OZP

Research Article

Quality management in social sciences research
Quality management in the Survey of Health, Ageing and Retirement in Europe (SHARE)

Nicole Halmdienst[1,*], Michael Radhuber[1,*]

1 Johannes-Kepler University Linz, Austria

* nicole.halmdienst@jku.at / michael.radhuber@jku.at

Abstract

Austria has been a member country of SHARE since its inception in 2004. In this paper, we address quality management in surveys and highlight three components – contract, sampling, and fieldwork management – that are fundamental for high data-quality. We provide an overview of a SHARE wave and discuss our approach to data-quality management based on the example of SHARE management in Austria. Results confirm that focusing on fieldwork quality management has the potential to improve overall data quality.

Keywords

Survey, survey management, quality management, fieldwork monitoring, interviewer behavior, survey contract, sampling

Qualitätsmanagement in der empirischen Sozialforschung
Ein Ansatz für ein umfassendes Qualitätsmanagement im Survey of Health, Ageing and Retirement in Europe (SHARE)

Zusammenfassung

Seit Beginn von SHARE zählt Österreich zu den Mitgliedsländern. In diesen 13 Jahren als SHARE-Mitglied konnte viel Erfahrung im Management und der Umsetzung der Studie gesammelt werden. Drei Komponenten – Vertrag, Stichprobenziehung und Feldmanagement – sind der Schlüssel um in einer großen standardisierten Studie wie SHARE erfolgreich zu sein und hohe Datenqualität zu liefern. In diesem Artikel geben wir einen Überblick über den Verlauf einer SHARE-Welle und diskutieren unseren Ansatz für ausgezeichnetes Datenqualitätsmanagement am Beispiel der SHARE-Studie in Österreich. Unsere Erfahrungen zeigen, dass man mit gezieltem Feldmanagement und Interviewer-Feedback die Datenqualität steigern kann.

Schlüsselwörter

Studie, Studienmanagement, Qualitätsmanagement, Feldmonitoring, Interviewer-Verhalten, Vetragsgestaltung, Stichprobenziehung

The authors have declared that no competing interests exist.

October 31, 2018 I *innsbruck* university press, Innsbruck
OZP – Austrian Journal of Political Science I ISSN 2313-5433 I http://oezp.at/
Vol. 47, issue 2 I DOI 10.15203/ozp.2021.vol47iss2

1. Introduction

The Survey of Health, Ageing and Retirement in Europe (SHARE) is the biggest interdisciplinary and longitudinal survey in social sciences in Europe. SHARE aims at providing an extensive research database for a better understanding of the relationship between the economic situation, family, social networks, and general health of ageing individuals, over the age of 50 years.

Since its commencement in 2004, more than 120,000 individuals in 27 European countries and Israel have been interviewed. Austria has been a cornerstone of SHARE since its beginning and is currently one of the SHARE frontrunners in terms of innovation and funding.

Excellent data quality is an essential requirement for excellent research. A core task of SHARE, therefore, consists of accompanying and supervising survey agencies entrusted with fieldwork tasks. This article aims to outline the quality management within SHARE, with a special focus on Austrian instruments that are implemented in addition to international measures. We begin by providing a synopsis of the agenda of a SHARE wave. Next, based on our experience, we describe three essential components of success in a standardized international survey: contract, sampling, and fieldwork management.

As is often the case, the best learning effects arise from learning by doing and, sometimes, from failing. Although we as authors were primarily responsible for survey operations in Austria, our experiences and findings relate not only to Austria, but to the international cooperation of the current 27 European SHARE member countries and Israel.

2. The organization of SHARE

SHARE preparations regularly initiate about two years before the first interview is conducted for the particular wave in consideration. The first task of SHARE scientists consists of designing the questionnaire for the new wave. As soon as the generic questionnaire is set up, all member countries engage in translations with the aid of an online instrument. After translation, national CAPI (Computer Assisted Personal Interview) instruments are compiled for intense testing by country scientists (Malter 2015).

All SHARE interviews are carried out as face-to-face CAPIs. Fieldwork agencies engaged in each member country provide interviewers and logistic support for fieldwork management while software and background technical infrastructure is provided by SHARE. These fieldwork agencies are commonly hired in open tender bids based on harmonized procurement rules.[1]

Subsequently, selected interviewers are trained and instruments and materials are evaluated for the first time in the field during the pretest, where around 100 interviews are conducted in each country. After error management and questionnaire refinement, the translating, programming, and testing processes start all over (Malter 2015).

Six months after the pretest, the second field test, called the "field rehearsal," is set in motion.[2] Significant changes in the questionnaire are allowed between the pretest and field rehearsal. However, after the field rehearsal, the generic questionnaire is left untouched. Before the onset of the main fieldwork, the SHARE questionnaire is approved by the SHARE questionnaire board and extensively tested in theory and practice. Translations are verified and approved by survey scientists as well as by external linguists. The survey instrument must pass various test-runs as well as the two practical tests during pretest and field-rehearsal.

Approximately two years after the beginning of a new wave, the main fieldwork phase commences. During fieldwork, the survey agency carrying out the data-collection is supported, monitored, and supervised by SHARE Central (SHARE headquarters, a statutory seat of SHARE-ERIC, currently in Munich, Germany) and scientific country teams. The principal duty of SHARE Central is to coordinate and supervise the efforts of scientific country teams, while country teams are responsible for in-country management and survey-agency supervision.

Once fieldwork is accomplished, all SHARE teams are involved in data cleaning and preparations for the first public data release. It is common practice that a new wave is kicked off while the previous wave is still in fieldwork or in post processing.

3. Contracting survey tasks

In most cases, survey interviews will not be carried out by researchers themselves, but by trained interviewers and specialized or commercial survey agencies. In this section, we focus on topics essential for outsourcing core

1 Procurement rules based on SHARE-ERIC Statutes apply to all SHARE-ERIC member countries. These Statutes can be found here: http://www.share-project.org/fileadmin/pdf_documentation/SHARE-ERIC/SHARE-ERIC_consolidated_version_27_04_2017.pdf

2 To be up to date, we adapted the annotation of field tests: Before wave 7, the "pretest" was called "pilot" and did not have to be conducted by the main agency. The second field test, the "field rehearsal" was called "pretest" in former waves.

survey tasks to specialized survey agencies (for additional arguments, also refer to Cibelli Hibben et al. 2018).

First, it is important to carefully evaluate the agency or company to contract for the specific survey objectives. In many economies, the "market-research" sector is very competitive and public tender calls often attract multiple bidders of very different characteristics. It is not uncommon for small survey agencies to bid for bigger projects. However, the experience of survey agencies in conducting survey tasks at the required scale is one of the most important criteria for success. Tender bids assessment should, therefore, focus on survey agencies' experience in handling surveys of the requested scale in addition to the bid contents.

Second, the question of legal ownership of the sample arises. In many countries, while the data collected would be the sole property of the purchaser, it is a common practice that the sample remains in the legal ownership of the survey agency, and thus, completely out of the control of researchers. However, we strongly argue in favor of carrying out the sample selection process ourselves as researchers or survey purchasers. If, for various reasons it will not be possible to draw a sample without engaging a survey agency, tight controls of the sample selection process are indispensable. The (geographic) sample distribution is a major cost driver for survey tasks. Thus, if not strictly bound by contract and controlled for, there are numerous incentives for survey agencies to "interfere" with the sample selection process. It should be noted that usually controls of the sample selection process will not be possible if the question of legal ownership of the sample has not been specifically addressed in the contract. Thus, the property rights of the data collected and of the sample data (address files) require special attention when drafting survey contracts.

Third, the question of ex-post interview controls carried out by survey agencies needs to be addressed. Interview controls, e.g., by re-contacting a certain number of randomly drawn respondents per interviewer and inquiring about the place and time of interview, questions posed, and interviewer friendliness, are considered standard in current survey quality management (Lyberg/Biemer 2008). If possible, such verification calls should be carried out by independent, third-party survey agencies. Alternatively, in some cases, call protocols or call recordings may serve as substitutes.

The protocol of dealing with interviewers presenting anomalies in ex-post verification calls and/or other data quality checks should be agreed on in advance. In some cases, it might be a good idea to agree on an extension of ex-post controls to all interviews carried out by interviewers with questionable control records, and a possible exclusion of all affected interviews. The exclusion of such interviewers from subsequent waves of data collection is another option to be considered.

Fourth, we also suggest including specific start and end dates, as well as other milestones such as due dates in the agreement. In many cases, reducing undesirable variation in data due to external effects requires the data collection period to be as short as possible. Penalty payments may facilitate compliance with contractual deadlines.

Lastly, interviewer effects may pose further concerns. For an excellent summary of interviewer effects and practical considerations for survey management see Cibelli Hibben et al. (2018). To secure a minimum number of active interviewers for a certain survey, a three percent (or similar) clause was often included in SHARE contracts: No interviewer shall carry out more than three percent of all interviews collected for the survey.

3.1 The collection of paradata

Many professional surveys such as the Health and Retirement Study (HRS), which is the SHARE sister study in the US, employ their own survey software that enables the collection of paradata and offers several other control possibilities. Paradata, also called keystroke data, are automatically collected data on time, duration, and sometimes even the place where the survey is conducted by interviewers. They allow for an independent ex-post evaluation of interview length and accurate reading of question texts.

For most smaller-scale surveys sourced out to survey agencies, programming of its own survey software might not be feasible. In such cases, researchers only provide survey agencies with a proper questionnaire, while the programming and technical implementation of the survey questionnaire is entrusted to the survey agency. As such, scientists are able to exercise hardly any control or supervision options during the fieldwork period. To avoid possible cheating by interviewers or survey agencies, in addition to regular deliveries of the survey data collected, it is also advisable to agree on the regular or automatic delivery of survey paradata during the fieldwork period.

3.2 Respondent and contact procedures

How often and at what times shall interviewers contact respondents before their lack of response is considered a hard refusal? Most surveys require at least six to eight contact attempts at different times of the day and week before a respondent can be considered to have made a soft or hard (final) refusal. It is part of the work of contract editors to include specific clauses on the modes,

numbers, and timing of contact attempts to be made by survey agencies or interviewers contracted for the fieldwork.

In all cases, data-protection rules have to be taken into account. According to Article 21 of the General Data Protection Regulation (GDPR) in Europe, once a respondent objects to processing his or her personal data, all contact procedures have to stop; in some cases, the respondent's data may even have to be excluded from the sample.

3.3 Valid interviews

Which interviews count as valid interviews and which are not, and hence, will not be remunerated? Every survey contract needs to detail some rules on how to proceed with incomplete interviews and interviews containing false data.

3.4 Bonus payments for high retention rates

Finally, every survey purchaser is advised to stipulate some form of bonus payments targeting high retention rates as well as the reduction of non-response errors. Expenses of survey agencies tend to grow with higher retention rates as they often have to employ specialized interviewers to convert unwilling respondents. Moreover, respondent incentives, another issue of concern to survey purchasers, also grow with higher retention rates. Lastly, bonus payments should also trickle down to interviewers, as they bear the bulk of the effort involved in conducting survey fieldwork.

4. Sampling methods and fieldwork reality

Sampling is one of the most crucial factors of survey quality. Without an adequate sample design and practical implementation of that design, the whole survey enterprise runs the risk of scientific uselessness. "Conclusions drawn from a poorly designed survey […] can be completely misleading" (Lohr 2008, 147).

According to Groves and Lyberg (2010), four possible error sources have to be coped with when carrying out a survey:

4.1. Coverage error,
4.2. Sampling error,
4.3. Non-response error, and
4.4. Measurement error.

These four parts make up the so called total survey error (TSE). In this context, TSE may be specified as:

$$TSE = \varepsilon_{Coverage} + \varepsilon_{Sampling} + \varepsilon_{Non\text{-}response} + \varepsilon_{Measurement}$$

Coverage error occurs when the sampling frame excludes parts of the population of interest. Sampling error occurs because a sample is taken instead of measuring the entire population (Lohr 2008). Non-response error arises when respondents are contacted for the survey, but provide no or only partial data.

Finally, measurement error results from inaccurate responses to questions or inaccurate measurements. As outlined further below, it is the principal objective of survey quality management to control and reduce measurement error to a minimum. Failure to take into account these different error sources may lead to bias and the survey may not adequately represent the population of interest. In what follows, we address the four error components of surveys as discussed in Lohr (2008), and add further considerations with regard to survey quality management.

4.1 Coverage error

A registry or database containing the entire population of interest for the survey is a necessary condition for carrying out an adequate sampling procedure. Incomplete or inadequate population registries such as telephone books are sometimes used to draw a survey sample. The exclusion of certain population groups from the population of interest, e.g., persons unlisted in telephone registers, leads to coverage bias. Inherently, since we cannot know who was excluded from our sampling process by using incomplete registers, we are not able to mathematically compute the resulting coverage bias. It is certainly possible to try to minimize this bias by post-stratification methods such as age-by-sex or race-by-sex categories. However, in all cases, we end up with a more or less pronounced representativity bias of the survey. In such instances, all our estimations end up as nothing more than "good guesses."

Nowadays, population registries are considered state-of-the-art for drawing samples for scientific surveys. Wherever population registers are not available or legally accessible, other population-linked databases may be employed, such as postal registries or address books containing a list of all households in certain geographic entities. However, two problems may arise from such approaches: First, such registers are often far from perfect and, in many cases, miss out on some parts of the population. Second, many surveys focus on subcategories of a certain population. Often, information on the characteristics of interest of the population (e.g., only respondents over the age of 50 in SHARE) is not included in sample databases on household or postal delivery points. Such lack of information makes screening processes after sampling unavoidable.

However, the practical implementation of the process of screening after sampling can be very problematic

and error prone. First, adequate screening can be very costly, since in many cases it requires multiple individual verifications of every sample point. Second, if no detailed and reproducible documentation of the screening process is provided, interviewers entrusted with screening tasks face a very big incentive to cheat during screening. Let's consider, for example, the case of an address book or postal register containing a list of every household in a certain geographical unit. If nobody answered door A, why keep returning to retry at door A instead of just knocking at doors B and/or C (i.e., doors, that are not included in the drawn sample) to check if some interview-eligible person may be at home? In many cases, it will be very difficult, if not impossible, to rule out such bad practices.

4.2 Sampling error

The other ingredient for achieving survey representativity is probability sampling. According to Lohr (2008), probability sampling is a sampling method in which each respondent is assigned a probability for being selected in the sample. At its purest, it takes the form of simple random sampling.

Without probability sampling, the sampling error cannot be reliably computed and meaningful estimates of the underlying baseline population cannot be developed. Other methods of random sampling such as clustering or stratification can provide other statistical or financial advantages but also tend to increase sampling errors. Similarly, non-random sampling methods such as quota sampling are widely used in market research. However, it should be noted that such approaches may in some cases come close to population averages, but can never be applied to reliably compute population estimators.

Larger sample sizes are required to minimize sampling errors. Above certain thresholds, sample sizes can be computed independent of population sizes since marginal increases in sample size are expressed as a decreasing function of population size. The "standard" formula for computing sample sizes is as follows (Quatember 2015, 47):

Equation 1

$$n_{req} = \frac{u^2_{1-\frac{\alpha}{2}}.N.p.(1-p)}{\varepsilon^2.(N-1)+u^2_{1-\frac{\alpha}{2}}.p.(1-p)}$$

where N is the reference population, p is the relative size of the subpopulation of interest, ε is the required margin of error, and u is the required confidence interval (e.g., inferred from Student's t-distribution). For very big reference population sizes, we can rewrite *Equation 1* by multiplying with 1/N and letting N tend to infinity:

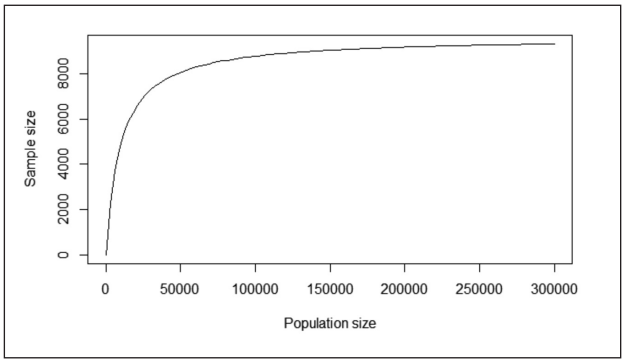
Equation 2

$$\lim_{N\to\infty} n_{req} = \frac{u^2_{1-\frac{\alpha}{2}}.p.(1-p)}{\varepsilon^2}$$

We thereby obtain a formula independent of the size of the reference population that is valid only for big populations.

To demonstrate this step graphically (see Figure 1), we set ε=0.01**pop* (1% error margin in terms of the baseline population), $u_{1-\frac{\alpha}{2}} = 1.96$ (for a confidence interval of 95% in a distribution akin to the t-distribution), and p=0.5 in *Equation 1* and obtain the following function for sample size dependent on the baseline population. We see that the required sample size grows steeply until a certain threshold of about 50,000 members of the baseline population; thereafter, it remains relatively stable at sample sizes of a little over 8,000 respondents.

Figure 1: Required sample size by population size for a 1% margin of error and 95% confidence interval

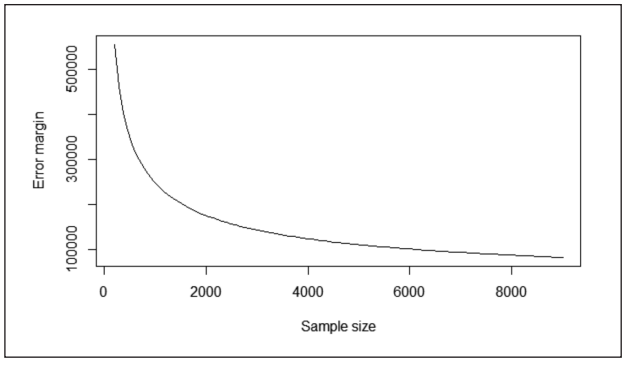


Thus, most formulas for calculating required sample sizes completely ignore population sizes. In SHARE, sample sizes of 6,000 individuals per country have been set as the objective target for participating countries.

In Figure 2, we illustrate the margin of error for different sample sizes with a baseline population of 8 million people, which is roughly equivalent to the population of the Republic of Austria. Variance and confidence intervals are left untouched. The margin of error – expressed in absolute terms of the baseline population – decreases as the sample size increases. In this case, most scientific appliances in survey research require sample sizes of at least 2,000 and ideally up to 6,000 or 8,000 respondents. The required sample size also depends on the aim of the study. If cross-country comparisons are not enough and country level analysis or investigation

of subgroups is required, the sample size should be significantly increased.

Figure 2: Margin of error by sample size for a given baseline population of 8 million people and a 95% confidence interval



4.3 Non-response error

Non-response error occurs when some respondents with common characteristics share a higher probability of refusal for survey participation than others. In such cases, non-response error can lead to biased estimates since the sample does not accurately reflect the characteristics of the baseline population. For panel studies, this "unit non-response" poses an even higher risk to the scientific usability and reliability of the survey. If some individuals are more likely than others to drop out of the survey between consecutive waves, the representativity of the sample might be lost altogether after some time. Indeed, there is evidence that the problem of survey attrition has worsened over time (Watson/Wooden 2009).

As Watson and Wooden (2009) have highlighted in their article, there exists mounting evidence of some common determinants that lead to higher attrition rates among respondents. For instance, response rates are almost always higher for women than for men. The youngest as well as the oldest exhibit significantly lower response rates compared to middle-aged individuals. Single persons and single households have a higher probability of survey attrition. Also, the presence of children within households leads to higher attrition rates.

On the other hand, higher education leads to somewhat higher response rates contrary to home-ownership, which has a negative effect on survey participation. The evidence on income is mixed, whereas location mostly has a relatively strong effect on survey participation with residents of bigger cities being less prone to respond or participate in surveys (Watson/Wooden 2009).

In SHARE, we have experimented with different measures to decrease the risks posed by unit non-response and survey attrition. We find that monetary incentives have strong and presumably near-linear effects on participation rates (Börsch-Supan/Krieger 2013). These findings are also aligned with a major meta-study conducted by Singer et al. (1999). Another experiment to evaluate the consequences of non-response on survey quality comprised the administration of an ultra-short questionnaire to people who dropped out; however, this received limited success.

Minimizing unit non-response has always been one of the main priorities within the SHARE survey. Interviewers were bound to undertake at least eight contact attempts on different days and times of day per sample household. Even more importantly, as survey administrators, we always concentrated our efforts on interviewer training since we learnt that interviewers were the best means to reduce non-response rates and survey attrition. As scientific researchers and, thus, "employers," being in direct contact with interviewers, training on respondent handling, and informing about the principal objectives of the survey has always been a major focus in SHARE (Malter 2013).

Where we have failed so far is to implement appropriate monetary incentive schemes to reward interviewers for high quality data and for minimizing unit non-response rates. This might also be due to the fact that interviewers are not hired directly by SHARE researchers but by intermediary survey agencies, often under precarious conditions. In some cases, incentive schemes have been put in place by survey agencies themselves, but almost exclusively only when fieldwork progress was slow and contractual deadlines were already near.

4.4 Measurement error

Measurement error occurs when a respondent's answer to a question is inaccurate and, thus, departs from the "true" value. One way to avoid measurement error is to ask clear and understandable questions. Special attention should, therefore, be dedicated to a well-designed and well-tested questionnaire.

A special sort of measurement error may arise when respondents are confronted with sensitive questions. Depending on the social setting of the interview, in some cases, respondents may opt for wrong answers to sensitive questions. Here, the method of data collection may lead to measurement error. In fact, in some cases, there may be good reasons to rely on interviewing methods without interviewers.

Besides sensitive questions, interviewers can also cause measurement errors when confronted with questions that they, themselves, do not understand. Good interview training is essential to resolve all interviewer questions with regard to the questionnaire and its contents (Lessler et al. 2008). Additionally, survey instruments should allow for special interviewer instructions and background information for every survey question.

5. Targeting fakes and errors through fieldwork quality management

Throughout data collection, "[…] systematically validating the work of field staff is a requirement for the responsible collection of survey data." Therefore, analyzing interview paradata is a key tool to detect and prevent falsification in computer assisted interviews (Johnson et al. 2001, 1). A sound fieldwork management and monitoring strategy is essential for reducing survey errors and achieving high-quality data.

The major objective of SHARE fieldwork management consists of avoiding measurement error and minimizing unit non-response. In general, monitoring is implemented at both country and interviewer level through the international coordination of SHARE and the scientific country teams.

International quality and fieldwork management is of obvious relevance to the project but is not elaborated on here. SHARE survey methods and fieldwork monitoring are documented thoroughly by Kneip et al. (2015) and Malter (2013), and can also be looked up in SHARE methodology volumes (Börsch-Supan/Malter 2013; 2015; 2017; Börsch-Supan/Jürges 2005). Regular fieldwork monitoring provided by SHARE Central focuses on response and contact statistics.

In face-to-face interviews, standardized interviewer behavior is essential to data-quality (Loosveldt 2008). Standardized interviewing requires the interviewer to read the questions accurately as worded, and to follow the given script (Schaeffer 2018; Schwarz et al. 2008). Although the impact of accurate reading of questions on survey data quality is unclear, "[…] good interviewer behavior should not only be measured in terms of response rates, but more closely monitor their actual behavior in the interaction with respondents" (Bergmann/Bristle 2016, 25). Kreuter (2018a; 2018b) argues that paradata on question reading times can be exploited to improve interview guidance and to identify fakes. "The lack of standardized practice and protocols across interviewers, as well as taking 'shortcuts' and outright falsification, can contribute to significant interviewer effects" (Mneimneh et al. 2018, as cited in Cibelli Hibben et al. 2018, 280). For this reason, in SHARE we investigate further and assess interview data per interviewer.

The following section is split into 3 parts. Following the description of paradata and the information flow in the SHARE project, we discuss fieldwork quality management at the interviewer level. The last part of this section presents our experiences with interviewer back checks.

5.1 Employment of paradata

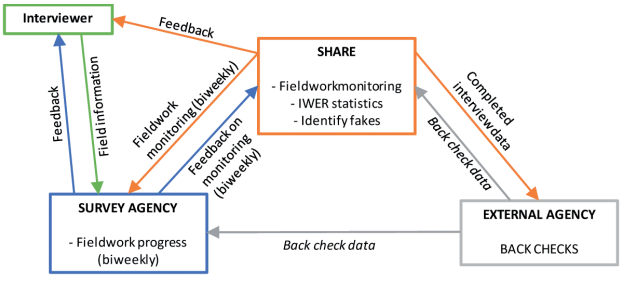
Following Mohler et al. (2008), we distinguish three types of survey data: numeric data, metadata, and paradata. Numeric data are simple survey question responses. Metadata are mostly descriptive data to document details on survey design, questionnaire definitions, interviewer training, and other background information. Finally, paradata are data gathered during the data collection process as briefly discussed above. Nowadays, almost every survey collects additional information on the survey process in the form of metadata and paradata.

In SHARE, we exploit paradata for fieldwork monitoring. SHARE interview software enables collecting a lot of additional information during the data collection process. The software automatically records each user action with the respective time stamp so that the time spent (in seconds) at every question can be calculated (Bristle 2015). A lot of other information collected by SHARE software is dependent on interviewer interactions including, for example, logging of contacts and contact attempts by interviewers (Martens et al. 2015; Wijnant et al. 2013).

SHARE survey agencies only have limited access to collected interview- and paradata. Therefore, the information flow of processed fieldwork data to agencies is crucial for steering the data collection process. All collected data are synchronized every two weeks with CentERdata in the Netherlands, the central data-processing unit of SHARE (Malter 2013). Therefore, the necessary time-frame for interview related feedback to the agencies ranges from one to a maximum of three weeks.

Figure 3 provides a rough overview of this information flow from SHARE to survey agencies in Austria. SHARE regularly provides a fieldwork monitoring report and detailed statistics at the interviewer level to the survey agency. On a monthly basis, feedback is also sent directly to interviewers. Additionally, an external agency, contracted for back checks on interviewer performance, receives regular updates on respondent data and returns the results of control checks conducted.

Figure 3: SHARE fieldwork control information flow in Austria



5.2 Data quality checks at the interviewer level

To perform quality checks at the interviewer level in Austria, we focus on interview duration, question reading measurements, and average item non-response by interviewer. Interview duration is considered a proxy variable for fabricated or shortened interviews. Question reading measurements have been implemented to control for the reading protocol of interviewers, i.e., the complete reading out of question texts by interviewers. Exact reading of the question is a key component in surveys to avoid measurement error from shortened or skipped text (Loosveldt 2008). Identical wording and meaning of questions is of even higher importance in longitudinal and cross-country harmonized surveys (Hoffmeyer-Zlotnik/Warner 2018). Item non-response by interviewer serves as a proxy variable for interviewer motivation and helps to detect systematic shortening of the questionnaire.

However, such measurements also have to be handled with caution. According to Kreuter (2018a; 2018b), no universally accepted indicators and mechanisms exist so far. All the variables specified above can therefore only provide lead evidence for problems in the field and have to be judged in the relevant interview context.

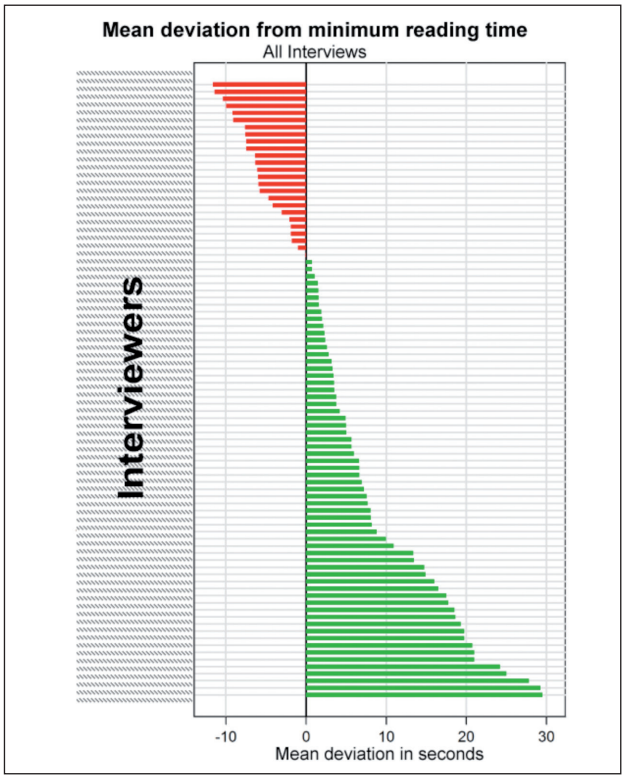
In SHARE, the survey agency is advised to contact underperforming interviewers and discuss possible improvements in interviewing techniques. Additionally, the Austrian SHARE team sends out personalized summary emails with individual performance data directly to all interviewers. As fieldwork progresses, we also estimate trends over time to examine interviewer adaptations to past feedback.

5.2.1 Question reading protocol

We choose three questions without respondent interaction from the SHARE interview. These are typically introductory texts containing important information and definitions on the questionnaire that follows. For these texts, we define a minimum reading time – fastest reading time that can still be understood by a listener – in seconds. The SHARE instrument records the time spent at each question in seconds. With help of these data, we compute the mean deviation from the minimum reading time across interviews and by interviewer. This enables us to control for interviewer adherence to reading protocols and to single out interviewers who systematically read questions too fast or even skip entire text passages.

Figure 4 shows the mean deviation from the minimum reading time by interviewer. This graph is extracted from the Austrian data quality report for SHARE wave 5. In total, around 72% of interviewers active in wave 5 show good reading times while the rest are below the minimum. The graph exhibits extremes in both directions. However, our major objective is to motivate interviewers with the "worst" reading times to improve their interviewing techniques. We do not care about the other extreme (long reading times), as these may be rooted in many different motives. Studying the exact reading times of "under-performing" interviewers for each question over time reveals that many interviewers are systematically below the minimum – an indication that they systematically cut question texts.

Figure 4: Sample graph of mean deviation from minimum reading time extracted from the final data quality report of SHARE wave 5



5.2.2 Item non-response

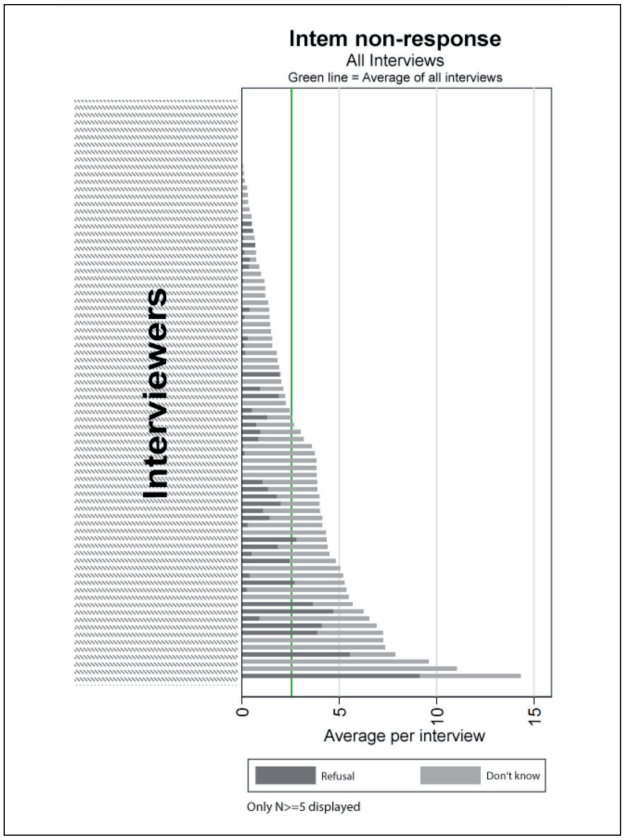
Additionally, we investigate item non-response at the interviewer level. Our objective is to filter out interviewing errors or errors in non-response codes. It is the duty of interviewers to encourage respondent participation and to inspire trust that all data are protected. Although the general policy is based on the respondents' freedom of answering or choosing not to do so, "don't knows" or "refusals" could also be exploited by the interviewer to shorten the interview.

For interviewer feedback, we evaluate the mean number of questions coded as "don't know" and/or "refusal" by interview and interviewer. Figure 5 presents

the respective graph from the final data quality report of wave 5. On average, 2.5 questions are not answered in a complete SHARE interview in Austria. Eventually, we are only concerned about interviews with a very high deviation from the mean; some interviewers present average deviations of 3 to 6 times the mean. The agency is then required to re-engage these interviewers, enquire what problems may have emerged in the field, and investigate the causes of these overshoots. Often, it is simply a problem of wrong coding; in other cases, the interviewer may be operating in areas that are known for high item non-response.

One might argue that instead of motivating respondents to correctly answer questions, interviewers are induced to provide wrong answers instead of "don't knows" or "refusals" due to such checks. In practice, however, such behavior could lead to significant problems since wrong answers will often lead to routings with incongruous questions. How should interviewers then respond to such inappropriate follow-up-questions? In other cases, forging answers is simply not possible, as was the case with social security numbers: When such data was collected in SHARE, it had to pass hash checks before being accepted by the interview software. Inventing such numbers was simply impossible.

Figure 5: Sample graph of item non-response from the final data quality report of SHARE wave 5



5.3 The effect of quality management on data quality

Overall, conducting checks and providing regular feedback to interviewers on their quality measures during fieldwork appears to bear a positive effect on survey data quality. Table 1 presents an overview of average quality indicators from waves 4 to 6. In wave 4, when problems were detected and deeper quality checks and feedback mechanisms were implemented for the first time, only every second interviewer possessed acceptable question reading times. Back then, on average six questions had not been answered per interview. Two years later, in wave 5, interviewers were already more focused on the reading protocol; data quality checks showed that 72% of interviewers displayed acceptable reading times and the average item non-response rate fell to 2.7 items. Again, there was consistent but lesser improvement from wave 5 to 6. In wave 6, three of four interviewers exhibited excellent reading times and the average number of non-answered questions was 1.3.

Table 1: Development of quality control indicators at the interviewer level

	Wave 4 (2011)	Wave 5 (2013)	Wave 6 (2015)
Question reading report: *acceptable reading time* [(a)]	55 %	72 %	76 %
Motivation report: *item non-response* [(b)]	6.1	2.7	1.3

[(a)] Share of interviewers with good reading time
[(b)] Mean number of refusal and „don't know" responses per interview

Nonetheless, all quality measures leave room for discussion and have to be interpreted with caution. Motivating interviewers to be active in the field and to deliver completed interviews while not criticizing them excessively for their interviewing behavior, is like walking on a thin red line. Another open question is whether these methods are ultimately causal for good data quality. We are not able to deliver final answers to these questions yet, but it is our belief and experience that if scientific research were to close its eyes to such issues, data quality could and would be a lot worse.

5.4 Independent verification as key for data quality

Another important contribution to survey quality is regular verification of interviews collected in the field to filter out fake interviews or tricky interviewers that do not follow the script (Johnson et al. 2001; Lyberg/Biemer 2008). In SHARE, at least 20% of every interviewer's interviews have to be back checked by computer assisted telephone interviews (CATI). These back checks are

typically conducted by the survey agency that has been assigned the fieldwork itself, applying its own system (Malter et al. 2016).

As is well known, controlling oneself can never compare to external control standards from third parties. SHARE Austria has, therefore, decided to outsource the required CAPI back checks to independent control-agencies. In SHARE wave 6 (2015), an external survey agency has been contracted for the first time, to conduct audit checks on SHARE interviews in Austria.

On a biweekly basis, we provide new data and receive feedback on CATI checks. The primary survey agency conducting the fieldwork has direct access to feedback data provided by the control agency. Free and secure data exchange between all players in the field is a necessary requirement for this cooperation to work out successfully.

When implementing external back checks for the first time, we experienced that telephone numbers supplied by interviewers were often wrong or missing altogether. This made it difficult for the audit agency to conduct adequate back checks and it was even impossible to carry out any interview verifications for a few interviewers. Two possible explanations come to mind: In the first place, household respondents themselves may refuse to provide telephone numbers. Second, interviewers may keep telephone numbers secret and refrain from recording them in the SHARE software to prevent households from being redistributed to other interviewers or even to veil inconsistencies.

Part of our solution to these issues was to instruct the audit agency to collect telephone numbers themselves. Upon completion of fieldwork for wave 6, we also required the principal survey agency to update telephone numbers on households in the SHARE software. In the end, this was made possible by additional payment incentives for interviewers who updated respondent contact information. This facilitated contacting households for back checks in subsequent waves. We established an overview of every interviewer with suspiciously high number of non-auditable interviews and requested the primary survey agency to take action on these employees. As is always the case, when all efforts to shed light on suspicious cases fail, such interviews will not be accepted nor paid for by SHARE, and the interviewer may be excluded from further fieldwork in current and future waves.

Retrospectively, without sourcing out audit tasks to an external agency, we would never have been aware of the problems linked to non-contactable households and suspicious interviewer behavior. Nevertheless, having independent checks adds further challenges to survey management, infrastructure, and personnel resources. Most importantly, the country team requires access to sample data. In the Austrian case, the sample management server was located directly at the premises of the scientists, with the survey agency operating the system via remote access. This guaranteed permanent access to important information on fieldwork progress by SHARE.

6. A more personal conclusion on the approach to good data quality

As researchers, it is important to not merely outsource the entire data collection process and receive back data with neither additional information nor independent evaluation on data quality. Researchers need to focus not only on the theoretical or statistical aspects of the survey setup, but also need to get their "feet on the ground" and check what is being done in the field and how. This is where quality management should cross every researcher's mind. It is one thing to define probability sampling methods on paper; it is another thing to monitor how survey agencies and interviewers actually recruit new respondents, sometimes beyond the researcher's control.

In our experience, the practical implementation of survey data collection is fundamental to data quality and, thus, to high-quality scientific research. In the business world, quality management is in everyone's mind. In scientific survey research, it should be even more so.

In this paper, we attempt to contribute to the discussion on survey quality management by highlighting some basic elements of survey design, as well as by sharing our practical experiences and findings in carrying out the SHARE survey in a European member country. Our objective consists of developing a harmonized "standards" framework laying out the minimum requirements for survey data quality management.

A quick guideline for survey quality management:

1. CONTRACT: A contract with a survey agency should clarify any issues regarding sample and data ownership, interviewers, interview back checks, how to deal with suspicious interviews and interviewers, and fieldwork deadlines. Moreover, respondent contact procedures and – if necessary – access to paradata have to be agreed upon. A definition of what actually makes a complete interview should not be forgotten.

2. SAMPLING: A register sample and a sufficient number of observations (6,000-8,000 observations) are considered gold standard in survey research. If possible, carry out the sampling process yourself or work closely with the survey agency. If register sampling is not an option, insist on detailed screening and documentation information.

3. QUALITY MANAGEMENT: Besides careful questionnaire development and intensive interviewer training, implement an efficient fieldwork monitoring strategy. In addition to response and contact statistics, study interviewer behavior by examining interview paradata. Back checks by a third party are more reliable than back checks conducted by the principal agency itself.

Acknowledgements

Thank you to all participants of the PUMA-Symposium: "Umfrageforschung in Österreich" (Vienna 2016), and the PUMA-Workshop: "Potenzielle Interviewereffekte und praktische Monitoringstrategien im Feld" (Vienna 2017), for their interest and helpful comments. We also want to convey our thanks to the national funders of SHARE Austria, the Federal Ministry of Education, Science and Research, and the Federal Ministry of Labour, Social Affairs, Health and Consumer Protection. Last but not least many thanks to the international SHARE community for their extraordinary efforts.

Literature

Bergmann, Michael/Johanna *Bristle* (2016), Do interviewers' reading behaviors influence survey outcomes? Evidence from a cross-national setting, SHARE Working Paper Series, 23-2016.

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

Börsch-Supan, Axel/Ulrich *Krieger* (2013), Investigating response behavior, in: Börsch-Supan, Axel/Frederic Malter (ed.), 53-61.

Börsch-Supan, Axel/Frederic *Malter* (ed.) (2013), SHARE wave 4: Innovations & methodology, Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Börsch-Supan, Axel/Frederic *Malter* (ed.) (2015), SHARE wave 5: Innovations & methodology, Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Börsch-Supan, Axel/Frederic *Malter* (ed.) (2017), SHARE Wave 6: Panel innovations and collecting Dried Blood Spots, Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Bristle, Johanna (2015), Measuring interview length with keystroke data, in: Börsch-Supan, Axel/Frederic Malter (ed.), 165-176.

Cibelli Hibben, Kristen/Beth-Ellen *Pennell*/Lesli *Scott* (2018), Interviewer effects in multicultural, multinational, and multiregional surveys, Quality Assurance in Education (just-accepted).

de Leeuw, Edith D./Joop J. *Hox*/Don A. *Dillman* (ed.) (2008), International handbook of survey methodology, New York: Psychology Press.

Groves, Robert M./Lars *Lyberg* (2010), Total survey error: Past, present, and future, in: *Public opinion quarterly,* Vol. 74(5), 849-879.

Hoffmeyer-Zlotnik, Jürgen H. P./Uwe *Warner* (2018), Harmonization for Cross-National Comparative Social Survey Research: A Case Study Using the "Private Household" Variable, in: Vannette, David L./Jon A. Krosnik (ed.), 79-86.

Johnson, Timothy P./Vincent *Parker*/Cayge *Clements* (2001), Detection and Prevention of Data Falsification in Survey Research, in: Survey Research, Vol. 5(3), 1-2.

Kreuter, Frauke (2018a), Getting the Most out of Paradata, in: *Vannette*, David L./Jon A. *Krosnik* (ed.), 193-198.

Kreuter, Frauke (2018b), Paradata, in: *Vannette*, David L./Jon A. *Krosnik* (ed.), 529-536.

Kneip, Thorsten/Frederic *Malter*/Gregor *Sand* (2015), Fieldwork monitoring and survey participation in fifth wave of SHARE, in: Börsch-Supan, Axel/Frederic Malter (ed.), 101-157.

Lessler, Judith T./ Joe *Eyerman*/Kevin *Wang* (2008), Interviewer training, in: *de Leeuw*, Edith D./Joop J. *Hox*/Don A. *Dillman* (ed.), 442-460.

Lohr, Sharon L. (2008), Coverage and sampling, in: *de Leeuw*, Edith D./Joop J. *Hox*/Don A. *Dillman* (ed.), 147-168.

Loosveldt, Geert (2008), Face-to-Face Interviews, in: *de Leeuw*, Edith D./Joop J. *Hox*/Don A. *Dillman* (ed.), 292-318.

Lyberg, Lars/Paul P. *Biemer* (2008), Quality assurance and quality control in surveys, in: *de Leeuw*, Edith D./Joop J. *Hox*/Don A. *Dillman* (ed.), 593-622.

Malter, Frederic (2013), Fieldwork monitoring in the survey of health, ageing and retirement in Europe (SHARE), Survey Methods: Insights from the Field (SMIF).

Malter, Frederic (2015), Questionnaire development in the fifth wave of SHARE, in: Börsch-Supan, Axel/Frederic Malter (ed.), 16-17.

Malter, Frederic/Karin *Schuller*/Axel *Börsch-Supan* (2016), SHARE Compliance Profiles – Wave 6, Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Martens, Maurice/Iggy *van der Wielen*/Arnaud *Wijnant*/Gregor *Sand* (2015), Software Innovations in SHARE Wave 5, in: *Börsch-Supan*, Axel/Frederic *Malter* (ed.), 51-59.

Mneimneh, Zeina/Lars. E. *Lyberg*/Shri *Sharma*/Mahesh *Vyas*/ Dhananjay *Bal-Sathe*/Yasmin *Altwaijri* (2018), Case Studies on Monitoring Interviewer behavior in international and multinational surveys, cited in: *Johnson*, Timothy P./Beth-Ellen *Pennell*/Ineke *Stoop*/Brita *Dorer*, Advances in Comparative Survey Meth-

ods: Multinational, Multiregional and Multicultural Contexts (3MC), Wiley, Hoboken.

Mohler, Peter P./Beth-Ellen *Pennel*/Frost *Hubbard* (2008), Survey Documentation: Toward Professional Knowledge Management in Sample Surveys, in: *de Leeuw*, Edith D./Joop J. *Hox*/Don A. *Dillman* (ed.), 568 – 592.

Quatember, Andreas (2015), Datenqualität in Stichprobenerhebungen: eine verständnisorientierte Einführung in Stichprobenverfahren und verwandte Themen, Berlin Heidelberg: Springer-Verlag.

Schaeffer, Nora C. (2018), Survey Interviewing: Departures from the Script, in: Vannette, David L./Jon A. Krosnik (ed.), 109-119.

Schwarz, Norbert/Bärbel *Knäuper*/Dapha *Oyserman*/Christine *Stich* (2008), The Psychology of Asking Questions, in: *de Leeuw*, Edith D./Joop J. *Hox*/Don A. *Dillman* (ed.), 36-60.

Singer, Eleanor/John *van Hoewyk*/Nancy *Gebler*/Kathrine A. *McGonagle* (1999), The effect of incentives on response rates in interviewer-mediated surveys, in: Journal of Official Statistics, Vol. 15(2), 217.

Vannette, David L./Jon A. *Krosnik* (ed.) (2018), The Palgrave Handbook of Survey Research, Palgrave McMillan, Springer.

Watson, Nicole/Mark *Wooden* (2009), Identifying factors affecting longitudinal survey response, in: Lynn, Peter (ed.), Methodology of longitudinal surveys, United Kingdom: John Wiley & Sons, 157-182.

Wijnant, Arnaud/Maurice *Martens*/Marcel *Das* (2013), Software Innovation in SHARE Wave Four, in: *Börsch-Supan*, Axel/Frederic *Malter* (ed.), 62-73.

Authors

Nicole Halmdienst, born 1985, is a senior scientist and project manager at Johannes-Kepler University Linz. She is the quality manager of SHARE—the Survey of Health, Ageing and Retirement in Europe. Her main field of research is econometrics and statistics, evidence-based policy analysis, business, and regional development.

Michael Radhuber, born 1979, is a senior scientist and project manager at Johannes-Kepler University Linz. He is the operational manager of SHARE—the Survey of Health, Ageing and Retirement in Europe. His main field of interest is regional and development economics, economics of climate change, and law.