Survey Research Methods (2022)
Vol. 16, No. 2, pp. 147-163
doi:10.18148/srm/2022.v16i2.7672
European Survey Research Association

Observing interviewer performance in slices or by traces: A comparison of methods to predict interviewers' individual contributions to interviewer variance

Celine Wuyts
Centre for Sociological Research
KU Leuven, Belgium

Geert Loosveldt
Centre for Sociological Research
KU Leuven, Belgium

The interviewing practice of survey interviewers has long been recognized as an important determinant of measurement error in survey data. In the current article, we compare two approaches that can be used to identify interviewers whose task performance might be inadequate and damaging to data quality. The first approach assesses interviewing behavior through the use of audio-recorded interviews. Behavioral assessments capture actual behavior in an interview, but typically rely only on "slices" of observed behavior. The second approach is based on interview time paradata, a type of "trace" data that can easily be aggregated in summary measurements such as average interview speed at the interviewer level. In the current study, we use data from the Dutch-speaking subsample of interviewers employed in two survey rounds of the European Social Survey in Belgium to evaluate how successful the two above approaches are for predicting interviewers' contributions to interviewer variance. The results show that interviewers who deviate from a larger number of standardized interviewing practices in one (early) audio-recorded interview, as well as those who tend to accelerate their interviewing speed over the course of an interview, tend to contribute more to interviewer variance. The two types of performance assessments appear to be independent, additive predictors of interviewers' variance contributions. While statistically significant, the effects are modest in size. The implication for practice is that interviewer monitoring would benefit from well-considered combinations of both behavioral and paradata-based assessments

Keywords: Interviewer effects; interviewer monitoring; survey interviewing; paradata

1 Introduction

Interviewer effects on responses recorded from survey respondents have long been evident and have been extensively studied (e.g. Cannell, 1968; Hyman, 1954; Rice, 1929), but interviewer-related measurement error remains a relevant concern. Numerous studies have reported on the degree of similarity of responses to survey questions obtained by the same interviewer, as commonly measured by the intra-interviewer correlation (Kish, 1962), and many have made attempts to explain these intra-interviewer correlations as well as to estimate them (West & Blom, 2017). Interviewers' performance in the task of questionnaire administration is generally recognized as a crucial factor in their contribution to interviewer-related measurement error.

In the current article, we compare two approaches that can be used to identify at-risk interviewers; that is, those whose task performance might be inadequate and damaging to data quality. The first approach assesses interviewers' interviewing practice observed through audio-recorded interviews. While interviewers would ideally be scored for multiple recorded interviews, and on the basis of detailed behavior coding, both the breadth (number of interviews) and depth (detail of behavior) of assessments are often constrained in practice. We can therefore consider this approach as observing "slices" of interviewer behavior, borrowing the expression from the social-psychological concept of "thin slicing," or judgement based on short observations of behavior (Ambady & Rosenthal, 1992). The second approach uses interview time paradata, which is easily aggregated in summary measurements such as average interview speed at the interviewer level. Interview time stamp data has also been referred to as a type of "trace" data. In the same way that interaction with a physical environment leaves physical traces, interaction with a digital environment leaves digital traces, for example in the form of time stamps. It is possible for these approaches to provide complementary information and to produce different assessments of interviewer performance.

Contact information: Celine Wuyts, Centre for Sociological Research, KU Leuven, Parkstraat 45—bus 3601, 3000 Leuven, Belgium (email: celine.wuyts@kuleuven.be)

The question that we address is whether interviewer assessments based on "slices" of behavior (audio recordings) and on "traces" of behavior (interview time paradata) are associated with interviewers' effects on the responses recorded from survey respondents, and if so, which of the two approaches has the superior predictive power. We evaluate how successful these methods are with regard to predicting individual contributions to interviewer variance, for the Dutch-speaking subsample of interviewers employed in two survey rounds of the European Social Survey in Belgium.

2 Coding and capturing interviewer performance

As sources of information on interviewer performance, behavior coding and interview time paradata represent very different traditions in the practice of interviewer monitoring, and may be contrasted in terms of scope, cost and timeliness, as well as effectiveness. In this section, we briefly review the relevant literature in support of each method of assessment.

2.1 Coding interviewer behavior

Systematic coding of interviewer behavior was initially advanced as an appropriate method for evaluating and monitoring interviewer performance (Cannell et al., 1975). Cannell and colleagues argued that behavior coding from audio recordings produces objective, balanced, and reliable data on relevant aspects of interviewer performance, and is therefore superior to alternative methods such as checking interview data for consistency and completeness (many shortcomings in performance are thereby not detectable) and supervisor judgement in field observation (subjective, and positive aspects of performance are prone to be ignored). Behavioral assessments capture actual interviewer performance in the interview process in a *direct* way, typically in terms of compliance with the adopted (standardized) interviewing protocol. Although coding schemes can be expanded and customized to align with any preferred interviewing style or set of interviewing instructions, behavior coding is inherently well suited to evaluating *standardized* interviewing. Standardized interviewing should ensure that "any differences in the answers can be correctly interpreted as reflecting differences between respondents rather than differences in the process that produced the answer" (Fowler, 1990, p. 14). Because actual interviewer behavior is observed, behavioral assessments allow clear, well-grounded feedback to be provided to interviewers, justified by concrete examples.

A range of coding schemes and strategies have been developed, with varied focuses and aims (see Ongena & Dijkstra, 2006, for an overview). Behavior coding has not only been widely adopted in interviewer monitoring (e.g. Mathiowetz & Cannell, 1980), but has also quickly gained acceptance as a tool for systematic question evaluation and pre-testing (Oksenberg & Kalton, 1991). It has also been used to study the interview process more generally, as well as the causes of interactional problems (e.g. Brenner, 1982; Dijkstra & Ongena, 2006; Loosveldt, 1997). Such methodological investigations into the nature of the interview process tend to favor high levels of detail. Sequential utterances are typically coded from interview transcripts. In practical applications of behavior coding, cost and time considerations—as well as effectiveness in terms of identifying at-risk questions or interviewers—will weigh heavily on the choice of the scheme and coding strategy. Live and recorded coding on the basis of rough-and-ready coding schemes is more economical (Ongena & Dijkstra, 2006), and also yields more timely assessments and feedback than coding based on transcripts. Even at a low level of detail, however, behavior coding remains a time-intensive activity that requires extensive resources. Coded behavior can be assessed early in the data collection period from one or several of an interviewer's first completed interviews (or even one of their training or pilot interviews), so that relevant interviewer information is promptly available.

When used for interviewer monitoring, behavior coding can be useful to the extent that inadequate interviewing practice (which actually affects responses) is reliably detected; that is, to the extent that observed interviewing quality is predictive of measurement quality for individual interviewers (Sharma, 2019). However, empirical support is inconsistent regarding the link between observations of interviewing practice and response quality. On the one hand, there are traditional question wording experiments (e.g. Kalton & Schuman, 1982), which demonstrate that response distributions shift when survey questions are slightly differently worded, formatted, or ordered by design. These results feed concerns that unscripted changes and non-neutral probing by interviewers could be similarly damaging. Leading behavior by interviewers has been shown to threaten response quality (Smit & van der Zouwen, 1997). On the other hand, several studies drawing on external records show that incorrect question reading is on the whole unrelated to the accuracy of recorded responses to factual questions (Belli & Lepkowski, 1996; Dykema, Lepkowski, & Blixt, 1997). This suggests that unscripted changes to question wording may be mostly benign.

Relatedly, a number of studies have examined whether common interviewing practice is related to response quality across survey questions (Schaeffer & Dykema, 2011). These studies tend to disprove the assumption that questions that are frequently read incorrectly result in lower response quality in terms of interviewer effects on response distributions (Groves & Magilavy, 1986; Mangione, Fowler, & Louis, 1992) or in terms of test-retest reliability (Hess, Singer, and Bushery, 1999; see Maitland and Presser, 2016 for a counterexample). Nevertheless, a weak or even nonexistent association between response quality and some interviewer behavior at the question level does not preclude a possible association between response quality and interviewer behavior at the interviewer level. Interviewers who perform their task of interviewing poorly would make changes to well written as well as poorly written questions, and possibly when interviewing "easy" respondents as well as when interviewing respondents who do not fully appreciate their intended role in a standardized survey interview. Questions that are incorrectly read by a large number of interviewers may simply have been poorly written, or be complex or lengthy, tempting all interviewers to remedy their clarity or improve their ease of expression, without overly detracting from measurement quality. We may consequently expect *interviewers* with unstandardized interviewing performance to contribute more to measurement error, even if *questions* characterized by unstandardized interviewing performance may not be more prone to such error.

An early examination of interviewing performance in relation to response quality at the interviewer level was reported by Groves and Magilavy (1986). They expected squared deviations from the overall mean to be larger for interviewers for whom more "incorrect" behavior was observed. Their data did not support the hypothesis, possibly because the interviewers were well trained and monitored, and interviewer effects were typically small. The current study builds on this example and research tradition.

2.2 Interview time paradata

Interview time paradata is available in abundance for many current survey projects. Interview durations were initially recorded by interviewers, but time paradata at any level of detail is increasingly being automatically captured by the software used in computer-assisted data collection (Couper, 1998), and its collection can thus be essentially costless. Time paradata should also be reasonably accurate, especially if event-cued (rather than actively recorded or voice-cued) (West & Sinibaldi, 2013). Time stamps may well be the most common type of paradata on the interview process (Olson & Parkhurst, 2013); however, in contrast to behavioral assessments, which measure interviewer behavior in a direct way, the interview process itself remains opaque when only time stamps are observed. Interview time measurements may represent *indirect* measurements of interviewer behavior, at best. It is in the aggregate, at the interviewer level, that we would expect general systematic behavior to be reflected. In addition, although interview time paradata is collected from the very first completed interview and accumulates over the course of the data collection period, interviewer-level aggregates may only stabilize to the point of capturing systematic deviating interviewing behavior when a relatively large number of interviews have been completed. If interview time measurements are used for interviewer monitoring, this progression over the course of the fieldwork needs to be taken into consideration.

Interview time has often been assumed to be related to response quality (e.g. Cannell, Miller, and Oksenberg, 1981; Hox, 1994; Olson and Peytchev, 2007; and the recent literature review on satisficing by Roberts, Gilbert, Allum, and Eisner, 2019). Time measurements—such as interview duration, interview pace (minutes per question) or speed (questions per minute)—computed from time stamp paradata may even serve as Key Performance Indicators to track the expected costs of data collection and/or flag unusually short or long interviews (Jans, Sirkis, & Morgan, 2013). Survey practitioners are particularly wary of short interviews that result from insufficient effort being applied ("satisficing", Krosnick, 1991) by respondents (e.g. Zhang & Conrad, 2014) and/or interviewers (e.g. Japec, 2007). However, the relationships between response times and response quality (at the question level), between interview length and interview quality (at the respondent level), and between interviewer-level time measurements and interviewer task performance are not unequivocal (Olson & Parkhurst, 2013).

The notion that interview duration (or other time-related measurements) can serve as an indicator of interviewer performance is certainly not new (e.g. Steinkamp 1964). The idea that this duration captures relevant interviewer behavior derives from early anecdotal evidence of interviewers who rush their work, and is supported by the consistent finding of strong variability in interview duration between interviewers (Hox, 1994; Kirchner & Olson, 2017; Loosveldt & Beullens, 2013a; Olson & Peytchev, 2007; Vandenplas, Loosveldt, Beullens, & Denies, 2018). In the context of standardized interviewing, one would expect the interview duration to be determined by the length, format, and complexity of the questionnaire, and to vary around the mean interview duration in accordance with the respondents' cognitive capacity and motivation (Loosveldt & Beullens, 2013a). Any impact of the interviewer should in principle be minimal if interviewers "apply the same basic task rules" (Loosveldt & Beullens, 2013b). However, it has been observed that in practice "some interviewers read very quickly, others speak slowly and distinctly. Some give respondents time to consider their answers after they have given them, while others begin asking the next question as soon as one answer has been given" (Fowler, 1990). Interviewers can affect the character and pace of their interactions with respondents not only by their pace of (correct) question reading and of progressing through the questionnaire, but also by, for example, their inclination to rephrase questions, to suggest answers or even skip questions altogether, to probe incomplete answers, and to digress into discussions with respondents. Such differences in interviewing practice across interviewers are expected to leave traces in interview time paradata, as well as to affect responses.

Metrics based on time stamp paradata can easily be tracked during data collection, but using such metrics for monitoring interviewers only makes sense if these metrics actually capture interviewers' task performance and are thereby sufficiently predictive of measurement quality. Examples of interview time measurements used either as correlates of improper interviewing behavior (as observed), or as predictors of interviewer-related measurement error (as estimated from the data) are sparse. Olson, Smyth, and Kirchner (2020) recently contributed a powerful demonstration of the first type, showing the association between interview time measurements and interviewing behavior in telephone surveys. They observed that wording deviations simultaneously increased the question reading time (in seconds) and the reading speed (in words per second). Other recent empirical studies support the link between interview time measurements and indicators of measurement quality at the interviewer level (e.g. Vandenplas et al., 2018). A recent study by Vandenplas, Beullens, and Loosveldt (2019) demonstrated that both slow and fast interviews show elevated interviewer effects on the recorded answers to survey questions (compared with interviews completed at a moderate speed). These findings show that some aspects of interviewing practices may indeed be reflected in the interview time and accordingly justify the use of relevant paradata to flag irregular interview interactions.

2.3 Addressing the gaps within and between interviewing monitoring traditions

In general, there is much more literature *describing* interviewer effects than there is literature successfully *explaining* interviewer effects (Blom & Korbmacher, 2013). And while interviewers' interviewing practice is recognized as a crucial factor of measurement quality, very few studies have empirically investigated the link between interviewer performance in the interview and measurement quality. This is in part because interviewer performance in the interview can only be measured either indirectly (e.g. using time stamp data) or at a high cost (transcribing and coding of audio-recorded interviews).

Not only do few studies within either the behavior coding tradition or the paradata tradition attempt to explain interviewer-related measurement error, the two research traditions also remain distinct. The present study addresses the empirical question concerning interviewer effects on measurement quality, which lingers in both the behavior coding tradition and in the paradata tradition, in a single setup. We thereby aim to bridge the gap between the insights of "slices" (audio recordings) and "traces" (interview time paradata) of interviewer performance, and to inform best practices in interviewer monitoring.

3 Data and methods

We draw on data from two rounds of the European Social Survey (ESS) in Belgium (European Social Survey, 2012, 2014). In both Round 6 (2012–2013) and Round 7 (2014–2015), approximately 3,200 sample units were allocated to about 150 interviewers overall, and about 1,800 interviews were administered (with a response rate of 59% in Round 6 and 57% in Round 7). The fieldwork conditions for the two rounds were not identical, but most of the design remained constant, in terms of the target population (resident population aged 15 or above), sampling frame (National Register), mode (personal interviews), contact protocol (advance letter, four personal visits spread by day of the week and time of day), survey agency (TNS Dimarso), etc. (see European Social Survey (2016a, 2016b) for further details). The interviewers were all given training in standardized interviewing by the survey agency and close to 80% had more than two years' experience working as an interviewer (Wuyts & Loosveldt, 2017). All attended a half-day project briefing, during which the principles and practice of standardized interviewing were reviewed and the specifics of the ESS questionnaire were discussed. For most of the Dutch-speaking interviewers who worked for the ESS in Round 6 (hereafter referred to as the ESS6_BEDUT) and Round 7 (hereafter referred to as the ESS7_BEDUT), adherence to the standardized interviewing protocol was evaluated and coded from audio-recorded interviews. Although in some cases multiple audio recordings were captured, only a single audio recording was evaluated unless this assessment flagged issues too severe to let pass without reconsideration.

3.1 Analytic sample

Interviewers. An audio recording was obtained and coded for 88 Dutch-speaking interviewers (90%) in the ESS6_BEDUT and the 80 Dutch-speaking interviewers (94%) in the ESS7_BEDUT. As the same survey agency was contracted, the two groups of interviewers partially overlap. About half of them (44 interviewers) worked in both survey rounds.

Respondents. The selected interviewers administered 1,044 and 973 interviews in respectively the ESS6_BEDUT and the ESS7_BEDUT.

Before describing the interviewer-level explanatory variables and modeling approach, the following subsections outline how questionnaire items were selected and a proxy measurement of interviewers' contributions to measurement error was derived.

3.2 Selection of questionnaire items

A random intercept model was estimated for each item that was measured on at least a four-point ordinal scale in the ESS questionnaire, excluding the sociodemographic module F. The model is specified as $y^{st}_{ij} = \beta x^{T}_{ij} + u_j + \varepsilon_{ij}$, with y^{st}_{ij} the (standardized) response recorded for respondent i interviewed by interviewer j, u_j the interviewer random intercept with $u_j \sim \mathcal{N}(0, \sigma^2_u)$, and ε_{ij} the residual error term with

$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. x_{ij}^T represents a row vector of respondent-level control variables (respondent age, gender, education level, type of domicile—i.e. big city, suburbs, town, country village, or countryside—and NUTS2 region) with β a corresponding vector of coefficients. Respondent age is a numeric variable. Respondent gender is represented by a single dummy variable. Education level was recoded into three broad categories ("up to lower secondary education", "upper secondary education or advanced vocational education", and "tertiary education"). Degree of urbanization was measured with five response categories ("big city", "suburbs or outskirts of big city", "town or small city", "country village", and "farm or home in countryside"). There are 11 statistical regions at the NUTS2 level in Belgium, corresponding to the ten provinces and the Brussels-Capital region. The Dutch-speaking interviewers, to which the present analysis is restricted, worked almost exclusively in the five Flemish provinces.

The conditional intra-interviewer correlation, $\rho_{\text{int}} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\varepsilon^2}$, expresses the proportion of the total variance in the responses recorded for the given item that is due to systematic differences between the interviewers, after controlling for possible differences in their respondent sample composition in terms of age, gender, education level, urbanicity, and region.

The specified random intercept model with the respondent characteristics was initially estimated for 139 items in the ESS6_BEDUT and 110 items in the ESS7_BEDUT. The between-interviewer variance is small for most items. In order to avoid items with negligible interviewer variance distorting the results, items for which the estimated interviewer variance component is smaller than 3% of total variance were dropped. This minimal selection criterion left 39 items (28%) for the ESS6_BEDUT and 25 items (23%) for the ESS7_BEDUT, of which three items were repeated in both survey rounds. The intra-interviewer correlation estimates for the 61 selected questionnaire items, ordered from largest to smallest (on average if repeated in both survey rounds), are presented in the Appendix (Table A1). For the selected items, the intra-interviewer correlation estimates range between 0.03 and 0.10 (Mean = 0.05, Std. Dev. = 0.02). These estimates suggest that interviewer effects are weak to moderate in Belgium, compared with other countries in the European Social Survey (Beullens & Loosveldt, 2016).

3.3 Interviewer-related measurement error and attributing measurement error to interviewers

Interviewer-related measurement error for a given survey question is commonly regarded and quantified in a narrow sense, namely in terms of interviewer variance. After all, a direct evaluation of response accuracy across interviewers would require respondents' "true values" to be available for comparison with their recorded responses. External records containing such "true values" are normally not available even for factual questions, and are inconceivable for the vast array of survey questions on beliefs and attitudes. The recorded responses to a survey question therefore generally cannot be assessed in terms of response accuracy (e.g. Sudman & Bradburn, 1974).

Interviewer variance is an aspect of measurement error that is quantifiable, and not only a cause for concern per se, but also a signal for the risk of interviewer error more broadly. It is composed of all interviewers' deviations from the overall mean response, each of which is the product of an interviewer's individual combination of physical characteristics, interviewing style, mannerisms, and personal idiosyncrasies (Hagenaars & Heinen, 1982). However, interviewer variance only captures correlated response effects and it is important to keep in mind that interviewers can also affect responses erratically (e.g. by making different question reading mistakes for each of their respondents) and/or uniformly (e.g. by *all* making the same question reading mistake for *all* respondents). Measurement errors of these types would be caused by interviewers, but cannot be captured by interviewer variance components because they do not result in systematic differences between interviewers.

Variance components are relatively easy to compute without requiring any external data sources. We therefore consider interviewers' individual contributions to between-interviewer variance as an approximate measurement of interviewers' contributions to measurement error more broadly. For each of the selected questionnaire items, we extract the difference between the interviewer-specific intercepts and the overall intercept from the estimated random intercept model. These are the conditional means of the interviewer random effects, $\widehat{u}_j$. The larger the discrepancy for a particular interviewer j (in either direction and after accounting for the composition of respondent groups; that is, the larger $|\widehat{u}_j|$, the further away the interviewer-specific mean response is from the overall mean response, and consequently the more the interviewer tends to contribute to the interviewer variance for that item. We thereby obtain a series of measurement error contribution estimates for each interviewer, with the number of estimates equal to the number of selected questionnaire items in each round (39 estimates for the ESS6_BEDUT interviewers, 25 estimates for the ESS7_BEDUT interviewers, and 64 estimates for interviewers who worked in both survey rounds).

Figure 1 illustrates the interviewers' estimated interviewer variance contributions. Each row of dots corresponds to one interviewer, and each dot corresponds to one variance contribution estimate for that interviewer: one estimate for each item in the questionnaire. Some interviewers have variance contribution estimates for both the ESS6_BEDUT and the ESS7_BEDUT, while others only have estimates for one survey round. The average variance contribution for each interviewer is indicated by an asterisk, and the interviewers are ordered accordingly. Those shown near the bottom of the plot consistently contribute little to interviewer variance across survey items, while those near the top of the plot contribute a great deal to interviewer variance, at least for some survey items. The contributions to interviewer variance appear as a gradient across interviewers, without clearly delineable groups. There is one interviewer with an uncommonly large average variance contribution, but none of the interviewers consistently contributes large proportions of interviewer variance across all survey items.

3.4 Interviewers' observed deviation from standardized interviewing protocol

The Belgian interviewers were required to make an audio recording of at least one of their first three interviews, in order to ensure that interviewers who deviated strongly from the standardized interviewing protocol (as endorsed by the European Social Survey) could be identified early in the fieldwork. Interviewers who were unable to record one of their first three interviews had to record one in their first assigned set (usually 18 sample units). Each interviewer was given written feedback from the fieldwork supervisors on the basis of an assessment of the recorded interview. Interviewers for whom interviewing practice was judged as likely to be damaging to data quality were excluded from receiving additional assignments and were removed from the project.

The evaluation of the Dutch-speaking interviewers' recordings was based on a checklist of 29 items. Not every form of possible interviewer behavior was covered; only those deemed relevant to interviewing quality, and these were coded only in general, at the level of the interview as a whole. The checklist covers interviewer behavior with regard to reading questions (e.g. reading all the questions), clarifying the respondents' task (e.g. referring to showcards), objectivity (e.g. not leading or steering), manner of speaking (e.g. reading questions clearly), and the interaction with respondents (e.g. allowing sufficient time). Each checklist criterion was evaluated dichotomously ("OK" or "not OK"). An "OK" code was assigned when in general the basic standardized interviewing instruction was properly followed during the interview. Small errors or deviations were ignored. Table 1 shows the full set of checklist criteria, and the relative number of audio-recorded interviews (therefore interviewers) for which a deviation was observed.

With "OK" coded as 0 and "not OK" coded as 1, an overall count of deviations from the standardized interviewing protocol is calculated as the sum of the checklist items. This overall deviation count thus captures the number of different deviating behaviors observed in one audio-recorded interview (therefore, for one interviewer). Although the checklist is applied to only one audio-recorded interview per interviewer and each checklist item by itself could be coded imperfectly, the overall count can be assumed to be a reasonable indicator of interviewers' tendency to depart from standardized interviewing. Since interviewers are aware of the audio recording process, they may be encouraged to act more in line with what they know is expected of them than they would during unrecorded interviews. We assume that the observation effect influences all interviewers to a similar degree. Interviewers who tend to be less compliant with the protocol would thus achieve higher scores on average than interviewers who strongly adhere to it (and continue to do so).

The deviation count (Figure 2) ranges between zero (respectively 13% and 9% of interviewers in the ESS6_BEDUT and the ESS7_BEDUT) and 16. Although only a small number of interviewers had a perfect score of zero, most scored moderately well, with on average four deviating behaviors. Particularly common—even among interviewers who closely but not perfectly adhered to the interviewing protocol—are reading text that *should not* be read out, or not reading out text that *should* be. Specifically, this refers to reading out response options when they should not be read (there is a showcard), not reading out all the response options when they should be read (there is no showcard), reading interviewer instructions out loud, and not reading out questions completely.

In order to gauge the types of deviating behavior most predictive of interviewers' interviewer variance contributions, we furthermore evaluated three component deviation indicators. We constructed binary indicators to represent (1) reading questions as written (checks 3, 4, 5, 15, 23, and 25), (2) remaining neutral (checks 7, 11, 12, 18, 19, 20, and 29), and (3) maintaining an appropriate pace (checks 21 and 26). Each indicator equals 1 if the evaluation of the audio recording produced at least one flag among the respective checklist items. At least one type of deviation in reading out questions as written was flagged for 50% and 73% of interviewers, in remaining neutral for 55% and 54%, and in keeping an appropriate pace for 6% and 17% in the ESS6_BEDUT and the ESS7_BEDUT respectively.

3.5 Interviewers' interview speed

Since Round 5, the ESS has required interview and questionnaire module time stamp paradata to be collected in all participating countries. In most countries, this type of data is captured by timers implemented in the CAPI program. Interview and questionnaire module durations and related measurements such as interview speed (the number of questionnaire items administered per minute) can be derived from this timer paradata for each individual respondent. In the current study, we focus on two measurements of interview speed: the overall speed of the interview and the change in speed (acceleration) over the questionnaire. Interview speed is calculated as the number of (applicable) questionnaire items divided by the interview duration (in minutes). Acceleration is

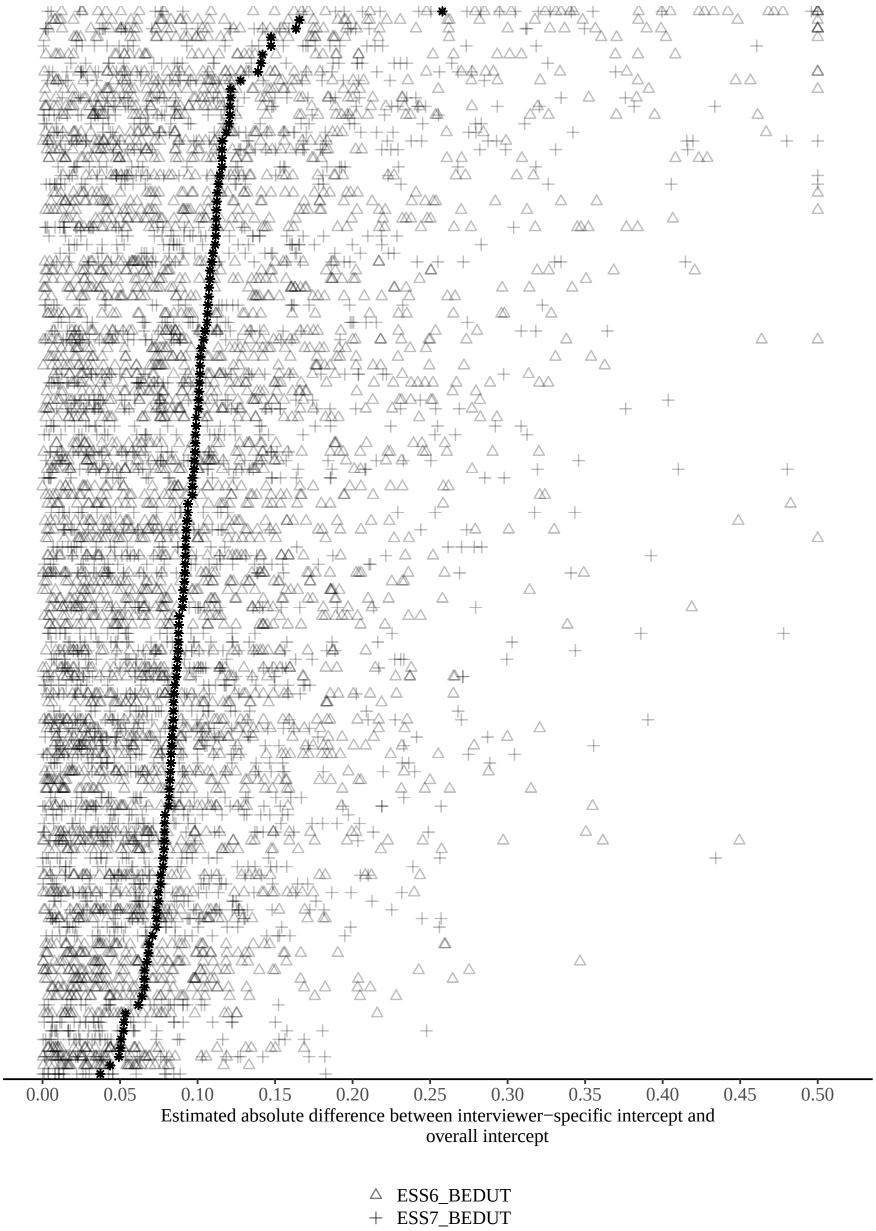

Figure 1. Interviewer-specific contributions to interviewer variance. Note: Each row of dots corresponds to one interviewer. Variance contribution estimates exceeding 0.5 in absolute value are trimmed at that value ($N = 19$ out of 5,430 estimates, or $< 0.04\%$). The average of each interviewer's variance contribution estimates is indicated by an asterisk (*), and interviewers are ordered accordingly from top to bottom.

expressed as the slope parameter estimate of a linear regression for module speed (weighted by the number of applicable items in the module) as a function of module position. Both measurements are derived over the main questionnaire modules A to E (excluding the sociodemographic module F). From these respondent-level interview speed measurements, we extract two corresponding interviewer characteristics. We take the average over all completed interviews at the interviewer level to observe the interviewers' average interview speed and the interviewers' average acceleration over the questionnaire.

We have previously noted that one relevant feature of

Table 1
Checklist of standardized interviewing protocol and deviation frequency

		Interviewers for whom deviation was observed (in %)	
	Check	ESS6_BEDUT	ESS7_BEDUT
1	Reads introduction	13	11
2	Speaks the language of the interview fluently	0	8
3	Reads questions completely	36	51
4	Does not add anything to questions	16	33
5	Reads all applicable questions	5	9
6	Reads introductory sentences before questions	39	28
7	Repeats questions in case of irrelevant or unclear answer,	-	-
	or when requested by the respondent	9	20
8	Does not read out interviewer instructions	7	26
9	Reads references to showcards	6	15
10	Does not read out showcards	39	34
11	Asks for additional explanation if the answer is not	-	-
	one of the available options	36	29
12	Does not provide example answers	27	18
13	Does not read out Refusal, Don't know, and Other	1	1
14	Does not read out additional options within brackets	2	3
15	Reads out all options in case no showcard available	20	34
16	Probes at least once in the case of Refusal or Don't know	5	3
17	Probes at least once in the case of all-that-apply questions	31	15
18	Is not leading or steering	30	23
19	Does not give his/her opinion	3	4
20	Asks the respondent to interpret the question him/herself in	-	-
	case they ask for explanation	7	21
21	Does not read out questions too slowly or too quickly	1	16
22	Does not read out questions too loudly or too quietly	0	1
23	Reads out questions clearly	5	6
24	Is agreeable to listen to	3	6
25	Reads out all the questions in the same way, without apology	6	5
26	Gives respondent sufficient time to answer	6	6
27	Gives short confirmations	0	0
28	Is friendly and interested	0	3
29	Does not give value judgements, approvals, disapprovals	1	1

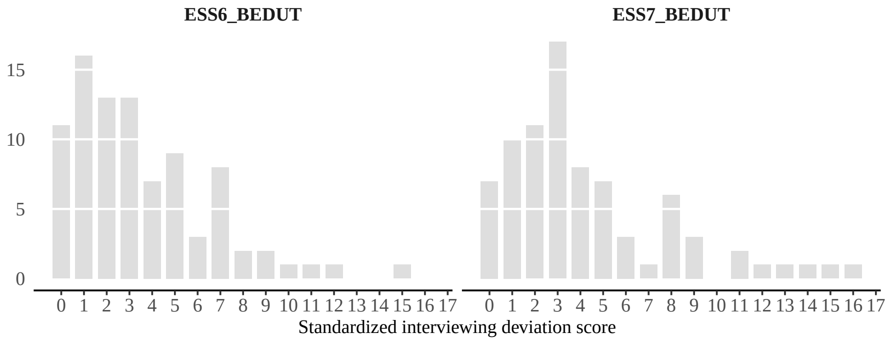

Figure 2. Distribution of the interviewers' standardized interviewing deviation count

the evaluation in actual practice is the timeliness of the interviewer-level information. As interview time tends to decrease over the first few completed interviews (Olson & Peytchev, 2007) and is subject to respondent-to-respondent variability, interviewer-level measurements may take a while to stabilize. We therefore also explored how many interviews have to be completed and taken into account before interviewer-level measurements derived from time stamp paradata become sufficiently stable for interviewer performance monitoring. To this end, the interviewer-level speed measurements were also computed by taking into account an increasing number of completed interviews. After one completed interview, the interviewer-level measurements simply equal those of the first respondent of each interviewer; after two completed interviews, the interviewer-level measurements are calculated as the averages over the first two respondents of each interviewer, and so forth. Figure 3 shows the intermediate interview speed measurements for three interviewers by way of illustration. In the main analysis we use interviewers' average interview speed and average acceleration over the questionnaire, computed over all completed interviewers, irrespective of the number of interviews completed. Building on the main analysis, we make use of the intermediate interview speed measurements in order to assess after how many completed interviews the interviewers' average interview speed and average acceleration over the questionnaire (up to that point) may be used to predict interviewers' contributions to interviewer variance.

3.6 Modeling approach

The interviewers' interviewer variance contribution estimates $|\hat{u}_{jkr}|$ (see Section 3.3), derived from the basic model fits for multiple survey questions, are themselves modeled in a cross-classified multilevel model. The variance contribution estimates are nested within survey rounds $r \in (6, 7)$, survey items $k = 1 \dots K$, and interviewers $j = 1 \dots J$. The baseline model is specified as:

$$|\hat{u}_{jkr}| = \gamma_0 + \gamma_1 R_{jkr} + \gamma_2 m_{jr} + v_j + s_k + \varepsilon_{jkr} \quad \text{(Model 0)}$$

with v_j and s_k an interviewer-level and an item-level random intercept with $v_j \sim \mathcal{N}(0, \sigma_v^2)$ and $s_k \sim \mathcal{N}(0, \sigma_s^2)$, respectively. A dummy variable with the ESS7_BEDUT data identified by $R_{jkr} = 1$ is included to account for any systematic difference between the two survey rounds. The number of interviews m_{jr} conducted by each interviewer j is included as interviewer-level control variable to account for the expected increase in average interview speed and average acceleration over the questionnaire as interviewers complete a greater number of interviews (Olson & Peytchev, 2007).

Model A and B add, respectively, the interviewer-level standardized interviewing deviation count, and the interviewer-level average interview speed and average acceleration over the questionnaire. Model A* is an alternative specification with three binary indicators for specific types of deviating behavior: (1) not reading questions as written, (2) not remaining neutral, and (3) not maintaining an appropriate pace. Model C combines the interviewer-level standardized interviewing deviation count, average interview speed, and average acceleration over the questionnaire in one model. Model C* is an alternative specification with the three binary indicators of deviating behavior replacing the total deviation count. All of the model specifications are of the form

$$|\hat{u}_{jkr}| = \gamma_0 + \gamma_1 R_{jkr} + \gamma_2 m_{jr} + \boldsymbol{\gamma} \mathbf{X}_{jr} + v_j + s_k + \varepsilon_{jkr}$$

with $\mathbf{X}_{jr}$ being the vector of independend variables and $\boldsymbol{\gamma}$ their regession coefficients. Table 2 gives an overview of the independent variables included in $\mathbf{X}_{jr}$.

Table 2
Model specifications

Model	Interviewer-level explanatory variables
A	Standardized interviewing deviation count
A*	Indicator for not reading questions as written Indicator for not remaining neutral Indicator for not maintaining an appropriate pace
B	Average interview speed Average acceleration over the questionnaire
C	Standardized interviewing deviation count Average interview speed Average acceleration over the questionnaire
C*	Indicator for not reading questions as written Indicator for not remaining neutral Indicator for not maintaining an appropriate pace Average interview speed Average acceleration over the questionnaire

4 Results

On the basis of the Model 0 estimates, and thus after controlling for the number of completed interviews per interviewer, the variance in interviewers' individual variance contributions can be decomposed into the proportion that is due to systematic differences between interviewers (5.8%), the proportion due to systematic differences between survey questions (7.2%), and the residual part.

Table 3 presents the standardized, fixed parameter estimates for the effects of the interviewer-level measurements of interest, the standardized interviewing deviation count (Model A), and the two interview speed measurements (Model B) on interviewers' contributions to interviewer variance. Both models have an improved fit compared with

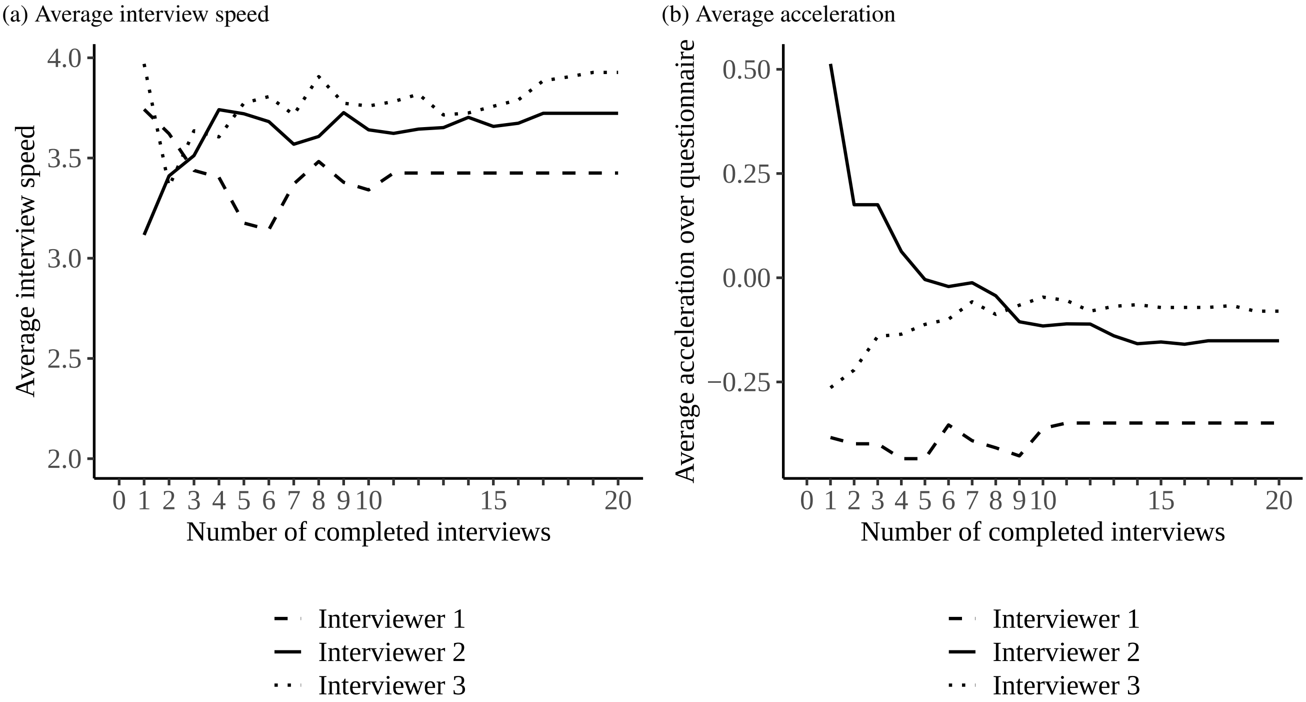

Figure 3. Examples for speed measurement over the questionnaire computed over the first x number of completed interviews

the baseline Model 0, which accounts only for the cross-classified structure, the difference between the two survey rounds, and the number of completed interviews per interviewer (Model A: LR = 7.37, p = 0.0066; Model B: LR = 19.29, $p < 0.0001$).

The estimates for Model A indicate that interviewers for whom a larger number of different deviating behaviors are observed in the audio-recorded interview do tend to contribute more to interviewer variance. The estimated effect of this higher occurrence is, however, modest. It is useful to interpret the effect size in terms of how a shift in an interviewer's interviewing deviation score would translate into a shift in the empirical distribution of variance contributions. For example, the standardized coefficient of 0.0547 means that for an interviewer, an increase from zero to ten deviating behaviors (which corresponds to about three standard deviations in the interviewing deviation score) is expected to result in a change of only about 0.16 standard deviations in terms of variance contributions. We also tested the interaction between the standardized interviewing deviation count and the number of completed interviews, and find that the effect of deviating behaviors (measured early in the fieldwork) on interviewers' contributions to interviewer variance is all the greater for interviewers who completed a larger number of interviews (results not shown).

The alternative specification of Model A* is informative with regard to the types of deviating behavior that are most predictive. We observe that both interviewers who are flagged for non-neutrality and those who are flagged for inadequate interview pace tend to contribute more to interviewer variance than interviewers who are not flagged for such behaviors. However, interviewers who are flagged for incorrect question reading do not appear to contribute any more to interviewer variance than those who are not.

The estimates for Model B indicate that interviewers' interview speed is not (linearly) associated with their contribution to interviewer variance. We also tested an alternative specification with quadratic terms (results not shown), but no improvement in model fit was thereby achieved. Interviewers with a higher acceleration (or lower deceleration) over the questionnaire, however, do tend to contribute more to interviewer variance. The estimated effect of acceleration in interviewing over the questionnaire is modest, but is somewhat more pronounced than the effect of the standardized interviewing deviation score. The standardized coefficient of 0.1073 means that compared with a consistent interview speed over the questionnaire, a substantial acceleration of three standard deviations (an acceleration of 0.6) is expected to result in an interviewer increasing by about 0.32 standard deviations in terms of variance contributions. We also tested the interaction between the two interviewer-level interview speed measurements and the number of completed interviews (results not shown). We find that for interviewers who completed a small number of interviews, average interview speed has an effect on interviewers' contributions to interviewer variance over and above the effect of average acceleration over the questionnaire, but the effect of interview speed shrinks to zero for interviewers who completed larger number of interviews. The effect of acceleration over the questionnaire is robust to the interviewers' workloads.

In this analysis, we computed the interviewer-level interview speed measurements (interviewers' average interview speed and average acceleration over the questionnaire) over all the completed interviews. As previously noted, however, there is a possibility that interviewer-level interview speed measurements may only stabilize—and thereby become informative regarding interviewers' usual interviewing practice—after a sufficiently large number of interviews have been completed. Intermediate interview speed measurements (computed over the first x interviews) should be more useful for interviewer monitoring. In order to assess when the intermediate interview speed measurements become sufficiently stable to predict interviewers' contributions to interviewer variance, we subsequently refitted Model B with the interview speed measurements computed after the first completed interview, after the first two completed interviews, etc. In order to ensure strict comparability across subsequent models, this analysis was restricted to the 67 interviewers in the ESS6_BEDUT and the 53 interviewers in the ESS7_BEDUT (76% and 66%, respectively) who completed at least eight interviews. The conclusion, however, is robust regarding the choice of this cutoff point. Figure 5 shows how the model fit (expressed by the difference in AIC compared with Model 0 estimated for the same group of interviewers) improves as an increasing number of interviews is taken into account. The model attains its maximal predictive power once the interviewer-level interview speed measurements are computed as averages over the first four interviews. At this point, the parameter estimates closely approximate the Model B estimates with the interview speed measurements computed at the end of the fieldwork—that is, over all the interviews completed by a given interviewer – and the model fit is very similar. We therefore retain the interview speed measurements (interviewers' average interview speed and average acceleration over the questionnaire) as computed over all completed interviews in the analysis.

Although both Model A and Model B have an improved fit over Model 0 and show statistically significant effects for relevant interviewer-level explanatory variables, they do not have overwhelming explanatory power. The interviewer variance component, relative to total variance, is reduced by only fractions of a percentage point for Model A and half a percentage point (from 5.8% to 5.3%) for Model B. Model B is slightly superior to Model A in terms of model fit (LR = 11.92, p = 0.0006), but both models leave much unexplained.

The combined Model C has a better fit than Model A (LR = 19.48, p < 0.0001) or Model B (LR = 7.57, p = 0.0059). The fixed parameter estimates are not appreciably altered in Model C and the standardized interviewing deviation count and acceleration over the questionnaire appear to be independent, additive predictors of interviewers' variance contributions. A similar conclusion holds for the alternative specification Model C*, which demonstrates the independent effects of being flagged for inadequate interview pace and acceleration over the questionnaire.

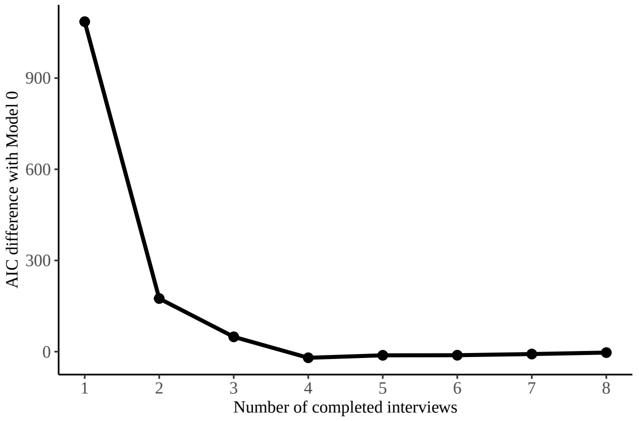

Figure 4. Model fit for Model B refitted with intermediate interview speed measurements

5 Conclusions and discussion

While it is reasonable to assume that interviewer-related measurement error is driven by interviewers' imperfect adherence to the standardized interviewing protocol, and interviewer performance monitoring is common in survey practice, there is only limited research in which the impact of survey interviewers' task performance measurements on aspects of data quality has been empirically investigated. The current study addresses this gap and thereby aims to support the practice of interviewer monitoring in survey research. We compared two approaches for observing interviewers' interviewing task performance: in slices (single audio-recorded interviews) and by its traces in survey paradata (time stamp data). We assessed the predictive power of interviewer-level measurements derived from these sources (respectively, a standardized interviewing deviation count, and average interview speed and average acceleration over the questionnaire) with regard to interviewers' contributions to interviewer variance in a series of questionnaire items.

Drawing on survey data and paradata for Dutch-speaking interviewers in the European Social Survey in Belgium, our results corroborate the presumed association between interviewing task performance and interviewers' variance contributions. We observed that interviewers who score higher on an overall count of different deviating behaviors—derived

Table 3

Standardized coefficients for models explaining interviewers' contributions to interviewer variance

	Model 0		Model A		Model A*		Model B		Model C		Model C*	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Standardized interviewing deviation count	-	-	0.055**	0.020	-	-	-	-	0.055**	0.020	-	-
At least one flag for incorrect question reading	-	-	-	-	−0.031	0.020	-	-	-	-	−0.037	0.021
At least one flag for deviation from neutrality	-	-	-	-	0.046*	0.020	-	-	-	-	0.051*	0.020
At least one flag for inadequate interview pace	-	-	-	-	0.046*	0.019	-	-	-	-	0.046*	0.019
Average interview speed	-	-	-	-	-	-	0.003	0.025	0.014	0.025	−0.000	0.026
Average acceleration over questionnaire	-	-	-	-	-	-	0.107***	0.024	0.107***	0.024	0.113***	0.025
Number of interviews completed by interviewer	0.169***	0.020	0.177***	0.020	0.161***	0.021	0.161***	0.020	0.168***	0.020	0.153***	0.021
Round 7	−0.006	0.028	−0.009	0.028	−0.002	0.029	0.062	0.035	0.065	0.035	0.070*	0.035
Var(v_j)	0.000418		0.000414		0.000445		0.000382		0.000387		0.000413	
Var(s_k)	0.000515		0.000515		0.000516		0.000515		0.000516		0.000516	
Var(ϵ_{jkr})	0.006265		0.006259		0.006248		0.006254		0.006245		0.006234	
Number of observations	5,430		5,430		5,430		5,430		5,430		5,430	
Number of interviewers	124		124		124		124		124		124	
Number of questionnaire items	61		61		61		61		61		61	
AIC	-11,834		-11,840		-11,840		-11,850		-11,855		-11.857	

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

from a simple checklist of standardized interviewing rules applied to one (early) audio-recorded interview—contributed more to interviewer variance related to questionnaire items, as observed at the end of the data collection. We also found that while average interview speed has little predictive power, interviewers who accelerate the interview speed more over the questionnaire, also contributed more to interviewer variance. To assess interviewers' systematic tendencies in responses to survey questions, their average acceleration over the questionnaire was shown to be at least as predictive as the standardized interviewing deviation count determined on the basis of their audio-recorded interviews. Moreover, the interview speed measurements computed after about four completed interviews were completed were already maximally predictive of the extent to which individual interviewers ultimately contributed to interviewer variance. The results show that interviewers who deviate more from the standardized interviewing protocol, as well as those who tend to accelerate their interviewing speed over the course of an interview, tend to employ a more damaging interviewing practice and to contribute more to at least this component of measurement error. The two types of performance assessments appear to be independent, additive predictors of interviewers' variance contributions. We also take note that being flagged, on the basis of a recorded interview, for deviating from neutrality or for an inadequate interview pace was a statistically significant predictor of interviewer variance contributions in the present study, while being flagged for incorrect question reading appeared unrelated.

Although statistically significant, the effects are modest in size, suggesting that definite conclusions about interviewers' interviewing practice cannot be drawn from either type of assessment or both combined. A possible explanation for the low predictive power of the task performance measurements is that the survey under investigation is of reasonably high quality, with an experienced and well-trained interviewer workforce, close interviewer monitoring, and apparently relatively weak interviewer effects. The effects may also be attenuated by the unreliability of the performance measurements. It is evident that interview time measurements ("traces" of behavior) are a few steps removed from actual interviewer behavior. Drawing on single audio recordings, captured at the beginning of data collection, we are able to measure actual interviewer behavior, but these one-off assessments ("slices" of behavior) may not be sufficiently representative of the interviewers' behavior throughout subsequent interviews. Having at least one more audio recording from each interviewer would provide a sense of how reliable a count of deviations from the standardized interviewing protocol is. However, audio recordings are burdensome and costly to collect and process. At least until automatic transcription and error recognition becomes a viable option, the added value of additional audio recordings has to be carefully weighed against their cost.

The results of this study are promising, but in need of replication. We focused on interviewers within a single country (Belgium) and of a single language group (Dutch), for which we evaluated and coded adherence to the standardized interviewing protocol on the basis of audio-recorded interviews in the European Social Survey. We would expect to observe stronger effects if interviewers are less carefully recruited and selected, and less adequately trained and monitored. The relative predictive power of the two types of interviewer performance assessments may also be different in other countries and for other interviewer staff. Additional research on the benefits and costs of these and other interviewer monitoring practices would help survey managers to make informed decisions to maintain or improve survey quality within budgetary limits.

These findings suggest that interview time paradata may be very useful for screening interviewers for closer monitoring during the fieldwork. Audio recordings may be appropriate with regard to removing or retraining interviewers with the worst interviewing practice early during the data collection. Given their high cost, however, resources may be more efficiently allocated when comprehensive assessments of audio-recorded interviews are targeted toward at-risk interviewers identified on the basis of the associated interview paradata. In the survey context under study, we may advise prioritizing assessment of one or multiple audio-recorded interviews conducted by interviewers with the strongest acceleration in interview speed over the questionnaire. The first four complete interviews should be sufficient to compute this. Behavioral assessments may also need to be focused much more on reacting behavior than on question-reading behavior.

Another type of "trace" data of interviewer performance is keystroke data (Olson & Parkhurst, 2013). Keystroke data contains a record of all input from the keyboard, mouse and/or touchscreen. Like interview time stamp data, it is essentially costless to collect, and keystroke data is even more information rich. Keystroke data capturing interviewer actions like skips, changes, and errors may contribute important information about irregular interview interactions. Aggregated to the interviewer level, keystroke-based measures may be used to flag interviewers' suspicious interviewing practice. This type of paradata was not available for analysis in the present study, but would be a rich area for future research.

It is also important to reiterate the point that interviewer variance constitutes only one component of interviewer-related measurement error. It is quantifiable and easily derived from survey data alone, and therefore very convenient to use for empirical investigations into sources of interviewer-related measurement error. However, measurement error resulting from erratically inadequate task performance (where no *systematic* differences *between* interviewers can be observed) are not captured by individual interviewers' contributions to interviewer variance. Our operationalization of interviewers' contributions to measurement error entails an important limitation of this study. A more extensive assessment of interviewer-related measurement error necessitates external data sources, which are rarely available and anyhow absent for attitudinal questions. Investigations into interviewer variance contributions as exemplified in this article are more widely feasible and may help guide further research.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, *111*(2), 256.

Belli, R. F., & Lepkowski, J. M. (1996). Behavior of survey actors and the accuracy of response. In *Health survey research methods: Conference proceedings* (pp. 69–74). Citeseer.

Beullens, K., & Loosveldt, G. (2016). Interviewer effects in the European Social Survey. *Survey Research Methods*, *Vol 10*, No 2 (2016). doi:10.18148/SRM/2016.V10I2.6261

Blom, A. G., & Korbmacher, J. M. (2013). Measuring interviewer characteristics pertinent to social surveys: A conceptual framework. *Survey Methods: Insights from the Field*, 16.

Brenner, M. (1982). Response effects of 'role-restricted' characteristics of the interviewer. *Response behavior in the survey interview*, 131–165.

Cannell, C. F. (1968). *The influence of interviewer and respondent: Psychological and behavioral variables on the reporting in household interviews*. US Public Health Service.

Cannell, C. F. et al. (1975). A technique for evaluating interviewer performance.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological methodology*, *12*, 389–437.

Couper, M. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section of the ASA at JSM1998*, 41–49.

Dijkstra, W., & Ongena, Y. (2006). Question-answer sequences in survey-interviews. *Quality and Quantity*, *40*(6), 983–1011.

Dykema, J., Lepkowski, J. M., & Blixt, S. (1997). The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. *Survey measurement and process quality*, 287–310.

European Social Survey. (2012). ESS round 6 data file edition 2.4. doi:10.21338/ESS6E02_4

European Social Survey. (2014). ESS round 7 data file edition 2.2. doi:10.21338/ESS7E02_2

European Social Survey. (2016a). ESS-6 2012 documentation report. edition 2.4. doi:10.21338/NSD-ESS6-2012

European Social Survey. (2016b). ESS-7 2014 documentation report. edition 3.2. doi:10.21338/NSD-ESS7-2014

Fowler, F. (1990). Standardised survey interviewing. Sage Publications, Newbury Park.

Groves, R. M., & Magilavy, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public opinion quarterly*, *50*(2), 251–266.

Hagenaars, J. A., & Heinen, T. G. (1982). Effects of role-independent interviewer characteristics on responses. *Response behaviour in the survey-interview*, 91–130.

Hess, J., Singer, E., & Bushery, J. (1999). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, *11*(4), 346–360.

Hox, J. J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological methods & research*, *22*(3), 300–318.

Hyman, H. H. (1954). *Interviewing in social research*. The university of chicago press.

Jans, M., Sirkis, R., & Morgan, D. (2013). Managing data quality indicators with paradata based statistical quality control tools: The keys to survey performance. *Improving Surveys with Paradata: Analytic Uses of Process Information*, 191–229.

Japec, L. (2007). Interviewer error and interviewer burden. *Advances in telephone survey methodology*, 185–211.

Kalton, G., & Schuman, H. (1982). The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society: Series A (General)*, *145*(1), 42–57.

Kirchner, A., & Olson, K. (2017). Examining changes of interview length over the course of the field period. *Journal of Survey Statistics and Methodology*, *5*(1), 84–108.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American statistical association*, *57*(297), 92–115.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*(3), 213–236.

Loosveldt, G. (1997). Interaction characteristics of the difficult-to-interview respondent. *International Journal of Public Opinion Research*, *9*(4), 386–394.

Loosveldt, G., & Beullens, K. (2013a). 'how long will it take?' An analysis of interview length in the fifth round of the European Social Survey. In *Survey research methods* (Vol. 7, pp. 69–78).

Loosveldt, G., & Beullens, K. (2013b). The impact of respondents and interviewers on interview speed in face-to-face interviews. *Social Science Research*, *42*(6), 1422–1430.

Maitland, A., & Presser, S. (2016). How accurately do different evaluation methods predict the reliability of survey questions? *Journal of Survey Statistics and Methodology*, *4*(3), 362–381.

Mangione, T. W., Fowler, F. J., & Louis, T. A. (1992). Question characteristics and interviewer effects. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, *8*, 293–293.

Mathiowetz, N. A., & Cannell, C. F. (1980). Coding interviewer behavior as a method of evaluating performance. In *Proceedings of the section on survey research methods* (pp. 525–528). American Statistical Association.

Oksenberg, L., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of official statistics*, *7*(3), 349.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. *Improving surveys with paradata: Analytic uses of process information*, *580*, 43.

Olson, K., & Peytchev, A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly*, *71*(2), 273–286.

Olson, K., Smyth, J. D., & Kirchner, A. (2020). The effect of question characteristics on question reading behaviors in telephone surveys. *Journal of Survey Statistics and Methodology*, *8*(4), 636–666.

Ongena, Y. P., & Dijkstra, W. (2006). Methods of behavior coding of survey interviews. *Journal of Official Statistics*, *22*(3), 419.

Rice, S. A. (1929). Contagious bias in the interview: A methodological note. *American Journal of Sociology*, *35*(3), 420–423.

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, *83*(3), 598–626.

Schaeffer, N., & Dykema, J. (2011). Response 1 to Fowler's chapter: Coding the behavior of interviewers and respondents to evaluate survey questions. *Question evaluation methods: Contributing to the science of data quality*, 23–39.

Sharma, S. (2019). *Paradata, interviewing quality, and interviewer effects* (Doctoral dissertation).

Smit, J. H., & van der Zouwen, J. (1997). Suggestive interviewer behaviour in surveys: An experimental study. *Journal of Official Statistics*, *13*(1), 19.

Sudman, S., & Bradburn, N. M. (1974). Response effects in surveys: A review and synthesis.

Vandenplas, C., Beullens, K., & Loosveldt, G. (2019). Linking interview speed and interviewer effects on target variables in face-to-face surveys. In *Survey research methods* (Vol. 13, pp. 249–265).

Vandenplas, C., Loosveldt, G., Beullens, K., & Denies, K. (2018). Are interviewer effects on interview speed related to interviewer effects on straight-lining tendency in the European Social Survey? An interviewer-related analysis. *Journal of Survey Statistics and Methodology*, *6*(4), 516–538.

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of survey statistics and methodology*, *5*(2), 175–211.

West, B. T., & Sinibaldi, J. (2013). The quality of paradata: A literature review. *Improving surveys with paradata: Analytic uses of process information*, 339–359.

Wuyts, C., & Loosveldt, G. (2017). *The interviewers in the European Social Survey round 5 to 7—Belgium*. Centre for Sociological Research; Leuven.

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. In *Survey research methods* (Vol. 8, pp. 127–135).

Appendix
Tables

(Appendix table follows on next page)

Table A1
Intra-interviewer correlation estimates for selected questionnaire items

Item	Label	ESS6_BEDUT	ESS7_BEDUT
iprspot	Important to get respect from others	0.1044	0.0974
prtdgcl	How close to party	0.0869	-
dfegcon	Different race or ethnic group: contact, how often	-	0.0861
tvpol	TV watching, news, politics, current affairs on average weekday	-	0.0852
dspplvt	Voters discuss politics with people they know before deciding how to vote	0.0815	-
iplylfr	Important to be loyal to friends and devote to people close	0.0763	-
freehms	Gays and lesbians free to live life as they wish	0.0728	-
gvctzpv	The government protects all citizens against poverty	0.0725	-
qfimedu	Qualification for immigration: good educational qualifications	-	0.0695
cttresa	The courts treat everyone the same	0.0677	-
gvexpdc	The government explains its decisions to voters	0.0667	-
votedir	Citizens have the final say on political issues by voting directly in referendums	0.0666	-
alcbnge	Frequency of binge drinking for men and women, last 12 months	-	0.0665
gvcodmc	In [country] the government is formed by coalition	0.0643	-
ipsuces	Important to be successful and that people recognize achievements	0.0598	-
euftf	European Union: European unification go further or gone too far	-	0.0592
ipshabt	Important to show abilities and be admired	0.0590	-
dfprtal	Different political parties offer clear alternatives to one another	0.0545	-
impenv	Important to care for nature and environment	0.0531	-
qfimlng	Qualification for immigration: speak [country's] official language	-	0.0524
gptpelcc	In [country] governing parties are punished in elections when they have done a bad job	0.0523	-
flclpla	Feel close to the people in local area	0.0519	-
tmendng	Enthusiastic about what you are doing, how much of the time	0.0517	-
tnapsur	Take notice of and appreciate your surroundings	0.0517	-
eimpcnt	Allow many/few immigrants from poorer countries in Europe	-	0.0513
grdfincc	In [country] the government takes measures to reduce differences in income levels	0.0496	-
imbleco	Taxes and services: immigrants take out more than they put in or less	-	0.0494
slprl	Sleep was restless, how often past week	0.0484	-
algyplv	Allow many or few Gypsies to come and live in country	-	0.0484
noimbro	Of every 100 people in country how many born outside country	-	0.0474
imdfetn	Allow many/few immigrants of different race, ethnic group from majority	-	0.0442
ipstrgv	Important that government is strong and ensures safety	0.0391	0.0486

Continues on next page

Continued from last page

Item	Label	ESS6_BEDUT	ESS7_BEDUT
gvctzpvc	In [country] the government protects all citizens against poverty	0.0433	-
qfimchr	Qualification for immigration: Christian background	-	0.0432
impdiff	Important to try new and different things in life	-	0.0418
plinsoc	Your place in society	0.0417	-
rghmgpr	The rights of minority groups are protected	0.0409	-
ctstogv	The courts are able to stop the government acting beyond its authority	0.0409	-
impfun	Important to seek fun and things that give pleasure	0.0398	0.0414
gvrfgap	Government should be generous judging applications for refugee status	-	0.0404
deaimpp	Deal with important problems in life	0.0400	-
etapapl	Easy to take part in politics	-	0.0399
cptppol	Confident in own ability to participate in politics	-	0.0395
imwbcnt	Immigrants make country worse or better place to live	-	0.0388
pltaviec	In [country] politicians take into account the views of other European governments	0.0387	-
trtrsp	Feel people treat you with respect	0.0383	-
gincdif	Government should reduce differences in income levels	0.0380	-
lotsgot	There are lots of things I am good at	0.0378	-
votedirc	In [country] citizens have the final say on political issues by voting directly in referendums	0.0373	-
ipmodst	Important to be humble and modest, not draw attention	0.0372	-
ipeqopt	Important that people are treated equally and have equal opportunities	-	0.0364
prhlppl	Provide help and support to people you are close to	0.0363	-
ipgdtim	Important to have a good time	0.0359	-
ipadvnt	Important to seek adventures and have an exciting life	-	0.0353
ipudrst	Important to understand different people	0.0351	-
pltavie	Politicians take into account the views of other European governments	0.0351	-
happy	How happy are you	0.0342	-
psppsgv	Political system allows people to have a say in what government does	-	0.0339
enjlf	Enjoyed life, how often past week	-	0.0321
flapppl	Feel appreciated by people you are close to	0.0318	-
gvtrimg	Compared to yourself, government treats new immigrants better or worse	-	0.0306