit model are defined as:

$$ICC_j = \frac{\sigma_u^2}{\sigma_e^2 + \sigma_u^2 + \sigma_v^2} \quad ICC_k = \frac{\sigma_v^2}{\sigma_e^2 + \sigma_u^2 + \sigma_v^2}$$

where $\sigma_e^2 = \pi^2/3$. Varying level-2 and level-3 variances in ranges consistent with those estimated in Tables 2 and 3, we exploit finite sample behaviour of estimates for a range of values of the intraclass correlations: $ICC_j = [0.19, 0.26]$ and $ICC_k = [0.01, 0.035]$.

The true value of the parameters is reported in Table B1. Parameters are kept constant across model specifications apart from the level-2 and level-3 variances.

Table B1

True parameters' values used in the simulation analysis

Parameter	True value
β_0	1.00
β_1	0.8
β_2	-0.3
β_3	-0.7
β_4	0.4
β_5	-0.2

In particular, we investigate the finite sample behaviour of estimates when the number of groups (N_k) and the variance of the random effect at the third level (σ_v^2) are small. In the simulations, the following conditions vary:

- N_k assumes values in {5, 10, 15, 20, 25};

- σ_v^2 assumes values in {0.05, 0.1, 0.15} and the level-2 variance σ_u^2 takes values in {0.8, 1, 1.2}, in
- line with model estimates in Table 3.

To replicate the variability in the group size observed in the data, we allow for heterogeneity in the number of observations within each of the three levels. More precisely, the number of level-2 units within each level-3 group can take 5 values, $S_{jk} = \{30, 45, 60, 75, 90\}$, and this reproduces the variability in the number of interviewers per survey agency in the data. The number of level-1 units within each level-2 group can take 5 values, $S_{ijk} = \{10, 20, 30, 40, 70\}$, which replicates the distribution of the number of respondents per interviewers in the data.²⁹

Following Paccagnella (2011), for each combinations of the level-2 and level-3 variances we generate 1,000 simulated datasets (R). To generate the covariates we simulate from five standard independent Normal distributions. The binary variables at level-1 and 2 take value one if the underlying continuous variable is positive and zero otherwise. The binary variables at level-3 is obtained from the underlying standard Normal by imposing that the mean of the binary variable is 0.17, as for *daily_contact*.

The random components u_{jk} and v_k are obtained with R random draws from two independent Normal distributions with mean zero and variances σ_u^2 and σ_v^2 respectively.

Using the regression coefficients of Table B1, the generated regressors and the random components, we compute $\pi_{ijk} = \text{logit}(p_{ijk})$ and derive p_{ijk} by applying the inverse logit function. Finally, each value of the dependent variable Y_{ijk} is a random draw from a Bernoulli distribution with probability p_{ijk} .

To perform our simulation exercise we use Stata 15. The multilevel logit models are estimated with the *melogit* command. The integration method used to integrate the approximated likelihood over the random effects is the mode-curvature adaptive Gauss–Hermite quadrature with 7 integration points.

To gain knowledge on the accuracy of the estimates of model parameters and their standard errors, we report three summary measures: relative parameter bias, non-coverage rate and relative standard error bias (Paccagnella, 2011; Bryan and Jenkins, 2016; Vassallo, Durrant and Smith, 2017). The relative parameter bias is computed as the percentage difference between estimated and true parameters. The non-coverage rate (Mass and Hox, 2005) is used to assess the accuracy of the standard errors. It results from the average over model replications of a binary indicator that takes value one if the true parameter value lies outside the 95% estimated confidence interval. The estimates are accurate if the relative parameter bias is close to zero and the non-coverage rate is close to 5%. Given that non-coverage rate might reflect both parameter bias and standard error bias, following Bryan and Jenkins

²⁹These sets of five values are replicated according to the number of level-3 and level-2 groups.

(2016) and Rodriguez and Goldman (1995), we compute also standard error bias comparing the ‘analytical’ standard error - average of estimated standard errors over the replications - and the ‘empirical’ standard error - standard deviation of the estimated parameters based on the R replications (Greene, 2004).

Table B2

Results from baseline model simulations when the level two variance σ_u^2 is set to 1 and the level three variance σ_v^2 to 0.05.

Parameter \ N_k	5	10	15	20	25
Relative parameter bias (%)					
β_0	-0.39	0.30	-0.37	-0.13	0.64
β_1	0.26	-0.01	-0.11	-0.03	0.01
β_2	0.33	-0.27	0.09	0.10	0.04
β_3	-0.28	0.16	-0.13	0.26	0.14
β_4	0.11	-0.31	0.27	-0.08	-1.30
β_5	-0.26	0.00	-2.36	-4.50	2.21
σ_u^2	-0.75	-0.58	-0.56	-0.30	-0.14
σ_v^2	-44.29	-29.31	-17.29	-12.81	-10.70
Non-coverage rate					
β_0	0.13 *	0.10 *	0.10 *	0.07 *	0.08 *
β_1	0.05	0.06	0.05	0.05	0.04
β_2	0.04	0.05	0.06	0.05	0.06
β_3	0.05	0.05	0.04	0.05	0.04
β_4	0.05	0.06	0.05	0.05	0.04
β_5	0.21 *	0.12 *	0.08 *	0.09 *	0.07 *
σ_u^2	0.09 *	0.05	0.06	0.05	0.06
σ_v^2	0.43 *	0.29 *	0.21 *	0.17 *	0.15 *

Parameter \ N_k	5	10	15	20	25
Relative standard error bias (%)					
β_0	-17.60	-11.28	-10.10	-5.90	-2.19
β_1	-0.85	-2.99	-0.51	2.02	3.34
β_2	4.15	0.32	-1.73	0.22	-3.11
β_3	1.84	-1.07	0.14	-2.10	1.34
β_4	2.76	-3.93	1.57	.56	-0.63
β_5	-13.38	-13.92	-10.42	-12.06	-6.85
σ_u^2	-5.65	0.80	0.69	1.60	-2.12
σ_v^2	-17.30	-14.25	-8.50	-6.66	-6.47

Note:

* Significantly different from 0.05 at 5% significance level.

In Table B2 we report simulations results for the case in which $\sigma_u^2 = 1$ and $\sigma_v^2 = 0.05$, the closest scenario to our full model specification (Column 4 of Table 2). Focusing on the scenario with 10 groups at the third level, the relative bias is zero for most parameters with the exception of the level-3 variance, σ_v^2 , which is 29.31% downward biased. The non-coverage rate is significantly different from 0.05 for both β_5 (the coefficient of the level-3 dummy control) and σ_v^2 . This is the result of parameter bias and standard error bias: according to the third panel of Table B2 the standard errors of β_5 is underestimated by 13.92% and the standard error of σ_v^2 is underestimated by 14.25%.

Generally, both parameter and standard error biases decrease as the number of level-3 groups increases (with the sole exception of the relative parameter bias of β_5), but they remain far from the target value even with 25 groups. In the case of σ_v^2 the non-coverage rate is as high as 0.15 even with 25 groups.

By varying σ_u^2 and σ_v^2 , and thus the *ICC*, the downward bias of σ_v^2 ranges from 20% to 29% and it is lower when the *ICC* at the third level is higher. Results of these further simulations are available upon request.

Given the simulation results, in our application in Table 2 we are likely to underestimate the third level variance of the full model by about 29%.

The simulation results reveal that the distribution of the estimated level-3 control (β_5) shows large variability for all values of N_k . It is worth stressing, however, that the coefficient of *daily_contact* would remain statistically significant in our application even if its relative bias was equal to the 10th or the 90th percentile of the relative bias distribution, and accounting for the 14% underestimation of the standard error.

In addition to the baseline model specification in Equation 5.1, we replicate the simulation exercise (by varying the level-2 and 3 variances as in the baseline scenario) for two alternative model specifications: a null model (as in Model 0 of Table 2) and a model without the binary level-3 control (as in Model 3 of Table 2). The rationale is to understand whether estimation accuracy changes with the number and level of controls included and to provide some evidence on how we should expect the parameter bias to change, varying the model specification as in Table 2.

Focusing again on the model with 10 level-3 groups we report in Table B3 the simulation results for the null model with $\sigma_u^2 = 1.2$ and $\sigma_v^2 = 0.15$ (values close to those estimated in the first column of Table 2). Results are very similar when the model without level-3 control is considered instead. The downward bias in the estimation of the level-3 variance is reduced by about 50% and the same is true for the standard error bias. In particular, the negative parameter bias of σ_v^2 is between 11% and 18% and the standard errors bias is between 6% and 9% in the case of 10 level-3 units when we let the *ICC* vary within the specified range. This provides a rule of thumb to measure the downward bias in higher-level variances for different model specifications as reported in Table 3.

This result is somehow intuitive if we think that we need ‘large’ sample size to ensure consistency and efficiency in regression model parameter estimates. This extends also to level-3 parameters: we need a large number of groups, i.e. more information to exploit, to estimate additional level-3 effects reliably (Bryan and Jenkins, 2016).

Table B3

Results from null model simulations when the level-2 variance σ_u^2 is set to 1.2 and the level-3 variance σ_v^2 to 0.15.

Parameter \ N_k	5	10	15	20	25
Relative parameter bias (%)					
β_0	-0.69	.51	-0.40	0.37	0.58
σ_u^2	0.13	.18	-0.08	-0.05	0.14
σ_v^2	-24.31	-13.86	-6.27	-4.56	-4.06

Parameter \ N_k	5	10	15	20	25
Non-coverage rate					
β_0	0.14 *	0.10 *	0.09 *	0.08 *	0.07
σ_u^2	0.06	0.05	0.06	0.05	0.06
σ_v^2	0.31 *	0.22 *	0.14 *	0.13 *	0.11 *
Relative standard error bias (%)					
β_0	-14.93	-9.29	-8.60	-6.20	-1.29
σ_u^2	-3.66	-0.23	-1.33	2.79	-2.62
σ_v^2	-6.73	-7.64	-5.61	-2.41	-2.65

Note:

* Significantly different from 0.05 at 5% significance level.

We should point out that the 95% confidence interval that we use to derive the non-coverage rate is obtained from the inversion of the Wald test (as is normally done in Stata). Berkhof and Snijders (2001) show that the Wald test has low power in the context of variance component tests and should not be used to test for variance component significance. In fact, the Wald test relies on the assumption of asymptotic normality of the maximum likelihood estimator and this is problematic when the random effect variance is considered, in particular if its value is close to zero, as zero lies on the boundary of the parameter space (Mass and Hox, 2005).

Bottai (2003) examines the asymptotic behaviour of confidence intervals in the case in which information is zero at a critical point of the parameter space. He compares several ways to derive confidence intervals - inversion of the log-likelihood ratio test, of the Wald test and of the score test - and finds that the score test-based confidence intervals, that use expected information (instead of observed information), are the ones that perform better. As stressed in Bottai and Orsini (2004), the problem of inference about the variance of the random effect can be accommodated into this more general framework, because when the variance component is zero, the score function is identically zero, and information is zero.

Bottai and Orsini (2004) develop a Stata routine (xtvc) that allows testing the null that the random effect variance is equal to a specific value (including zero) and computes 'corrected' confidence intervals based on the inversion of the score test. This routine works for random-effects linear regression models and can be used after the xtreg command in Stata. The simulation results presented in the paper show that the observed rejection rate is close to the nominal 5% level, regardless of the number of groups considered. The confidence interval

obtained with the inversion of the score test is ‘slightly shifted to include greater values’ (Bottai and Orsini, 2004, p. 432) with respect to the Wald confidence interval.

In our simulations we use Wald-based confidence intervals³⁰, as is normally done in the literature - see for example the recent contribution by Vassallo, Durrant and Smith (2017) – even though they may be inaccurate as the variance at level three is set to relatively small values. Possible strategies to assess the level of inaccuracy would be extending the Bottai and Orsini’s routine to multilevel logit models, adopting a parametric bootstrapping strategy, or relying on alternative estimation procedures such as the Bayesian MCMC algorithm.

Generalization of the Bottai and Orsini routine to multilevel logit model requires working with a marginal likelihood (in which random effects are integrated out) that does not have in this case a closed form. This makes such a procedure computationally expensive. A parametric bootstrap method could be used to construct 95% confidence intervals for the variance components (Kuk, 1995 and Goldstein, 1996), but this would be even more computationally expensive.³¹

Alternatively, a Bayesian MCMC algorithm - with non-informative priors to ease comparability with maximum likelihood - could be used to perform the entire analysis. Such algorithm, that would also entail an extra computational burden to achieve convergence, would in fact directly provide confidence intervals for parameter estimates based on the posterior distributions. We know from Rodriguez and Goldman (2001) that in three-level logistic models parameter estimates using full maximum likelihood and Bayesian estimation are similar when random effects variances are large. To the best of our knowledge, a Bayesian MCMC procedure for the three-level logit model in the case where at least one variance is small has not been implemented in the literature, therefore we leave further investigation of this issue to future research.

To the extent that we can draw from the existing literature, we can expect the confidence intervals for random effects variances based on the Wald test to be smaller and shifted towards zero (see for example Turner, Omar and Thompson, 2001 and Browne and Draper, 2006).

References

- Berkhof J, Snijders T. 2001; Variance Component Testing in Multilevel Models. *Journal of Educational and Behavioral Statistics*. 26:133–152.
- Blom AG. 2012; Explaining cross-country differences in survey contact rates: Application of decomposition methods. *Journal of the Royal Statistical Society Series A*. 175(Part 1):217–242.
- Blom AG, de Leeuw ED, Hox JJ. 2011; Interviewer effects on nonresponse in the European Social Survey. *Journal of Official Statistics*. 27:359–377.
- Blom, AG, Lynn, P, Jäckle, A. Understanding Cross-National Differences in Unit Non-Response: The Role of Contact Data. 2008. ISER Working Paper 2008-01

³⁰Note that in our main model specification we test variance component significance using the adjusted likelihood ratio test (Section 5.2).

³¹For each iteration of the simulation process, a number of samples should be drawn from the model evaluated at current parameter estimates. The model should then be estimated for each sample and the confidence intervals constructed from the distribution of parameters estimates.

- Börsch-Supan A, Brandt M, Hunkler C, Kneip T, Korbmacher J, Malter F, Schaan B, Stuck S, Zuber S. 2013; Data resource profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*. 42:992–1001. [PubMed: 23778574]
- Bottai M. 2003; Confidence Regions When the Fisher Information Is Zero. *Biometrika*. 90:73–84.
- Bottai M, Orsini N. 2004; Confidence intervals for the variance component of random-effects linear models. *The Stata Journal*. 4(4):429–435.
- Branden, L, Gritz, RM, Pergamit, MR. The Effect of Interview Length on Attrition in the National Longitudinal Survey of Youth. 1995. NLS Discussion Paper. Report: NLS 95-28
- Browne WJ, Draper D. 2006; A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*. 1:473–514.
- Bryan ML, Jenkins SP. 2016; Multilevel modelling of country effects: a cautionary tale. *European Sociological Review*. 32(1):3–22.
- Burton, J; Laurie, H; Moon, N. Don't ask me nothin' about nothin', I just might tell you the truth – The interaction between unit nonresponse and item nonresponse. Paper presented at the International Conference on Survey Nonresponse; Portland, Oregon. 1999.
- Campanelli P, O'Muircheartaigh C. 1999; Interviewers, Interviewer Continuity, and Panel Survey Response. *Quality and Quantity*. 33:59–76.
- Campanelli P, O'Muircheartaigh C. 2002; The Importance of Experimental Control in Testing the Impact of Interviewer Continuity on Panel Survey Nonresponse. *Quality and Quantity*. 36:129–144.
- Couper MP, Kreuter F. 2013; Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society Series A*. 176:271–286.
- Davis P, Scott A. 1995; The effect of interviewer variance on domain comparisons. *Survey Methodology*. 21:99–106.
- Durrant GB, D'Arrigo J. 2014; Doorstep interactions and interviewer effects on the process leading to cooperation or refusal. *Sociological Methods Research*. 43:490–518.
- Durrant GB, Groves RM, Staetsky L, Steele F. 2010; Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*. 74:1–36.
- Durrant G, Kreuter F. 2013; The use of paradata in Social Survey Research. *Journal of the Royal Statistical Society Series A*. 176:1–3.
- Durrant G, Steele F. 2009; Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys. *Journal of the Royal Statistical Society Series A*. 172:361–381.
- Fielding A. 2004; Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality and quantity*. *European Journal of Methodology*. 38:425–433.
- Fricker, S; Creech, B; Davis, J; Gonzalez, J; Tan, L; To, N. Does length really matter? Exploring the effects of a shorter interview on data quality, nonresponse, and respondent burden. Paper presented at the Federal Committee on Statistical Methodology 2012 Research Conference; Washington DC, USA. 2012.
- Goldstein H. 1996; Consistent estimators for multilevel generalised linear models using an iterated bootstrap. *Multilevel Modelling Newsletter*. 8:3–6.
- Goldstein, H. *Multilevel Statistical Models*. 4th. Chichester: Wiley; 2011.
- Goldstein H, Rasbash J. 1996; Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, A*. 159:505–513.
- Goyder, J. *The Silent Minority Boulder: Nonrespondents on Sample Surveys*. Boulder: Westview Press; 1987.
- Greene W. 2004; The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econometrics Journal*. 7:98–119.
- Groves RM, Cialdini RB, Couper M. 1992; Understanding the decision to participate in a survey. *Public Opinion Quarterly*. 56:475–495.
- Groves, RM, Couper, MP. *Nonresponse in Household Interview Surveys*. New York: John Wiley and Sons, Inc; 1998.

- Hill DH, Willis RJ. 2001; Reducing Panel Attrition: A Search for Effective Policy Instruments. *The Journal of Human Resources*. 36:416–438.
- Hox, JJ. *Multilevel Analysis Techniques and Application*. 2nd. New York: Routledge; 2010.
- Hox, JJ, de Leeuw, E. The influence of interviewers' attitude and behavior on household survey nonresponse: An international comparison. In: Groves, RM, Dillman, DA, Eltinge, JL, Little, RJA, editors. *Survey Nonresponse*. New York: Wiley; 2002.
- Jäckle A, Lynn P, Sinibaldi J, Tipping S. 2013; The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys. *Survey Research Methods*. 7:1–15.
- Kim Y, Choi YK, Emery S. 2013; Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician*. 63:171–182.
- Kneip, T. *SHARE Wave 4: Innovations and Methodology*. München: MEA; 2013. Survey Participation in the Fourth Wave of SHARE, Malter, Frederic und Börsch-Supan, Axel; 140–155.
- Korbmacher JM, Schröder M. 2013; Consent when linking survey data with administrative records: The role of the interviewer. *Survey Research Methods*. 7:115–131.
- Krause N. 1993; Neighbourhood deterioration and social isolation in later life. *International Journal of Aging and Human Development*. 36:9–38. [PubMed: 8425749]
- Kreuter, F. *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, New Jersey: John Wiley and Sons; 2013.
- Kreuter, F, Couper, MP, Lyberg, LE. *Proc Jt Statist Meet American Statistical Association*. Alexandria: American Statistical Association; 2010. The use of paradata to monitor and manage survey data collection; 282–296.
- Krosnick JA. 1991; Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*. 5:213–236.
- Kuk AYC. 1995; Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society, Series B*. 57:395–407.
- Lemay M, Durand C. 2002; The effect of Interviewer Attitude on Survey Cooperation. *Bulletin De Methodologie Sociologique*. 76:27–44.
- Lepkowski, JM, Couper, MP. Nonresponse in the second wave of longitudinal household surveys. In: Groves, RM, Dillman, DA, Eltinge, JL, Little, RJA, editors. *Survey Nonresponse*. New York: Wiley; 2002.
- Lipps O, Benson G. 2005 Cross national contact strategies. Paper presented at the AAPOR – ASA Section on Survey Research Methods.
- Lipps O, Pollien A. 2011; Effects of interviewer experience on components of nonresponse in the European Social Survey. *Field Methods*. 23:156–172.
- Loosveldt G, Beullens K. 2013; The impact of respondents and interviewers on interview speed in face-to-face interviews. *Social Science Research*. 42:1422–1430. [PubMed: 24090842]
- Loosveldt G, Pickery J, Billiet J. 2002; Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics*. 18:545–557.
- Lugtig P. 2014; Panel attrition: Separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods and Research*. 14:699–723.
- Lynn, P. *Longer Interviews May Not Affect Subsequent Survey Participation Propensity*. 2013. Understanding Society Working Paper Series: 2013-07
- Lynn P, Kaminska O, Goldstein H. 2014; Panel Attrition: How Important Is It To Keep the Same Interviewer? *Journal of Official Statistics*. 30:434–457.
- Maas C, Hox J. 2004; Robustness issues in multilevel regression analysis. *Statistica Neerlandica*. 58:127–137.
- Malter, F, Börsch-Supan, A, editors. *SHARE Wave 4 Innovations & Methodology*. Munich: MEA, Max Planck Institute for Social Law and Social Policy; 2013.
- Moore J, Stinson L, Welniak E. 2000; Income measurement error in surveys: A review. *Journal of Official Statistics*. 16:331–361.

- Nicoletti C, Peracchi F. 2005; Survey response and survey characteristics: Microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 168:763–781.
- O’Muircheartaigh C, Campanelli P. 1999; A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 162:437–446.
- Paccagnella O. 2011; Sample size and accuracy of estimates in multilevel models. *New simulation results. Methodology*. 7(3):111–120.
- Perera AAPNM, Sooriyarachchi MR, Wickramasuriya SL. 2016; A goodness of fit test for the multilevel logistic model. *Communications in Statistics – Simulation and Computation*. 45:643–659.
- Pickery J, Loosveldt G. 2002; A multilevel multinomial analysis of interviewer effects on various components of unit nonresponse. *Quality and Quantity*. 36:427–437.
- Pickery J, Loosveldt G, Carton A. 2001; The effects of interviewer and respondent characteristics on response behavior in panel surveys: A multilevel approach. *Sociological Methods and Research*. 29:509–523.
- Rabe-Hesketh, S, Skrondal, A. *Multilevel and Longitudinal Modeling Using Stata*. 2nd. College Station, Texas: Stata Press; 2005.
- Rodriguez G, Goldman N. 1995; An assessment of estimations procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, A*. 158:73–89.
- Rodriguez G, Goldman N. 2001; An assessment of estimations procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*. 164:339–355.
- Self SG, Liang KY. 1987; Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*. 82:605–610.
- Sharp LM, Frankel J. 1983; Respondent burden: A test of some common assumptions. *Public Opinion Quarterly*. 47:36–53.
- Schröder, M. *SHARELIFE Methodology*. Mannheim: MEA; 2011. *Retrospective Data Collection in the Survey of Health, Ageing and Retirement in Europe*.
- Skrondal, A, Rabe-Hesketh, S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman and Hall/CRC; 2004.
- Turner RM, Omar RZ, Thompson SG. 2001; Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*. 20:453–472. [PubMed: 11180313]
- Vassallo R, Durrant GB, Smith PWF, Goldstein H. 2015; Interviewer effects on non-response propensity in longitudinal surveys: a multilevel modelling approach. *Journal of the Royal Statistical Society: Series A*. 178:83–99.
- Vassallo R, Durrant GB, Smith PWF. 2017; Separating interviewer and area effects by using a cross-classified multilevel logistic model: simulation findings and implications for survey designs. *Journal of the Royal Statistical Society: Series A*. 180:531–550.
- Watson, N, Wooden, M. Identifying factors affecting longitudinal survey response. In: Lynn, P, editor. *Methodology of Longitudinal Surveys*. Chichester, UK: John Wiley; 2009.
- West BT, Blom AG. 2017; Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*. 5:175–211.

Figure 1. Proportions of cooperation in w4 by survey agency

Note: 95% confidence intervals displayed. AT=Austria, BE-Fr=Belgium-Wallonia, BE-Fl=Belgium-Flanders, CH=Switzerland, CZ=Czech Republic, DE= Germany, DK=Denmark, ES=Spain, IT=Italy, NL=The Netherlands, SE=Sweden

Table 1

Sample definition

Number of observations released in SHARELIFE ^I	20,106
<i>Sample restrictions</i>	
Not part of assigned w4 sample	144
Household not contacted in w4	522
Deceased in w4	599
<i>Linkage restrictions</i>	
Non-linked with interviewer information	994
<i>Incomplete data restrictions</i>	
Missing data at interviewer level	15
Missing data at respondent level	887
Final number of respondents	16,945
Final number of interviewers	643
Final number of survey agencies	11

^I without Greece, Ireland, France and Poland

Table 2

Estimated multilevel models including respondent, interviewer and agency characteristics (dependent variable: cooperation)

Model	Intercept only	1	2	3	4
<i>Respondent characteristics</i>					
Female		0.088* (0.048)	0.099** (0.049)	0.099** (0.049)	0.100** (0.049)
Age		0.183*** (0.029)	0.170*** (0.029)	0.171*** (0.029)	0.171*** (0.029)
Age ²		-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)
Being in poor health		-0.218*** (0.050)	-0.195*** (0.051)	-0.195*** (0.051)	-0.194*** (0.051)
Single		0.182*** (0.068)	0.212*** (0.069)	0.212*** (0.069)	0.211*** (0.069)
Any proxy		-0.316*** (0.095)	-0.193** (0.097)	-0.196** (0.097)	-0.194** (0.097)
Years of education		0.047** (0.021)	0.039* (0.021)	0.040* (0.021)	0.045** (0.021)
Years of education squared		-0.002*** (0.001)	-0.002*** (0.001)	-0.002*** (0.001)	-0.003*** (0.001)
HH income – 1st quartile		0.164** (0.077)	0.236*** (0.078)	0.236*** (0.078)	0.234*** (0.078)
HH income – 2nd quartile		0.403*** (0.077)	0.383*** (0.078)	0.378*** (0.078)	0.376*** (0.078)
HH income – 3rd quartile		0.146* (0.077)	0.157** (0.078)	0.158** (0.078)	0.153** (0.078)
Living in a (semi-) detached house		0.283*** (0.057)	0.300*** (0.058)	0.296*** (0.058)	0.301*** (0.058)
Working		-0.105 (0.067)	-0.066 (0.067)	-0.065 (0.067)	-0.065 (0.067)
Living in an urban area		0.021 (0.073)	0.020 (0.074)	0.047 (0.073)	0.042 (0.073)
<i>Paradata at the respondent level</i>					
Interrupted response pattern(int. in wave 1 but not in wave2)		-0.991*** (0.083)	-0.991*** (0.083)	-0.973*** (0.083)	-0.977*** (0.083)

Model	Intercept only	1	2	3	4
Item nonresponse to monetary questions			-0.521 *** (0.089)	-0.527 *** (0.089)	-0.527 *** (0.088)
Length of interview (hours)			1.170 *** (0.274)	1.149 *** (0.273)	1.147 *** (0.272)
Length of interview ² (hours)			-0.368 *** (0.114)	-0.361 *** (0.114)	-0.362 *** (0.114)
Willingness to answer			0.444 *** (0.090)	0.441 *** (0.090)	0.451 *** (0.090)
Did not ask for clarification			0.257 *** (0.068)	0.261 *** (0.068)	0.264 *** (0.068)
Interviewer characteristics (w4)					
Age				-0.007 (0.005)	-0.004 (0.005)
Female				0.072 (0.105)	0.060 (0.104)
Experience (previous SHARE waves)				0.627 *** (0.113)	0.642 *** (0.109)
Interviewer-specific mean of contacts with HH until cooperation/refusal				-0.161 ** (0.068)	-0.134 * (0.069)
Rounding to a multiple of 5 for grip strength measure (too many)				-0.216 ** (0.105)	-0.238 ** (0.105)
Rounding to a multiple of 5 for grip strength measure (too few)				-0.788 *** (0.230)	-0.768 *** (0.228)
Agency control variables					
daily contact					0.714 *** (0.153)
Constant	1.872 *** (0.098)	-4.651 *** (1.023)	-5.445 *** (1.043)	-5.009 *** (1.079)	-5.388 *** (1.075)
σ_u^2 (interviewer level)	1.174	1.184	1.198	1.000	1.006
σ_v^2 (agency level)	0.073	0.093	0.086	0.089	0.007
N	16945	16945	16945	16945	16945

Standard errors in parentheses;

P-values for fixed effect covariates significance refer to Wald-type tests.

$p<0.001$.

$p<0.01$,

**

$p<0.05$,

*

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Estimated variance components, intraclass correlations and model fit statistics for different model specifications of the multilevel models of cooperation

Model	0	1	2	3	4
Intercept only	Respondent	R-paradata	Interviewer	Agency	
Variance components					
<i>Not scaled:</i>					
σ_e^2 (individual level)	3.29				
σ_u^2 (interviewer level)	1.174	1.184	1.198	1.000	1.006
σ_v^2 (agency level)	0.073	0.093	0.086	0.089	0.007
Scaled:					
σ_e^2 (individual level)	3.29	3.226	3.125	3.048	3.031
σ_u^2 (interviewer level)	1.174	1.161	1.138	0.926	0.927
σ_v^2 (agency level)	0.073	0.091	0.082	0.082	0.006
Intraclass correlation (scaled variances):					
ICC _j (interviewer level)	0.259	0.260	0.262	0.228	0.234
ICC _k (agency level)	0.016	0.020	0.019	0.020	0.002
Log likelihood	-6917.251	-6835.011	-6694.445	-6667.654	-6661.432
LR test against previous column model (degrees of freedom, p-value of LR test)		164.48 (14; 0.000)	281.13 (6; 0.000)	53.58 (6; 0.000)	12.44 (1; 0.000)
Model fit statistic AIC	13840.5	13704.02	13434.89	13393.31	13382.86

Observations: 16945 respondents, 643 interviewers, 11 agencies. ICC=intraclass correlation; AIC=Akaike information criterion.