Journal of Survey Statistics and Methodology (2023) ***00***, *1–26*

INTERVIEWER EFFECTS ON THE MEASUREMENT OF PHYSICAL PERFORMANCE IN A CROSS-NATIONAL BIOSOCIAL SURVEY

SOPHIA WALDMANN
JOSEPH W. SAKSHAUG*
ALEXANDRU CERNAT

Biosocial surveys increasingly use interviewers to collect objective physical health measures (or "biomeasures") in respondents' homes. While interviewers play an important role, their high involvement can lead to unintended interviewer effects on the collected measurements. Such interviewer effects add uncertainty to population estimates and have the potential to lead to erroneous inferences. This study examines interviewer effects on the measurement of physical performance in a cross-national and longitudinal setting using data from the Survey of Health, Ageing and Retirement in Europe. The analyzed biomeasures exhibited moderate-to-large interviewer effects on the measurements, which varied across biomeasure types and across countries. Our findings demonstrate the necessity to better understand the origin of interviewer-related measurement errors in biomeasure collection and account for these errors in statistical analyses of biomeasure data.

KEY WORDS: Biomeasure; Health survey; Interviewer effects; Intraclass correlation; Measurement error; Biosocial survey.

Sophia Waldmann is a Research Associate at the German Youth Institute (DJI), Nockherstraße 2, 81541 Munich, Germany. Joseph W. Sakshaug is Professor of Statistics at the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany, the Ludwig-Maximilian University of Munich, Germany, and the University of Mannheim, Germany. Alexandru Cernat is an Associate Professor of Social Statistics at the School of Social Sciences, University of Manchester, Humanities Bridgeford Street 2.13N, Manchester M13 9PL, UK.
We thank Lukas Olbrich for sharing his expertise in the implementation of the location-scale model in the context of interviewer effects and all those who commented on the presentation of this work at the 7th SHARE User Conference in Bled, Slovenia (October 2022). Data access was provided by SHARE-ERIC for scientific use and free of charge. Researchers may apply for access to these data at the SHARE Research Data Center (http://www.share-project.org/data-access.html). Code for replication is available upon request. The study design and analysis were not preregistered.
This article uses data from SHARE waves 1, 2, 4, 5, 6, 7, and 8 (DOIs: 10.6103/SHARE.w1. 710, 10.6103/SHARE.w2.710, 10.6103/SHARE.w4.710, 10.6103/SHARE.w5.710, 10.6103/SHARE.w6.710, 10.6103/SHARE.w7.711, 10.6103/SHARE.w8.100); see Börsch-Supan et al.

https://doi.org/10.1093/jssam/smad031

Statement of Significance

This study documents the presence of interviewer effects on the measurement of physical performance (or "biomeasures") collected in a cross-national setting. Our findings suggest non-negligible interviewer effects on biomeasures across countries, which has practical implications for statistical analysis of biomeasure data. Being the first comprehensive cross-country and cross-year comparison of interviewer effects on biomeasures, the presented results offer a basis for further study of interviewer effects on the quality of biomeasure data.

1. INTRODUCTION

The collection of physical health measures (or "biomeasures") in social surveys has become an established practice, now implemented in several large population-based biosocial surveys (Sakshaug et al. 2015), including Understanding Society (McFall et al. 2012), the English Longitudinal Study of Aging (Banks et al. 2014), the US Health and Retirement Study (Crimmins et al. 2015), and the US National Social Life, Health and Aging Project (Jaszczak et al. 2009). Examples of biomeasures collected in these surveys include anthropometric measures (e.g., height, weight, blood pressure), physical performance measures (e.g., grip strength, peak flow), and cardiovascular function (e.g., pulse rate), and biological specimens such as blood and saliva. The scientific relevance of collecting biomeasures in population-based biosocial surveys has become increasingly clear in the context of demographic change and growing social inequalities, as highlighted by Kumari and Benzeval (2021) who assert that: "Understanding the interaction between people's social and economic circumstances and their

(2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA No 211909, SHARE-LEAP: GA No 227822, SHARE M4: GA No 261982, DASISH: GA No 283646), and Horizon 2020 (SHARE-DEV3: GA No 676536, SHARE-COHESION: GA No 870628, SERISS: GA No 654221, SSHOC: GA No 823782, SHARE-COVID-19: GA No 101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, 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, OGHA_04-064, HHSN271201300071C, RAG052527A) and various national funding sources is gratefully acknowledged (see www.shareproject.org).

*Address correspondence to Joseph W. Sakshaug, Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany; E-mail: joe.sakshaug@iab.de.

health across the life span is essential to develop policies not only to improve the nation's health but also its social and economic capacities" (p. 26).

Social surveys that—in contrast to clinical studies—provide population representative data often rely on respondents' subjective self-reports to measure their diagnosed illnesses and physical health. Compared to self-reported health measures, biomeasures are taken to be more precise and objective. They allow for more comprehensive assessments of physical health and disease prevalence in the population, both of which can be analyzed in relation to social, economic, and environmental factors (Kumari and Benzeval 2021). Moreover, when the same biomeasures are collected in different countries, they enable cross-country comparisons of health and disease prevalence that inform global health policy discussions (Banks and Smith 2012; Franzese 2015; Barros et al. 2019; Vancampfort et al. 2019; Angel et al. 2022).

However, while providing research opportunities and benefits, biomeasures still have to be examined regarding the error sources typical for social survey data collection. With some self-administered exceptions (e.g., mail-in saliva samples, Dykema et al. 2017), biomeasures are mainly collected in face-to-face settings either by nurses or by interviewers (Sakshaug et al. 2015). These actors play a crucial role in the measurement process by explaining and demonstrating the measurement, administering the measurement (often in respondents' homes), and recording the results (Groves et al. 2009; Korbmacher 2014; Sakshaug et al. 2015). At the same time, their high involvement in the measurement process can lead to unintended interviewer/nurse effects on the measurement quality of the collected biomeasures (Cernat and Sakshaug 2020, 2021). Such interviewer/nurse effects can introduce additional uncertainty in population estimates and lead to false inferences if not accounted for in statistical analyses (Schnell and Kreuter 2005). Thus, it is important to study these effects which can inform improvements in measurement quality and harmonization of biomeasure collection within and across countries.

Against this background, the present study examines interviewer effects on the measurement of physical performance in the cross-national and longitudinal Survey of Health, Ageing and Retirement in Europe (SHARE), a leading data source for research on the interplay of social, economic, and health factors in Europe (SHARE-ERIC 2022). By its international character, the SHARE allows for comparisons of interviewer effects on biomeasurements across countries, which has not been the subject of previous research.

The following research questions (RQs) are addressed:

(1) What is the overall magnitude of interviewer effects on biomeasures collected in a cross-national and longitudinal biosocial survey?
(2) Does the magnitude of interviewer effects vary between different types of biomeasures, especially those that require more interviewer involvement?
(3) Are there differences in interviewer effects (i) between countries and (ii) across data collection waves?

2. THEORETICAL BACKGROUND

2.1 Literature Review

In contrast to biomeasures, measurement errors in self-reported health measures have been extensively studied (Boudreau et al. 2004; Raghunathan 2006; Gorber et al. 2007; Gil and Mora 2011; Davillas and Jones 2021; Footman 2021). Erroneous self-reporting has been attributed to social norms (Gil and Mora 2011; Davillas and Jones 2021), recall error (Boudreau et al. 2004), or undiagnosed conditions of which respondents are unaware (Leong et al. 2013; Petersen and Benzeval 2016). Social desirability bias in self-reports of body weight, sexual health, and substance use can be interviewer-related (Johnson and Parsons 1994; Heeb and Gmel 2001; Leone et al. 2021; Footman 2021), but less sensitive variables such as height also exhibit interviewer effects (Olbrich et al. 2022).

The potential pitfalls of self-reported health are the main motivation for the collection of objective physical health measures in population-based social surveys. According to Korbmacher (2014), there are two key benefits of collecting such biomeasures. First, they measure health status objectively without the risk of social desirability or recall error. Second, biomeasures can also reveal information about undiagnosed diseases such as diabetes. Moreover, the validation and calibration of self-reported survey data are sometimes based on biomeasures as a reference (Ezzati et al. 2006). However, to serve as a useful supplement to (or potential replacement of) self-reported health measures, biomeasures should be more accurate and less prone to measurement errors than their corresponding self-reports.

Most biomeasures are developed and validated in clinical settings, so the principal evaluation of their measurement properties and error sources is not necessarily a task for survey methodology. However, the adaption of classical clinical measures in large-scale social surveys might generate different error sources. In contrast to clinical studies, data collection in biosocial surveys is not always conducted by medical professionals, but also by trained "lay interviewers," and inside respondents' homes instead of a medical facility (Sakshaug et al. 2015; Guyer et al. 2017).

While much has been written about biomeasure participation (Gavrilova and Lindau 2009; Sakshaug et al. 2010; Dykema et al. 2017; Boyle et al. 2021; Pashazadeh et al. 2021), including interviewer or nurse effects on cooperation (Jaszczak et al. 2009; Korbmacher 2014; Sakshaug et al. 2015; Guyer et al. 2017; Cernat et al. 2021), very few studies have investigated interviewer or nurse effects on the measurement quality of biomeasures collected in biosocial surveys.

There is mixed evidence of observer-related measurement error in physical performance measures from clinical research. Typical measures such as timed walking, balancing, or sit-and-stand tests have yielded low measurement error variation across different observers (Cress et al. 1996; Durand et al. 2004;

Roberts et al. 2011; Stomfai et al. 2011; Bodilsen et al. 2015; Carsley et al. 2019). However, there are reports of observers having influence on measures of waist or hip circumference (Ulijaszek and Kerr 1999), blood pressure (Armstrong 2002; Bur et al. 2003; Dickson and Hajjar 2007), shoulder motion range (de Winter et al. 2004), and ratings of physical performance based on ladder climbing and trunk rotation (Durand et al. 2004).

In two investigations of large-scale biosocial surveys conducted in respondents' homes, considerable amounts of unexplained interviewer and nurse variation in biomeasurements were found (Cernat and Sakshaug 2020, 2021). The intraclass correlations (ICCs), a common measure of the estimable "interviewer effect" that characterizes interviewers' influence on a survey measurement, ranged from 0.03 to 0.30 for interviewers/nurses with larger effects occurring for the more complex biomeasures (e.g., touch test, timed balance and walk, grip strength, measures of lung capacity). ICCs above 0.10 are considered uncommon in the survey literature (Beullens and Loosveldt 2016) and give reason to assume that interviewers/nurses have varying influences on the measurements. These interviewer effects can have important implications for statistical analyses, comparable to the design effect caused by within-group homogeneity in cluster sampling (Schnell and Kreuter 2005). Within-interviewer correlations ρ_{int} increase the variance of population estimates by a factor of approximately $1 + \rho_{\text{int}}(m - 1)$, depending on the magnitude of ρ_{int} and the average interviewer workload m (Kish 1962). Thus, correlations within interviewer groups add unnecessary uncertainty to survey-based estimates. A possible consequence of multivariate analyses is incorrect inferences when the statistical model does not account for the given correlation structure (Schnell and Kreuter 2005).

2.2 Interviewer Influences on Physical Performance Measurements

There are multiple mechanisms by which interviewer effects can manifest in physical health measures and especially for physical performance measures. The required interaction and the fact that the measurements take place in the "unstandardized" homes of respondents make it difficult to control interviewers from varying their behavior (Cernat and Sakshaug 2021). Physical movements require a certain amount of space or include the furniture in a room, which can confront the interviewer with spontaneous decisions about necessary adjustments to the environment (Cernat and Sakshaug 2021). Another potential error source is the correct use of technical equipment and the recording of (potentially rounded) measurement values (Ulijaszek and Kerr 1999; Cernat and Sakshaug 2021). Further, the correct application of a measurement device often depends on the respondent's position and requires instruction and assessment by the person administering the measurement (Armstrong 2002; Dickson and Hajjar 2007).

Further tasks of interviewers include motivating respondents to perform the measurement to the best of their abilities, providing clarification and assistance when respondents have trouble performing the measurement, and reacting when they notice a mistake has been made. Interviewers' varying behavior can influence measurement results when they repeat the test multiple times (perhaps by respondent request), possibly leading to lower performance of the already exhausted respondent for each subsequent measurement.

Due to these potential influences, we expect physical performance measures to exhibit higher interviewer effects than self-reported (anthropometric) measures. This expectation will be tested while answering the first RQ about the magnitude of interviewer effects on biomeasures. However, not all biomeasures require the same degree of interviewer involvement and spontaneous adjustment to the situation and environment. Differences between biomeasures are the subject of the second RQ. Timed physical movements (e.g., chair stand, walking speed) are expected to be especially prone to interviewer variation as they require high interviewer involvement to set up the space, explain and demonstrate the movement, and operate the stopwatch. In contrast, performance measures that are administered using a specialized technical device (e.g., grip strength, peak flow) require less interviewer involvement and are therefore expected to be less prone to interviewer effects.

A further aspect that we consider part of the third RQ concerns differences in interviewer effects on biomeasures across countries and waves. The complex task of measuring physical performance requires careful interviewer selection and training alongside continuous supervision and monitoring, which depend on the financial and personnel resources of the national survey agencies. These resources might not be equal in all countries, even for the same survey (Börsch-Supan et al. 2013; Sakkeus et al. 2013; Markova et al. 2019).

This may result in differing levels of interviewer variation in biomeasurements across countries. However, when collected over multiple waves, measurement quality may improve as the survey agencies, interviewers, and respondents benefit from their experience and repeated training. Thus, while we expect interviewer effects to differ between countries, we also expect them to diminish over time.

3. DATA

The SHARE aims to provide "micro-level panel data of economic, social and health factors that accompany and influence ageing processes at the individual and societal levels" (Börsch-Supan et al. 2013, p. 993). Since the first wave in 2004/2005, data collection is repeated every second to third year and takes place in several European countries and Israel. The SHARE target population consists of all persons aged 50 or older who are domiciled in a SHARE country in the year of sampling. Further, it includes their partners living in the same

household. From the second wave onward, the sample consists of respondents from any earlier wave, country-specific refreshment samples, and the (baseline) samples of new SHARE countries in the respective wave. Biomeasure collection in SHARE started with the first wave in 2004/2005. We analyze all waves where biomeasure collection was performed through the eighth wave in 2019/2020 (for data citations, see Börsch-Supan 2020a,b,c,d,e,f, 2021), excluding the third wave SHARELIFE and the COVID-19 telephone surveys from 2020 onward. SHARELIFE is the third wave of SHARE fielded in 2008/2009. The SHARELIFE questionnaire focuses on the retrospective collection of respondents' life histories and differs strongly from the regular SHARE questionnaire. Grip strength is the only biomeasure collected in SHARELIFE.

The number of participating SHARE countries continuously increased from 12 in the first wave to currently 27 countries (see table A1 in the supplementary data online). Simultaneously, sample sizes and the number of interviewers grew from 30,424 completed interviews by 774 interviewers in wave 1 to a maximum of 77,261 respondents and 1,931 interviewers in wave 7. Individual-level response rates (response rate 3, American Association for Public Opinion Research 2016) in the baseline/refreshment samples from waves 1 to 7 ranged from 39.5 percent (wave 5) to 48.1 percent (waves 1 and 7) (Bergmann et al. 2019a, table A2 in the supplementary data online). The eighth wave was interrupted by the outbreak of the COVID-19 pandemic in Europe in March 2020. Fieldwork was suspended at a time when 70 percent of the expected longitudinal and 50 percent of the refreshment interviews had been conducted (Scherpenzeel et al. 2020). Retention rates of respondents from previous waves as well as the development of the longitudinal samples by country are documented in figure A1 in the supplementary data online and more comprehensively in Bergmann et al. (2019b, 2022).

Interviewer workloads have remained stable over time. The median number of interviews per interviewer ranges from 29 to 33. Information about the geographical areas of data collection is derived from primary sampling unit (PSU) identifiers. Interpenetration of interviewers and sampling units is acceptable for studying interviewer effects and improves over the waves (figure A2 in the supplementary data online and tables A3 and A4 in the supplementary data online). In total, 62 percent of the interviewers worked in at least two PSUs. In 32 percent of the PSUs, more than one interviewer worked. In most of the PSUs, one to five interviewers worked, and vice versa. Both standardized and detailed instructions for interviewers as well as interviewer trainings are implemented in SHARE (Börsch-Supan and Jürges 2005; Malter and Börsch-Supan 2013, 2015, 2017; Bergmann et al. 2019b; Bergmann and Börsch-Supan 2021).

We analyze four biomeasures that were collected in SHARE: timed chair stand, walking speed (twice), grip strength of both hands (twice), and peak expiratory flow (a measure of lung strength; twice). Table 1 provides a detailed description of the measures and interviewer tasks. Further details are given in

appendix B in the supplementary data online. Self-reported anthropometric measures (height and weight) are included as comparative measures. Biomeasure collection and the main interview are conducted using computer-assisted personal interviewing in respondents' homes (Das et al. 2005). Some of the biomeasures were only collected in certain waves and/or from subsamples of respondents. All physical performance measures are continuous variables.

The SHARE data provide a wide range of respondent characteristics that function as control variables in the modeling approach described in the next section. We used several sociodemographic variables, including respondents' age in the year of the interview, gender, educational level, current employment status, whether they were born in the country of the interview, and whether they have a partner. Further control variables relate to the respondents' living situation: whether they live alone, whether they own or rent their home, and the type of the building and area. The general health status of the respondent is accounted for by self-reported health and a question on long-term illness. To control for the respondent's earlier experience with biomeasure collection, we included a variable on whether it is the first SHARE interview the respondent ever participated in. The selection of these control variables was informed by the nonresponse and aforementioned biomeasure participation literature. Descriptive statistics of outcome and control variables are given in tables A5, A8, and A9 in the supplementary data online.

4. MODELING APPROACH AND ANALYSES

4.1 Detection of Interviewer Effects

The common model for detecting interviewer effects is a multilevel/hierarchical model (Hox et al. 1991; Hox 1994; O'Muircheartaigh and Campanelli 1998; Schnell and Kreuter 2005; West et al. 2018; Beullens et al. 2019). Based on the inclusion of group-specific error terms for the different levels, it allows one to decompose the unexplained variance of an outcome variable into different sources of variation in hierarchically structured data (Vassallo et al. 2017).

Data collected by interviewers naturally have such a hierarchical structure: the respondents (on the first level) are nested hierarchically within interviewers (the second level). As SHARE is a cross-national survey, the interviewers are themselves nested within countries (third level).

The decomposition of variance into the different levels also allows the estimation of ICCs. Within-interviewer correlations should be close to zero when respondents are randomly assigned to interviewers and there are no nonsampling errors. Therefore, the ideal condition to detect interviewer effects due to measurement error would be given in an experiment where interviewers and respondents are assigned randomly (Hox 1994) and there is no selective nonresponse.

Table 1. Biomeasures Collected in SHARE

Type	Biomeasure	Measurement	Unit	Device	Interviewer tasks	Waves
Physical performance	Chair stand (csfive_sqrt)	Once	Seconds	Stopwatch, normal chair	Explain and demonstrate; setup the test; record time for five stands	2, 5
	Walking speed 1, 2 (ws1_sqrt, ws2_sqrt)	Twice	Seconds	Stopwatch, tape measure, masking tape	Set up, explain, and demonstrate the walking course; record time and round to two decimals	1, 2
	Peak flow 1, 2 (pf1, pf2)	Twice	Liters per minute	Mini-Wright peak flow meter	Explain and demonstrate; motivate to blow as hard and fast as possible; record measured value	2, 4, 6
	Grip strength left 1, 2 (gripl1, gripl2); grip strength right 1, 2 (gripr1, gripr2)	Twice on each hand, in altering order	Kilograms	Smedley hand dynamometer (0–100 kg)	Explain and demonstrate; position respondent correctly; adjust dynamometer to hand size; let respondent practice with one hand; motivate to squeeze as hard as possible; record result to the nearest integer value	1–8
Self-report	Height (height)	Once		Questionnaire	Ask question, record value	1–8
	Weight (weight)	Once	Kilograms	Questionnaire	Ask question, record value	1–8

Otherwise, interviewer measurement effects can be mistaken for selection or area effects. Yet, to save time and travel costs, interviewers in large-scale face-to-face surveys are usually assigned to respondents that live in the same area or region. The standard approach in such non-experimental settings is to analyze interviewers and areas in a cross-classified multilevel model to disentangle area and interviewer effects (Hox 1994; Schnell and Kreuter 2005; West et al. 2018). Respondent characteristics (mentioned above) are included to control for selection.

4.2 The Cross-Classified Multilevel Model

The cross-classified multilevel model is defined as

$$y_{i(j,k)l} = \gamma + \sum \beta x_{i(j,k)l} + \nu_l + \eta_k + \upsilon_j + \varepsilon_i,$$

where the dependent variable y varies around the overall intercept γ by individual (i), area (j), interviewer (k), and country (l). Group-level error terms are estimated conditionally on the individual-level characteristics x with regression coefficients β to control for respondent selectivity and for respondent-induced measurement errors. The random effects terms ν_l (countries), η_k (interviewers), and υ_j (areas) are assumed to be normally distributed with mean zero and variance τ^2_ν, τ^2_η, and τ^2_υ, respectively. The residual error term ε_i also has mean zero. Following the location-scale modeling approach by Brunton-Smith et al. (2017) and Sturgis et al. (2021), we assume the residual variance σ^2_ε to vary between the interviewer groups. This is based on the idea that varying interviewer behavior does not necessarily bias the measurements by one interviewer in a certain direction (error term η_k) but can also contribute to a higher or lower dispersion of measurements within a group. The residual standard deviation is modeled in an additional log-linear equation $\ln(\sigma_{\varepsilon_i}) = a + u_k$, where a is the intercept and u_k is the varying interviewer-specific component of the residual standard deviation. The overall residual variance is estimated as $\sigma^2_\varepsilon = \exp(a + 0.5\sigma_{u_k})^2$, where σ_{u_k} is the standard deviation of the estimated interviewer-specific components u_k (Brunton-Smith et al. 2017).

The variance partition coefficient (VPC) for interviewers

$$\mathrm{VPC}_\eta = \tau^2_\eta / (\tau^2_\eta + \tau^2_\nu + \tau^2_\upsilon + \sigma^2_\varepsilon)$$

can be interpreted as the proportion of variance related to the interviewers (Goldstein et al. 2002; Gelman and Hill 2007). The within-interviewer correlation ρ_η—the correlation in error terms of two respondents assigned to the same interviewer, usually referred to as the ICC coefficient—in the cross-classified three-level model depends on whether the respondents of one interviewer are also from the same area or not (Snijders and Bosker 2012; West et al. 2015). The intraclass correlation of respondents from one country with the same interviewer but from a different area would be

$\rho_{\text{different}} = (\tau_\eta^2 + \tau_\nu^2)/(\tau_\eta^2 + \tau_\nu^2 + \tau_\upsilon^2 + \sigma_\varepsilon^2)$, while the ICC with the same interviewer and same area would be $\rho_{\text{same}} = (\tau_\eta^2 + \tau_\nu^2 + \tau_\upsilon^2)/(\tau_\eta^2 + \tau_\nu^2 + \tau_\upsilon^2 + \sigma_\varepsilon^2)$. These ICCs are relevant for the estimation of the variance inflation of population estimates according to Kish (1962), as described in section 2.1. However, to keep the interpretation and comparison of results as simple as possible, we only report VPCs in the results section. The rationale behind this is that VPC and ICC differ only in the numerator. As the variance components τ are always positive and the denominator remains the same, ICCs are always equal to or higher than the respective VPC. We therefore report the VPCs as an estimator for the lower bound of the within-interviewer correlation and the subsequent potential variance inflation.

The magnitude of variation of the residual variances σ_{ε_k} depends on the scale of the outcome variable and therefore cannot be compared between models for different outcomes. However, the estimation of heterogeneous residual variances allows the calculation of interviewer-specific VPCs (Sturgis et al. 2021):

$$\text{VPC}_k = \tau_\eta^2/(\tau_\eta^2 + \tau_\nu^2 + \tau_\upsilon^2 + \exp(a + u_k)^2).$$

The deviation of these interviewer-specific VPCs from the averaged VPC for the respective outcome and wave gives an idea of how relevant differences in within-interviewer variation are in comparison to the between-interviewer variation of error terms.

4.3 Estimation

Models were fitted in R version 4.1.2 (R Core Team 2021) using the brms package for the estimation of Bayesian hierarchical models with Markov Chain Monte Carlo (Bürkner 2017, 2018). The brms package works with the probabilistic programming language Stan (Stan Development Team 2022).

Draws from the posterior distribution were simulated using 5 chains with 5,000 iterations each and 2,500 as burn-in, resulting in a total of 12,500 iterations after warm-up. For the chair stand variables, the number of iterations had to be increased from 5,000 to 10,000 per chain to reach convergence. One problem in the country models was the occurrence of divergent transitions in the MCMC sampling process (Stan Development Team 2020). The walking speed country models were therefore excluded from the country comparison due to imprecise and unreliable estimations.

As prior distributions for the group-level standard deviations, weakly informative half-Student-*t* distributions were adopted from the default settings for the linear model in brms (Bürkner 2017). For the respondent-level regression coefficients, the default flat priors were replaced by scaled weakly informative priors as used in the rstanarm package (Gabry and Goodrich 2020). A

sensitivity analysis using the default flat priors (conducted for some selected models to save computation time) led to the same results as the models with the scaled weakly informative priors.

To answer the first and second RQs about the size of interviewer effects in the biomeasurements, three three-level models were estimated: (i) an empty model, containing only the group effects for countries, areas, and interviewers, (ii) a model including socio-economic respondent-level covariates to control for interviewer-specific selection mechanisms and measurement errors related to respondents instead of interviewers, and (iii) a model additionally including self-reported health and long-term illness as covariates to account for the general health status of the respondents. To inspect country and wave differences in interviewer effects (third RQ), two-level models including group effects for interviewers and areas were fitted separately for each country and wave. This implicitly introduces country-/wave-specific variances τ_{η_l} for interviewers and τ_{υ_l} for areas on the second level and, therefore, allows the estimation of country-/wave-specific interviewer and area VPCs. The country/wave models are conditioned on the same respondent characteristics as the overall three-level model.

All reported point estimates are obtained from the medians of the respective posterior samples. Reported credible intervals cover the 95 percent-highest-density intervals of those posterior distributions.

5. RESULTS

5.1 Magnitude of Interviewer Effects (RQ 1)

The overall amount of interviewer-related variance on biomeasures is evaluated on the basis of the three-level model including the complete set of respondent characteristics as covariates (model 3). Values of point and interval estimates for all models, outcomes, and waves are given in tables A5, A8, and A9 in the supplementary data online. Looking at all biomeasures, waves, and countries in model 3, interviewer VPCs range from 0.05 to 0.28. The interviewer VPCs indicate that up to 28 percent of unexplained variation in the physical performance outcomes are related to the interviewers conducting the measurements. Only 1 to 2 percent of the unexplained variance can be ascribed to the PSUs, which seems negligible compared to the interviewer variance components. The country VPCs reach values around 0.10 for some biomeasures as well as for self-reported height, indicating that there are either country-specific measurement errors or actual systematic differences in the physical composition of respondents. While both ideas are plausible, they are not a part of the RQs.

Figure 1 shows the proportions of the group-specific and residual variance components relative to the total unexplained variance for each biomeasure by

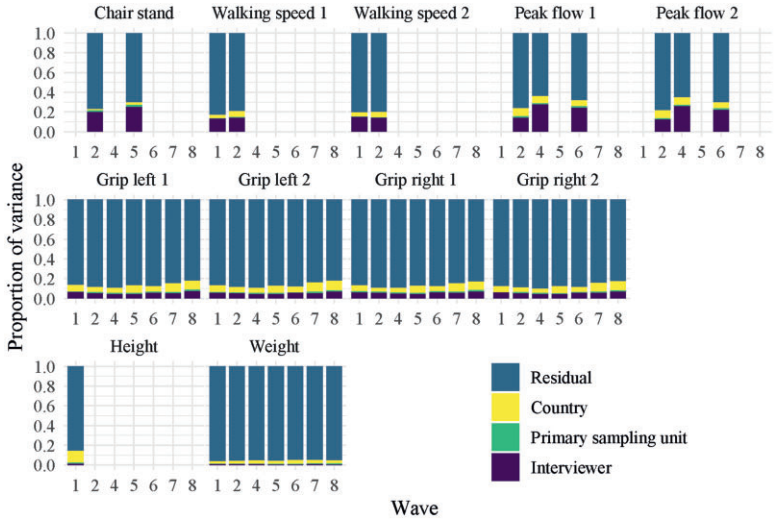

Figure 1. Variance Components for Interviewers, Countries, Primary Sampling Units, and Residuals in the Three-Level Model with All Covariates (Model 3). Not all biomeasures were collected in all waves. Chair stand was only collected in the second and fifth waves, walking speed in the first two waves, and peak flow in the second, fourth, and sixth waves. Self-reported height for all respondents was only collected in the first wave.

wave. With interviewer VPCs of 0.01 and 0.02 for self-reported height and weight, respectively, they have less potential for interviewer effects than the biomeasures themselves.

5.2 Differences between Types of Biomeasures (RQ 2)

While the gap in interviewer variation between the biomeasures and self-reported variables is very clear, the range of interviewer VPCs among the different biomeasures is also rather large. This leads to the second RQ, regarding varying levels of interviewer variation between different types of biomeasures.

Figure 2 illustrates that chair stand, walking speed, and peak flow have the highest proportions of interviewer variance, followed by grip strength and lastly the self-reported variables. The timed performance measures (chair stand, walking speed) were expected to exhibit higher interviewer variation due to their strong interviewer involvement, compared to the measures that rely on specialized technical devices (grip strength, peak flow) with less interviewer involvement. The chair stand test with VPCs above 0.2 and the grip strength measurements with VPCs below 0.1 are in line with this expectation. The walking speed measure also contains a considerable proportion of interviewer-related variance. Rather surprising are the results for peak flow,

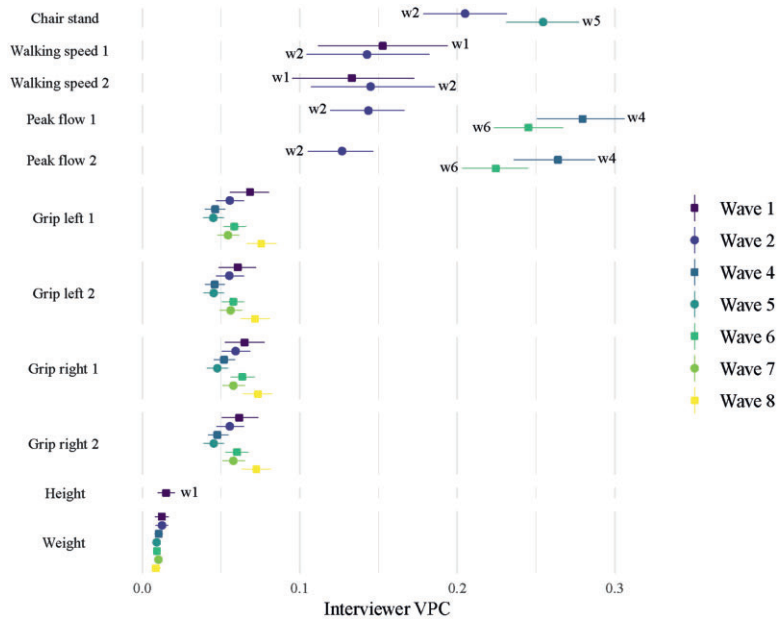

Figure 2. Interviewer VPCs in All Waves by Type of Biomeasure (with 95 Percent Credible Intervals). Not all biomeasures were collected in all waves. Chair stand was only collected in the second and fifth waves, walking speed in the first two waves, and peak flow in the second, fourth, and sixth waves. Self-reported height for all respondents was only collected in the first wave.

with VPC values lying above 0.1 in the second wave and above 0.2 in the subsequent waves. This is against the expectation that peak flow and grip strength have a comparable risk of interviewer effects because they are both based on a specialized technical device and require similar instructions. Possible causes of this unexpected empirical discrepancy will be addressed in section 6. Credible intervals of the peak flow measure overlap (and sometimes exceed) those of the chair stand and walking speed measures.

Another element of the three-level model was the estimation of variability in the unexplained within-interviewer variance. The dispersion of the interviewer-specific VPCs in figure 3 illustrates that while interviewer-specific VPCs for grip strength are stable, the consideration of heterogeneous residual variances leads to large dispersion of the VPCs for chair stand, walking speed, and peak flow test in the second wave. Plots for the other waves are shown in figure C2 in the supplementary data online. All plots indicate that heterogeneous (unexplained) variability across the respondents of different interviewers is especially relevant for the biomeasures that also have high overall interviewer variance proportions. This suggests that a relevant amplification or dampening of differences in the outcomes of their respondents by interviewers occurs

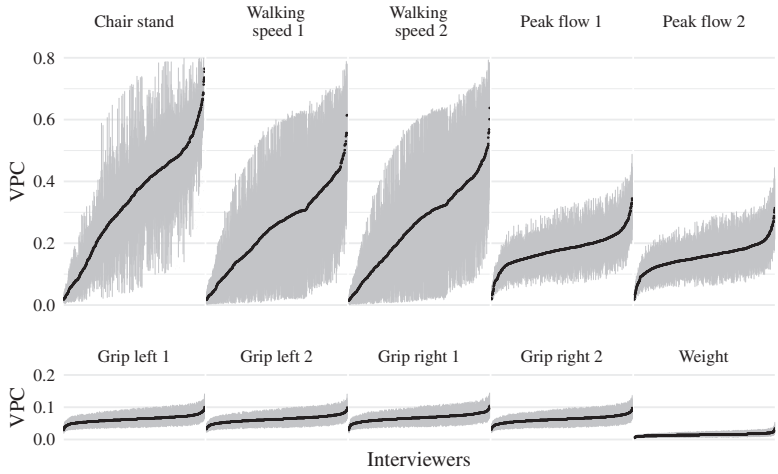

Figure 3. Interviewer-Specific VPCs in Wave 2 by Biomeasure (with 95 Percent Credible Intervals). A diagonal (instead of horizontal) pattern indicates heterogeneity of interviewer-specific residual variances.

especially with biomeasures that are generally more susceptible to interviewer effects.

5.3 Interviewer Effects across Countries (RQ 3a)

The third RQ concerns differences in interviewer effects across countries and waves. Both are assessed based on interviewer VPCs estimated in two-level models separately for each country. Not all countries participated in all waves.

Figure 4 shows the distribution of the country-specific interviewer VPCs for chair stand, peak flow, grip strength, and weight in waves 2–7. The specific point estimates of the interviewer VPCs by country are given in table C4 in the supplementary data online. For the chair stand test in wave 2, most credible intervals of the country-specific interviewer VPCs are overlapping. However, a group of countries with very high interviewer VPCs (Austria, Germany, Greece, Spain; 0.30–0.43) differs significantly from the countries below the average interviewer variance proportion (Belgium, France, Netherlands, Poland; 0.09–0.10). The pattern looks similar in the fifth wave but with different countries having high (Austria, Czech Republic, Spain, Italy; 0.37–0.45) and low (Switzerland, Germany, Estonia, France, Luxemburg; 0.05–0.15) VPCs. It should be noted that the smaller VPCs of around 0.10 are moderate but still not negligible.

Regarding the interviewer VPCs for peak flow, country differences are clearly visible and have less overlapping credible intervals. In wave 2, Spain, Greece, and Italy have proportions of interviewer-related variance above 20

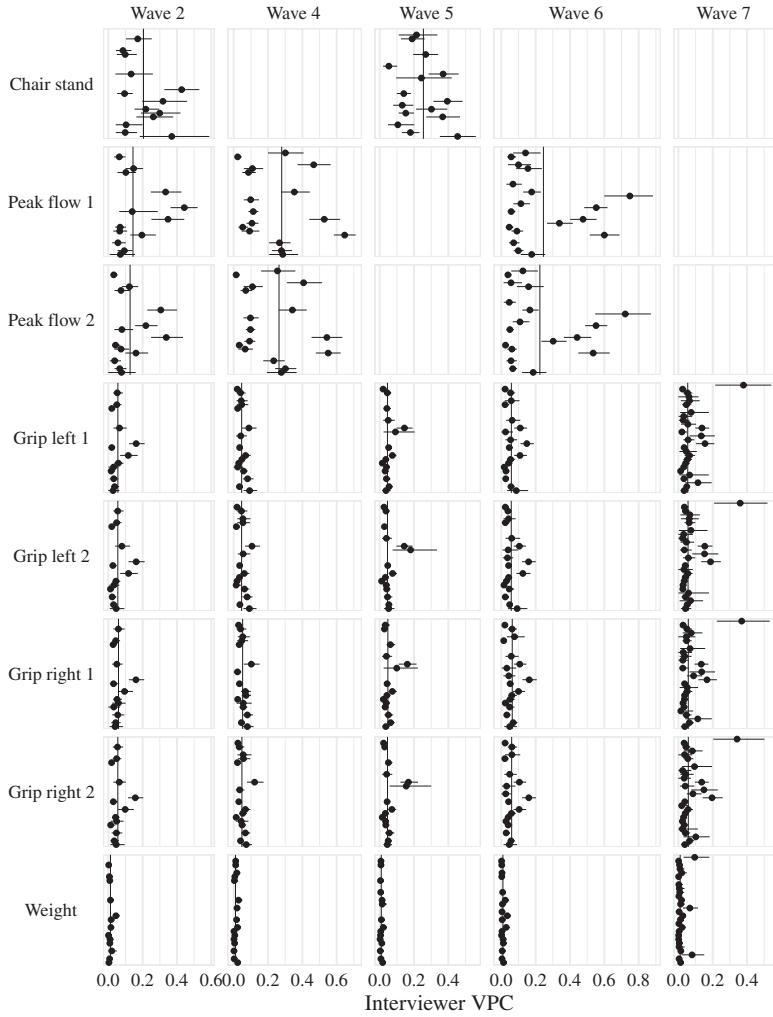

Figure 4. Distribution of Country-Specific Interviewer VPCs by Biomeasure and Waves (Each Point Denotes the Interviewer VPC in One Country, with a 95 Percent Credible Interval). Results for walking speed were excluded because VPCs could not be estimated precisely and reliably. The vertical black lines indicate the estimated interviewer VPC from the three-level model for the respective biomeasure and wave.

percent, while it reaches a maximum of 15 percent for Belgium, Switzerland, Germany, Denmark, and Sweden. Regarding grip strength, the majority of country-specific interviewer VPCs overlap with the mean or lie below it. It is possible, however, to identify single countries that exceed the mean

interviewer VPC and also the critical threshold of 0.10 (e.g., Greece and Spain in wave 2; Italy in wave 4; Israel and Italy in wave 5; Spain and Greece in wave 6; Bulgaria, Greece, Hungary, Italy, and Slovakia in wave 7). Interviewer variation increases in the later waves for grip strength as well as for the peak flow measurements. What should be noted is that it is not always the same countries that exhibit high interviewer VPCs, but it differs across waves and biomeasures.

While variation in the proportions of interviewer variance across countries was expected for all biomeasures, some of the estimated VPCs are concerningly high, reaching values of 0.50 and above. Plotting the estimated interviewer-specific error terms ($\widehat{\eta}_k$) for those countries suggests that high VPCs are at least partly driven by a few interviewers with extreme error terms rather than by generally high variation across interviewers (see figure C3 in the supplementary data online). Following this presumption, exemplary models for a few countries with very high interviewer VPCs were refitted under exclusion of those interviewers. As a result, interviewer variation significantly decreased in all countries (see table C5 in the supplementary data online). This means that at least part of the strong interviewer effects could be prevented by monitoring interviewer deviations and taking measures against those deviations during fieldwork (e.g., retraining interviewers, checking their measurement devices).

5.4 Interviewer Effects across Waves (RQ 3b)

The expectation regarding interviewer effects across waves was that they would decrease due to growing expertise of survey agencies and interviewers, and presumed methodological improvements in the survey process. Table 2 presents the average VPCs in chair stand, peak flow, and grip strength over all waves. Distinction is made between the interviewer VPCs that are averaged over the countries that participated in every wave of the respective biomeasure collection (i.e., excludes the entering of new countries) and all countries (new and old) that participated in any given wave. Despite the expectation of growing experience, interviewer variance proportions increased from the first to the second waves that chair stand and peak flow tests were conducted. From the second round (wave 4) to the third round (wave 6) of peak flow collection, the interviewer VPC decreases at a faster rate in the countries that participated in the earlier waves compared to the mean VPC of all countries. This could be an example of countries with more experience and improved procedures exhibiting lower interviewer effects. The interviewer-related variance in the grip strength variables reaches its maxima in the first and the eighth waves of SHARE. These patterns remain the same when controlling for changes in national survey agencies between the waves. Plots of the trends in the single

Table 2. Average interviewer VPCs for countries participating in every wave versus all countries (in parentheses)

	Wave 1	Wave 2	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Chair stand	–	0.20	–	0.27	–	–	–
(all)	–	(0.21)	–	(0.24)	–	–	–
Peak flow	–	0.13	0.24	–	0.17	–	–
(all)	–	(0.14)	(0.23)	–	(0.22)	–	–
Grip strength	0.07	0.05	0.06	0.06	0.06	0.05	0.07
(all)	(0.07)	(0.05)	(0.05)	(0.05)	(0.06)	(0.07)	(0.09)

countries are given in figure C4 in the supplementary data online. Trends are not always identical across the countries or go in the same direction.

6. DISCUSSION

6.1 Findings

This cross-national analysis of interviewer effects on the measurement of physical performance yielded four main findings. First, moderate-to-large proportions of VPCs ranging from 0.05 to 0.28 were detected across the four biomeasures: timed chair stand, walking speed, grip strength, and peak flow. This was in contrast to VPCs for the self-reported anthropometric measures, height and weight, which exhibited relatively smaller VPCs (0.01 and 0.02, respectively). Thus, the expectation that physical measures are particularly susceptible to interviewer effects was empirically supported and is in line with previous findings (Cernat and Sakshaug 2021). Second, differences in the magnitude of the interviewer effects were observed between the different types of biomeasures, with the largest VPCs detected for peak flow (0.13–0.28) followed by chair stand (0.20–0.25), walking speed (0.13–0.15), and grip strength (0.05–0.08). The ranking of biomeasures differed from previous findings in other biosocial surveys where nurse effects of around 15–20 percent were found for grip strength and peak flow (Cernat and Sakshaug 2020), and interviewer effects of around 10 percent were found for timed chair stand and timed walk (Cernat and Sakshaug 2021).

The timed performance measures (chair stand, walking speed) requiring more interviewer involvement were expected to have larger interviewer effects than the measures dependent on specialized technical devices (grip strength, peak flow). While this expectation held for grip strength, peak flow exhibited VPCs that were comparable to, and sometimes larger than, the timed movements. This surprisingly high interviewer variation and apparent difference between grip strength and peak flow measurements could be caused in different ways. One

explanation might be that peak expiratory flow strongly depends on the respondent's position during the test, and interviewers may differ in their instructions, in letting respondents practice in advance, or in their strictness when administering the test. Another cause for the strong interviewer variation could be peak flow device issues. According to the data collection documentation, all interviewers should be using the same peak flow meter device. Yet, there can be differences in device precision or some unnoticed malfunctions of single devices. Since one interviewer usually uses the same device throughout all their interviews, this could lead to confounding device effects appearing as interviewer variation.

Third, all biomeasures exhibited strong differences in interviewer VPCs between countries. These differences were especially pronounced for the chair stand and peak flow test. A rather surprising result was the very high interviewer VPCs in single countries for some biomeasures. Nevertheless, the country differences in interviewer effects are compatible with cross-national findings on interviewer effects in other survey measurements (Beullens and Loosveldt 2016; Cernat et al. 2019).

Lastly, there was no consistent pattern in the average interviewer VPCs across subsequent waves of data collection. The analysis and interpretation across waves are limited by the fact that we do not know the percentage of interviewers who worked in prior waves, which can differ across countries and waves. What should be noted is that an observed decrease in the VPCs for one biomeasure over time does not always coincide with a decrease in the VPCs of a different biomeasure collected within the same country. This means that overall training or experience does not automatically lead to better data quality for all biomeasure data. While we expected the opposite, there is some evidence that altered behavior by more experienced interviewers does not necessarily reduce measurement errors but can have different effects (Olson and Peytchev 2007). Further investigations of the role of interviewer experience, how it interferes with respondents' increased age, and the relevance of changes in fieldwork organization would require more information on the latter, as well as longitudinal interviewer IDs. This could be a subject of further research.

6.2 Limitations

As with all studies, this one has limitations that should be mentioned. First, the separation of area and interviewer effects relies on statistical controls instead of random allocation of interviewers to respondents. While this is a standard approach, there remains the risk of unobserved confounders. To minimize this risk, our models included two types of control variables: area identifiers, which are interpenetrated with interviewers, and a large number of respondent-level covariates based on the nonresponse and biomeasure participation literature. Regarding the respondent-level characteristics, a comparison of the models with and without covariates showed that some of the covariates are strong

predictors of the biomeasure outcomes, while the dimension and ranking of interviewer effects are unchanged. Regarding the potential confounding of interviewers and areas, the models with and without covariates provided only limited insights into whether the former explain the parts of potential area effects that are not controlled through the area identifiers. While the area variance components are partially explained by respondent characteristics for some biomeasures, the effects of the covariates on other biomeasures are varying and unclear. The consequences for our results, however, can be narrowed down when considering the separate analyses of the two-level models for the individual countries, some of which have a very good interpenetration of interviewers and areas, leading to better statistical control of potential confounders. The interviewer effects estimated for those countries do not show any consistent differences compared to the ones for countries with low interpenetration, giving the results presented above some validity. What should be noted further is that the existence of unobserved confounders in our model would mean that the observed within-group homogeneity is erroneously attributed to interviewers. The effect, however, on variance inflation and the need to consider ICCs in statistical analyses still holds. Another limitation is the absence of interviewer characteristics for most waves and interviewers that would allow for further explanations of interviewer variation. The same holds true for longitudinal interviewer IDs. These are areas for future work. Given these limitations, the empirical results still have implications for substantive analyses of biomeasure data and for the collection of the biomeasures themselves.

6.3 Practical Implications and Conclusions

Intra-interviewer correlations lead to interviewer effects in the form of variance inflation for population estimates. Even a relatively small interviewer ICC of, for example, 0.05 in combination with a median interviewer workload of 28 interviews can more than double (factor 2.35) the variance of mean grip strength in the population. Thus, the interviewer-related variances and homogeneities identified in the analyses above can lead to potentially strong interviewer effects on the variances of descriptive population estimates. This inflation of variances is equivalent to an undesired reduction of the analytic sample size. Country-specific interviewer measurement errors, as indicated by the interviewer VPCs differing across countries, could lead to differences in regression coefficients that are "erroneously attributed to real country differences" (Beullens and Loosveldt 2016) when interviewer clustering is not considered.

Recommendations for researchers working with biomeasures are to conduct sensitivity analyses that account for interviewer or nurse effects in the data and not take these measures to be automatically free from measurement error. Propositions for sensitivity analyses or adjustment methods are given in, for

example, O'Muircheartaigh and Campanelli (1998) and Fischer et al. (2019). An important precondition is that survey agencies make interviewer IDs available.

The observed interviewer variation has further practical implications for survey design and the organization of data collection. Knowing which biomeasures are prone to interviewer effects and a better understanding of the underlying mechanisms can inform improvements, for example, in interviewer instructions, recruitment, training, and fieldwork monitoring. The present study showed that, in particular, timed and context-dependent measurements, including walking speed and chair stand, but also the application of specialized measurement devices, such as the peak flow meter, are particularly susceptible to country-specific interviewer variation. One way to prevent interviewer variation could be to double-check all devices for measurement precision or malfunction before handing them to interviewers because it cannot be ruled out that part of the interviewer effects are actually device effects. Simultaneous fieldwork monitoring could be applied to re-train interviewers with many extreme values. A systematic documentation of difficulties with certain measurements that appear already during the interviewer trainings could be used to improve the training or the measurement procedure.

The existence of interviewer effects on biomeasures is not automatically an argument to drop them from survey programs or not use them in analyses. However, they can introduce additional uncertainty in population estimates and lead to false inferences when the clustering of observations in interviewers is not accounted for. This makes their documentation relevant and of interest for researchers planning to work with biomeasure data. From a methodological perspective, the presented approach and findings build a basis for further attempts to detect and address interviewer-related measurement errors in biosocial and cross-national surveys. Further research should attempt to not only detect but explain interviewer effects on biomeasures by means of, for example, interviewer characteristics or additional contextual variables regarding interviewer training and experience, the devices used, and fieldwork organization. These analyses could be extended by using longitudinal interviewer IDs, which would enable more robust analyses of longitudinal interviewer effects and interviewer learning effects.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

American Association for Public Opinion Research (2016), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.), Alexandria, VA: AAPOR.

Angel, B., Ajnakina, O., Albala, C., Lera, L., Márquez, C., and Bendayan, R. (2022), "Grip Strength Trajectories and Cognition in English and chilean Older Adults: A Cross-Cohort Study," *Journal of Personalized Medicine*, 12, 1230.

Armstrong, R. S. (2002), "Nurses' Knowledge of Error in Blood Pressure Measurement Technique," *International Journal of Nursing Practice*, 8, 118–126.

Banks, J., Nazroo, J., and Steptoe, A. (eds.) (2014), *The Dynamics of Ageing: Evidence from the English Longitudinal Study of Ageing 2002-2012 (Wave 6)*, London: The Institute for Fiscal Studies.

Banks, J., and Smith, J. P. (2012), "International Comparisons in Health Economics: Evidence from Aging Studies," *Annual Review of Economics*, 4, 57–81.

Barros, P. P., Pimentel-Santos, F., and Neto, Dias D. (2019), "Grip Strength across Europe - North/South and East/West Divides," in *Health and Socioeconomic Status over the Life Course: First Results from Share Waves 6 and 7*, eds. A. Börsch-Supan, Berlin/Boston: De Gruyter Oldenburg, pp. 327–336.

Bergmann, M., and Börsch-Supan, A. (eds.) (2021), *SHARE Wave 8 Methodology: Collecting Cross-National Survey Data in Times of COVID-19*, Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Bergmann, M., Kneip, T., de Luca, G., and Scherpenzeel, A. (2019a), "Survey Participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1–7," SHARE Working Paper Series, 41, Munich: SHARE-ERIC.

———. (2022), "Survey Participation in the Eighth Wave of the Survey of Health, Ageing and Retirement in Europe (SHARE)," SHARE Working Paper Series, 81, Munich: SHARE-ERIC.

Bergmann, M., Scherpenzeel, A., and Börsch-Supan, A. (eds.) (2019b), *SHARE Wave 7 Methodology: Panel Innovations and Life Histories*, Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Beullens, K., and Loosveldt, G. (2016), "Interviewer Effects in the European Social Survey," *Survey Research Methods*, 10, 103–118.

Beullens, K., Loosveldt, G., and Vandenplas, C. (2019), "Interviewer Effects among Older Respondents in the European Social Survey," *International Journal of Public Opinion Research*, 31, 609–625.

Bodilsen, A. C., Juul-Larsen, H. G., Petersen, J., Beyer, N., Andersen, O., and Bandholm, T. (2015), "Feasibility and Inter-Rater Reliability of Physical Performance Measures in Acutely Admitted Older Medical Patients," *PLoS One*, 10, e0118248.

Börsch-Supan, A. (2020a), "Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1," Release version: 7.1.0. SHARE-ERIC. Data set. doi:10.6103/SHARE.w1.710.

———. (2020b), "Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2," Release version: 7.1.0. SHARE-ERIC. Data set. doi:10.6103/SHARE.w2.710.

———. (2020c), "Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4," Release version: 7.1.0. SHARE-ERIC. Data set. doi:10.6103/SHARE.w4.710.

———. (2020d), "Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5," Release version: 7.1.0. SHARE-ERIC. Data set. doi:10.6103/SHARE.w5.710.

———. (2020e), "Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6," Release version: 7.1.0. SHARE-ERIC. Data set. doi:10.6103/SHARE.w6.710.

———. (2020f), "Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7," Release version: 7.1.1. SHARE-ERIC. Data set. doi:10.6103/SHARE.w7.711.

———. (2021), "Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8," Release version: 1.0.0. SHARE-ERIC. Data set. doi:10.6103/SHARE.w8.100.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., and Zuber, S.; SHARE Central Coordination Team (2013), "Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE)," *International Journal of Epidemiology*, 42, 992–1001.

Börsch-Supan, A., and Jürges, H. (eds.) (2005), *The Survey of Health, Aging, and Retirement in Europe: Methodology*, Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Boudreau, D. M., Daling, J. R., Malone, K. E., Gardner, J. S., Blough, D. K., and Heckbert, S. R. (2004), "A Validation Study of Patient Interview Data and Pharmacy Records for Anterhypertensive, Statin, and Antidepressant Medication Use among Older Women," *American Journal of Epidemiology*, 159, 308–317.

Boyle, J., Berman, L., Dayton, J., Iachan, R., Jans, M., and ZuWallack, R. (2021), "Physical Measures and Biomarker Collection in Health Surveys: Propensity to Participate," *Research in Social and Administrative Pharmacy*, 17, 921–929.

Brunton-Smith, I., Sturgis, P., and Leckie, G. (2017), "Detecting and Understanding Interviewer Effects on Survey Data by Using a Cross-Classified Mixed Effects Location-Scale Model," *Journal of the Royal Statistical Society*, 180, 551–568.

Bur, A., Herkner, H., Vlcek, M., Woisetschläger, C., Derhaschnig, U., Delle Karth, G., Laggner, A. N., and Hirschl, M. M. (2003), "Factors Influencing the Accuracy of Oscillometric Blood Pressure Measurement in Critically Ill Patients," *Critical Care Medicine*, 31, 793–799.

Bürkner, P.-C. (2017), "Brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software*, 80, 1–28.

———. (2018), "Advanced Bayesian Multilevel Modeling with the R Package Brms," *The R Journal*, 10, 395–411.

Carsley, S., Parkin, P. C., Tu, K., Pullenayegum, E., Persaud, N., Maguire, J. L., and Birken, C. S.; TARGet Kids! Collaboration (2019), "Reliability of Routinely Collected Anthropometric Measurements in Primary Care," *BMC Medical Research Methodology*, 19, 84.

Cernat, A., and Sakshaug, J. W. (2020), "Nurse Effects on Measurement Error in Household Biosocial Surveys," *BMC Medical Research Methodology*, 20, 45.

———. (2021), "Interviewer Effects in Biosocial Survey Measurements," *Field Methods*, 33, 236–252.

Cernat, A., Sakshaug, J. W., and Castillo, J. (2019), "The Impact of Interviewer Effects on Skin Color Assessment in a Cross-National Context," *International Journal of Public Opinion Research*, 31, 779–793.

Cernat, A., Sakshaug, J. W., Chandola, T., Nazroo, J., and Shlomo, N. (2021), "Nurse Effects on Non-Response in Survey-Based Biomeasures," *International Journal of Social Research Methodology*, 24, 487–499.

Cress, M. E., Buchner, D. M., Questad, K. A., Esselman, P. C., deLateur, B. J., and Schwartz, R. S. (1996), "Continuous-Scale Physical Functional Performance in Healthy Older Adults: A Validation Study," *Archives of Physical Medicine and Rehabilitation*, 77, 1243–1250.

Crimmins, E., Faul, J., Kim, J. K., and Weir, D. (2015), *Documentation of Biomarkers in the 2010 and 2012 Health and Retirement Study*, Ann Arbor: University of Michigan Survey Research Center.

Das, M., Vis, C., and Weerman, B. (2005), "Developing the Survey Instruments for SHARE," in *The Survey of Health, Aging, and Retirement in Europe*, eds. A. Börsch-Supan and H. Jürges, Mannheim: Mannheim Research Institute for the Economics of Aging (MEA), pp. 12–23.

Davillas, A., and Jones, A. M. (2021), "The Implications of Self-Reported Body Weight and Height for Measurement Error in BMI," GLO Discussion Paper, No. 919, Essen: Global Labor Organization (GLO).

de Winter, A. F., Heemskerk, M. A. M. B., Terwee, C. B., Jans, M. P., Devillé, W., van Schaardenburg, D.-J., Scholten, R. J. P. M., and Bouter, L. M. (2004), "Inter-Observer Reproducibility of Measurements of Range of Motion in Patients with Shoulder Pain Using a Digital Inclinometer," *BMC Musculoskeletal Disorders*, 5, 18–2474.

Dickson, B. K., and Hajjar, I. (2007), "Blood Pressure Measurement Education and Evaluation Program Improves Measurement Accuracy in Community-Based Nurses: A Pilot Study," *Journal of the American Academy of Nurse Practitioners*, 19, 93–102.

Durand, M.-J., Loisel, P., Poitras, S., Mercier, R., Stock, S. R., and Lemaire, J. (2004), "The Interrater Reliability of a Functional Capacity Evaluation: The Physical Work Performance Evaluation," *Journal of Occupational Rehabilitation*, 14, 119–129.

Dykema, J., DiLoreto, K., Croes, K. D., Garbarski, D., and Beach, J. (2017), "Factors Associated with Participation in the Collection of Saliva Samples by Mail in a Survey of Older Adults," *Public Opinion Quarterly*, 81, nfw045.

Ezzati, M., Martin, H., Murray, C. J. L., Skjold, S., and Vander Hoorn, S. (2006), "Trends in National and State-Level Obesity in the Usa after Correction for Self-Report Bias: Analysis of Health Surveys," *Journal of the Royal Society of Medicine*, 99, 250–257.

Fischer, M., West, B. T., Elliott, M. R., and Kreuter, F. (2019), "The Impact of Interviewer Effects on Regression Coefficients," *Journal of Survey Statistics and Methodology*, 7, 250–274.

Footman, K. (2021), "Interviewer Effects on Abortion Reporting: A Multilevel Analysis of Household Survey Responses in Côte D'Ivoire, Nigeria and Rajasthan, India," *BMJ Open*, 11, e047570.

Franzese, F. (2015), "Slipping into Poverty: Effects on Mental and Physical Health," in *Ageing in Europe—Supporting Policies for an Inclusive Society*, eds. A. Börsch-Supan, T. Kneip, H. Litwin, M. Myck, and G. Weber, Berlin/Munich/Boston: De Gruyter, pp. 139–148.

Gabry, J., and Goodrich, B. (2020), "Prior Distributions for rstanarm Models." Available at http://mc-stan.org/rstanarm/articles/priors.html. Accessed May 2, 2022.

Gavrilova, N., and Lindau, S. T. (2009), "Savilary Sex Hormone Measurement in a National, Population-Based Study of Older Adults," *Journal of Gerontology: Social Sciences*, 64B, 94–105.

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge: Cambridge University Press.

Gil, J., and Mora, T. (2011), "The Determinants of Misreporting Weight and Height: The Role of Social Norms," *Economics and Human Biology*, 9, 78–91.

Goldstein, H., Browne, W., and Rasbash, J. (2002), "Partitioning Variation in Multilevel Models," *Understanding Statistics*, 1, 223–231.

Gorber, S. C., Tremblay, M., Moher, D., and Gorber, B. (2007), "A Comparison of Direct vs. self-report Measures for Assessing Height, Weight and Body Mass Index: A Systematic Review," *Obesity Reviews*, 8, 307–326.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009), *Survey Methodology* (2nd ed.), Hoboken, NJ: Wiley.

Guyer, H., Ofstedal, M. B., Lessof, C., and Cox, K. (2017), *The Benefits and Challenges of Collecting Physical Measures and Biomarkers in Cross-National Studies*, Ann Arbor, MI: Institute for Social Research.

Heeb, J. L., and Gmel, G. (2001), "Interviewers' and Respondents' Effects on Self-Reported Alcohol Consumption in a Swiss Health Study," *Journal of Studies on Alcohol*, 62, 434–442.

Hox, J. J. (1994), "Hierarchical Regression Models for Interviewer and Respondent Effects," *Sociological Methods & Research*, 22, 300–318.

Hox, J. J., de Leeuw, E. D., and Kreft, I. G. G. (1991), "The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model," in *Measurement Error in Surveys*, eds. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, Hoboken, NJ: Wiley, pp. 439–461.

Jaszczak, A., Lundeen, K., and Smith, S. (2009), "Using Nonmedically Trained Interviewers to Collect Biomeasures in a National in-Home Survey," *Field Methods*, 21, 26–48.

Johnson, T. P., and Parsons, J. A. (1994), "Interviewer Effects on Self-Reported Substance Use among Homeless Persons," *Addictive Behavior*, 19, 83–93.

Kish, L. (1962), "Studies of Interviewer Variance for Attitudinal Variables," *Journal of the American Statistical Association*, 57, 92–115.

Korbmacher, J. (2014), "Interviewer Effects on Respondents' Willingness to Provide Blood Samples in SHARE," SHARE Working Paper Series, 20, Munich: SHARE-ERIC.

Kumari, M., and Benzeval, M. (2021), "Collecting Biomarker Data in Longitudinal Surveys," in *Advances in Longitudinal Survey Methodology*, ed. P. Lynn, Hoboken, NJ: Wiley, pp. 26–46.

Leone, T., Sochas, L., and Coast, E. (2021), "Depends Who's Asking: Interviewer Effects in Demographic and Health Surveys Abortion Data," *Demography*, 58, 31–50.

Leong, A., Chiasson, J.-L., Dasgupta, K., and Rahme, E. (2013), "Estimating the Population Prevalence of Diagnosed and Undiagnosed Diabetes," *Diabetes Care*, 36, 3002–3008.

Malter, F., and Börsch-Supan, A. (eds.) (2013), *SHARE Wave 4: Innovations and Methodology*, Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Malter, F., and Börsch-Supan, A. (eds.) (2015), *SHARE Wave 5: Innovations and Methodology*, Munich: MEA, Max Planck Institute for Social Law and Social Policy.
Malter, F., and Börsch-Supan, A. (eds.) (2017), *SHARE Wave 6: Panel Innovations and Collecting Dried Blood Spots*, Munich: MEA, Max Planck Institute for Social Law and Social Policy.
Markova, E., Yordanova, G., Theodoropoulos, N., Polycarpou, A., Rotkirch, A., Mäki, M., and Vokounová, D. (2019), "Becoming a New SHARE Country," in *SHARE Wave 7 Methodology*, eds. M. Bergmann, A. Scherpenzeel, and A. Börsch-Supan, Munich: MEA, Max Planck Institute for Social Law and Social Policy, pp. 63–79.
McFall, S., Booker, C., Burton, J., and Conolly, A. (2012), "Implementing the Biosocial Component of Understanding Society: Nurse Collection of Biomeasures," Understanding Society Working Paper Series. University of Essex.
Olbrich, L., Kosyakova, Y., and Sakshaug, J. W. (2022), "The Reliability of Adult Self-Reported Height: The Role of Interviewers," *Economics and Human Biology*, 45, 101118.
Olson, K., and Peytchev, A. (2007), "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes," *Public Opinion Quarterly*, 71, 273–286.
O'Muircheartaigh, C., and Campanelli, P. (1998), "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161, 63–77.
Pashazadeh, F., Cernat, A., and Sakshaug, J. W. (2021), "The Effects of Biological Data Collection in Longitudinal Surveys on Subsequent Wave Cooperation," in *Advances in Longitudinal Survey Methodology*, ed. P. Lynn, Hoboken, NJ: Wiley, pp. 100–121.
Petersen, J., and Benzeval, M. (2016), "Untreated Hypertension in the UK Household Population: Who Are Missed by the General Health Checks?," *Preventive Medicine Reports*, 4, 81–86.
R Core Team (2021), *R: A Language and Environment for Statistical Computing*, Vienna, Austria. Available at https://www.R-project.org/. Accessed May 2, 2022.
Raghunathan, T. E. (2006), "Combining Information from Multiple Surveys for Assessing Health Disparities," *Allgemeines Statistisches Archiv*, 90, 515–526.
Roberts, H. C., Denison, H. J., Martin, H. J., Patel, H. P., Syddall, H., Cooper, C., and Aihie Sayer, A. (2011), "A Review of the Measurement of Grip Strength in Clinical and Epidemiological Studies: Towards a Standardized Approach," *Age and Ageing*, 40, 423–429.
Sakkeus, L., Abuladze, L., Kézdi, G., Gál, R., Pita Barros, P., Delerue Matos, A., Mašič, Š. (2013), "Becoming a New SHARE Country," in *SHARE Wave 4*, eds. F. Malter and A. Börsch-Supan, Munich: MEA, Max Planck Institute for Social Law and Social Policy, pp. 11–17.
Sakshaug, J. W., Couper, M. P., and Ofstedal, M. B. (2010), "Characteristics of Physical Measurement Consent in a Population-Based Survey of Older Adults," *Medical Care*, 48, 64–71.
Sakshaug, J. W., Ofstedal, M. B., Guyer, H., and Beebe, T. J. (2015), "The Collection of Biospecimens in Health Surveys," in *Handbook of Health Survey Methods*, ed. T. P. Johnson. Hoboken, NJ: Wiley, pp. 383–419.
Scherpenzeel, A., Axt, K., Bergmann, M., Douhou, S., Oepen, A., Sand, G., and Börsch-Supan, A. (2020), "Collecting Survey Data among the 50+ Population during the COVID-19 Outbreak: The Survey of Health, Ageing and Retirement in Europe (SHARE)," *Survey Research Methods*, 14, 217–221.
Schnell, R., and Kreuter, F. (2005), "Separating Interviewer and Sampling-Point Effects," *Journal of Official Statistics*, 21, 389–410.
SHARE-ERIC (2022), "Publications Based on SHARE Data." Available at Publications (share-eric.eu). Accessed July 28, 2023.
Snijders, T., and Bosker, R. (2012), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London. Sage.
Stan Development Team (2020), "Brief Guide to Stan's Warnings." Available at https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup. Accessed May 2, 2022.
Stan Development Team (2022), "Stan Modeling Language Users Guide and Reference Manual." Available at https://mc-stan.org. Accessed May 2, 2022.
Stomfai, S., Ahrens, W., Bammann, K., Kovács, É., Mårild, S., Michels, N., Moreno, L. A., Pohlabeln, H., Siani, A., Tornaritis, M., Veidebaum, T., and Molnár, D; on behalf of the

IDEFICS Consortium (2011), "Intra- and Inter-Observer Reliability in Anthropometric Measurements in Children," *International Journal of Obesity*, 35, S45–S51.

Sturgis, P., Maslovskaya, O., Durrant, G., and Brunton-Smith, I. (2021), "The Interviewer Contribution to Variability in Response Times in Face-to-Face Interview Surveys," *Journal of Survey Statistics and Methodology*, 9, 701–721.

Ulijaszek, S. J., and Kerr, D. A. (1999), "Anthropometric Measurement Error and the Assessment of Nutritional Status," *British Journal of Nutrition*, 82, 165–177.

Vancampfort, D., Stubbs, B., Firth, J., Smith, L., Swinnen, N., and Koyanagi, A. (2019), "Associations between Handgrip Strength and Mild Cognitive Impairment in Middle-Aged and Olter Adults in Six Low- and Middle-Income Countries," *International Journal of Geriatric Psychiatry*, 34, 609–616.

Vassallo, R., Durrant, G., and Smith, P. (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 (Statistics in Society)*, 180, 531–550.

West, B. T., Conrad, F. G., Kreuter, F., and Mittereder, F. (2018), "Nonresponse and Measurement Error Variance among Interviewers in Standardized and Conversational Interviewing," *Journal of Survey Statistics and Methodology*, 6, 335–359.

West, B. T., Welch, K. B., and Galecki, A. (2015), *Linear Mixed Models: A Practical Guide Using Statistical Software* (2nd ed.), New York: CRC Press.