New Challenges for Interviewers when Innovating Social Surveys: Linking Survey and Objective Data

Dissertation an der Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München

Vorgelegt von:

Dipl.-Soz. Julie M. Korbmacher

30. September 2014

1. Berichterstatterin:	Prof. Dr. Frauke Kreuter
2. Berichterstatter:	Prof. Axel Börsch-Supan, Ph.D.
Verteidigung am:	15.12.2014

Acknowledgements

I have a long list of people to thank as they supported me in different ways during the years I worked on this dissertation. I express my gratitude to all of them even if not explicitly named in this acknowledgement.

First of all, I would like to thank my supervisors Frauke Kreuter and Axel Börsch-Supan for their valuable personal and academic support, the time they spend in discussing my work, and their patience.

I thank Frauke Kreuter for her encouraging guidance and the opportunity to be a member of the Frauke Kreuter Research Group (FKRG). Discussing and sharing research ideas, progress, and results on a regular basis was an inspiring and very valuable experience for me. This thank is addressed to all members of the FKRG: Stephanie Eckmann, Barbara Felderer, Antje Kirchner, Joe Sakshaug, Jennifer Sinibaldi, and of course Frauke Kreuter.

I thank Axel Börsch-Supan especially for his confidence in my work and the creative freedom I had in the realization and organization of the SHARE-RV projects and the interviewer survey. I am very grateful for the academic and social environment at MEA and I would like to thank all my colleagues. My special thanks goes to my mentor Matthias Weiss, and my colleagues Helmut Farbmacher, Thorsten Kneip, Johanna Bristle, and Luzia Weiss for their scientific and emotional support. I would also like to take the opportunity to thank my former colleagues and co-authors at the "old" MEA in Mannheim: Karsten Hank and Mathis Fräßdorf who did a great job in nudging me into the right direction, as well as Annelies Blom and Ulrich Krieger without whose support the implementation of the interviewer survey would not have been possible. I also thank the Institut für Statistik at the LMU for the invitation to present and discuss my work there.

I gratefully acknowledge financial support from the Munich Center for the Economics of Aging (MEA), the VolkswagenStiftung for funding the pilot studies of SHARE-RV and the biomarker project, as well as the Forschungsnetzwerk Alterssicherung for the funding of the continuation of the SHARE-RV project.

Finally, I thank my family and friends, especially my mother Hildegard Korbmacher and Christian Stumpf for their support and encouragement.

This dissertation uses data from SHARE wave 4 release 1.1.1, as of March 28th 2013 (DOI: 10.6103/SHARE.w4.111) or SHARE wave 1 and 2 release 2.6.0, as of November 29 2013 (DOI: 10.6103/SHARE.w1.260 and 10.6103/SHARE.w2.260) or SHARELIFE release 1, as of November 24th 2010 (DOI: 10.6103/SHARE.w3.100). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life),through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, $N°$211909, SHARE-LEAP, $N°$227822 and SHARE M4, $N°$261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions) .

Abstract

The combination of survey data with more objective information, such as administrative records, is a promising innovation within social science research. The advantages of such projects are manifold, but implementing them also bears challenges to be considered. For example, the survey respondents typically have to consent to the linking of external data sources and interviewers have to feel comfortable with this task.

This dissertation investigates whether and to what extent the interviewers have an influence on the willingness of the respondents to participate in two new projects within the *Survey of Health, Ageing and Retirement in Europe* (SHARE). Both projects had the goal to reduce the burden for respondents and to increase the data quality by linking the survey data with additional, more objective data. Both linkages required the interviewers to collect respondents' written consent during the interview.

The starting point of this dissertation is the question of what influences respondents' decisions to consent to link their survey answers with administrative data. Three different areas are considered: characteristics of the respondents, the interviewers, and the interaction between respondents and interviewers. The results suggest that although respondent and household characteristics are important, a large part of the variation is explained by the interviewers. However, the information available about interviewers in SHARE is limited to a few demographic characteristics. Therefore, it is difficult to identify key interviewer characteristics that influence the consent process.

To close this research gap, a detailed interviewer survey was developed and implemented in SHARE. This survey covers four different dimensions of interviewer characteristics: interviewers' attitudes, their own behavior, experiences in surveys and special measurements, and their expectations regarding their success. These dimensions are applied to several aspects of the survey process, such as unit or item nonresponse as well as the specific projects of the corresponding SHARE questionnaire.

The information collected in the interviewer survey is then used to analyze interviewer effects on respondents' willingness to consent to the collection of blood samples. Those samples are analyzed in a laboratory and the results linked with the survey data. Interviewers' experience and their expectations are of special interest, because as these are two characteristics that can be influenced during interviewer training and selection. The results in this dissertation show that the interviewers have a considerable effect on respondents' consent to the collection of biomarkers. Moreover, the information collected in the interviewer survey can explain most of the variance on the interviewer level.

A motivation for linking survey data with more objective data is the assumption that survey data suffer from recall error. In the last step, the overlap of information collected in the survey and provided in the administrative records is used to analyze recall error in

the year of retirement. The comparison of the two datasets shows that most of respondents remember the year they retired correctly. Nevertheless, a considerable proportion of respondents make recall errors. Characteristics can be identified which increase the likelihood of a misreport, However, the error seems to be unsystematic, meaning that no pattern of reporting the event of retirement too late or too early is found.

Zusammenfassung

Die Verknüpfung von Umfragedaten mit objektiveren Daten, wie zum Beispiel administrativen Daten, ist eine vielversprechende Innovation in der sozialwissenschaftlichen Forschung. Die Vorteile solcher Projekte sind vielfältig, jedoch birgt deren Umsetzung auch einige Herausforderungen, die berücksichtigt werden müssen. So müssen zum Beispiel die Befragten bereit sein an diesen zusätzlichen Projekten teilzunehmen, und auch die Interviewer müssen bereit sein, diese Projekte umzusetzen.

Diese Disseration beschäftigt sich mit der Frage ob und in welchem Ausmaß die Interviewer die Bereitschaft der Befragten beinflussen, an zwei neuen Projekten des *Survey of Health, Ageing and Retirement in Europe* (SHARE) teilzunehmen. Beide Projekte haben zum Ziel die Belastung für die Befragten zu reduzieren und die Datenqualität zu erhöhen indem die Befragungsdaten mit zusätzlichen objektiven Daten verknüpft werden. Beide Verknüpfungen setzten das Einverständnis der Befragten voraus, welches während des Interviews durch die Interviewer eingeholt wird.

Der Ausgangspunkt dieser Dissertation ist die Frage, was die Entscheidung der Befragten beeinflusst, der Verknüpfung der Umfragedaten mit administrativen Daten zuzustimmen. Drei unterschiedliche Bereiche werden hierbei berücksichtigt: Eigenschaften der Befragten, der Interviewer, und der Interaktion von Befragtem und Interviewer. Die Ergebnisse zeigen, dass Eigenschaften der Befragten zwar wichtig sind, jedoch ein großer Teil der Variation auf den Interviewer zurückzuführen ist. Da in SHARE nur einige demographische Informationen über die Interviewer verfügbar sind, ist es schwer Eigenschaften der Interviewer zu identifizieren, die den Zustimmungsprozess beeinflussen.

Um diese Forschungslücke zu schließen wurde im Rahmen dieser Dissertation eine detaillierte Interviewer Befragung konzipiert und in SHARE implementiert. Diese Umfrage umfasst die vier verschiedenen Aspekte: Einstellungen der Interviewer, das eigene Verhalten, Erfahrungen mit Umfragen und speziellen Messungen sowie die Erwartungen bezüglich ihres Erfolges. Diese Dimensionen wurden auf verschiedene Bereiche einer Umfrage, (wie zum Beispiel 'Unit' und 'Item nonresponse') sowie auf bestimmte Projekte der entsprechenden SHARE Welle angewandt.

Die in der Interviewerbefragung gesammelten Daten werden genutzt um den Einfluss des Interviewers auf die Bereitschaft des Befragten, der Entnahme von Blutstropfen zuzustimmen, zu untersuchen. Die Blutstropfen werden in einem Labor analysiert um die Ergebnisse mit den Befragungsdaten zu verknüpfen. Von besonderem Interesse sind hierbei die Erfahrungen des Interviewers sowie dessen Erwartungen. Es kann angenommen werden, dass diese Eigenschaften sowohl durch die Interviewer Schulung als auch im Rekrutierungsprozess beeinflusst werden. Die Ergebnisse zeigen, dass der Interviewer einen großen Einfluss auf die Bereitschaft der Befragten hat, der Entnamhe der Blut-

stropfen zuzustimmen. Zusätzlich zeigt sich, dass die Daten der Interviewerbefragung maßgeblich dazu beitragen, die Varianz zwischen den Interviewern zu erklären.

Eine Motivation, warum man Umfragedaten mit objektiven und qualitativ hochwertigen Daten verknüpft, ist die Annahme, dass Umfragedaten unter Erinnerungsfehlern der Befragten leiden. Im letzten Schritt wird die Überschneidung von Informationen aus den Umfragedaten und den administrativen Daten genutzt, um die Erinnerungsfehler bei dem Bericht des Renteneintrittsjahres zu analysieren. Der Vergleich der beiden Datenquellen zeigt, dass die meisten Befragten das Jahr ihres Renteneintritts richtig erinnern. Dennoch findet sich auch ein nicht zu vernachlässigender Anteil an Fehlern. Es können einige Eigenschaften, welche die Wahrscheinlichkeit eines Fehlers erhöhen, ausgemacht werden. Die Fehler scheinen aber unsystematisch zu sein, mit anderen Worten, es gibt keine Tendenz die falsche Jahreszahl bevorzugt in eine Richtung (zu früh oder zu spät) anzugeben.

Contents

1 Introduction

This dissertation studies *interviewer effects* and *measurement error* in social surveys. Surveys are an important instrument for empirical research in the social sciences. Unlike in the natural sciences, where one can often experimentally test how an external factor influences an outcome by isolating the unit of interest, manipulating the factor, and measuring the outcome, this is often not possible when analyzing societies or human behavior. We cannot isolate persons or societies and manipulate external factors to measure how these manipulations influence people's lives independently of other factors. But we can observe variations in external factors and simultaneously observe the behavior of people. The most prominent instrument to observe such variations is a survey. Surveys thus have an important role in understanding people and the society in which we live (Couper, 2013).

The goal of a survey is to produce data that are an error-free representation of all factors assumed to be relevant for the relationship of interest from a sub-population that is representative of the whole population under study. In an ideal world, we randomly sample a given number of units who are all willing to participate in the survey. They answer all questions and spare no effort to always give the correct answer. As a result, we would have data that are a copy of reality allowing us to analyze complex research questions and to reveal interdependencies. However, problems and errors can occur in nearly every step of a survey. Groves et al. (2009) summarize all potential errors in one model: "the survey life cycle from a quality perspective" which is sometimes labeled as the "total survey error" paradigm or the "total survey error" framework (Groves et al., 2009).

One big challenge when conceptualizing a survey is finding a good balance between information content and respondent burden. On the one hand, the more detailed is the information that is collected, the greater the potential to analyze different research questions. On the other hand, very long and exhausting surveys bear the risk that the respondents lose motivation, fatigue, use satisficing strategies such as answering don't know, start heaping or even break off the survey. All these reactions would adulterate the quality of survey data. A solution to this problem is linking survey data with external data sources of the same person such as administrative records. One goal of such procedures is to expand the survey data with detailed information while reducing the respondents' burden.

Even if a survey is short and easy, all relevant factors are included and respondents are highly motivated to answer the questions, other issues such as recall error and subjectively biased answers can influence the outcome. When asking respondents about facts that took place in the past, they differ in how accurate and detailed their memories

are (see Chapter 6). In addition, when thinking about questions about their own life circumstances, peoples' answers are influenced by how they perceive the situation. The self-reporting of respondents' health statuses is just one example for which the logical assumptions could be misleading - similar evaluations of the health status do not necessarily imply that these respondents are similarly healthy. Therefore, collecting objective information is another important challenge that motivates innovations in social surveys.

In recent years, more and more surveys put a lot of effort in developing and implementing new methods and technologies to reduce respondents' burden and to collect objective information.[1] Two examples of such innovations are the foundation of this dissertation: first, the enrichment of data collected in a survey with administrative records of the same persons and second, the collection of blood spots as objective health measures (biomarkers). Both innovations are implemented in the *Survey of Health, Ageing and Retirement in Europe* (SHARE). The goal of both projects is to link the survey data with high quality data that do not suffer from recall error or objective evaluations. The enrichment of survey data with administrative data or objective health measures has high potential for improving survey data quality, because the added data are much more detailed than survey questions could ever be and in the long term the response burden can be reduced as information that is included in external datasets does not have to be asked during the survey.

The benefits of such innovations are obvious, but there are also costs to be considered. In the case of SHARE, the respondents have to consent to these new parts of the questionnaire. As a result a new special form of nonresponse error which may be called 'consent error' has to be added to the total survey error framework.

In addition, SHARE, like most social surveys, is interviewer-mediated so that implementing new research methods also changes the job of the interviewers. In SHARE, the interviewers have to collect respondents' written consent during the interview. In contrast to all other questions, which are fully scripted, collecting consent demands much more spontaneity and flexibility from the interviewers. The interviewers' role changes from that of a classical survey interviewer to a much more complex and versatile one. New tasks are added to those of a standard interview, which includes making contact and gaining cooperation from the sampled unit, asking survey questions, conducting measurements, recording answers and measures, and maintaining the respondent's motivation throughout the interview (Schaeffer et al., 2010). Now interviewers also have to collect very sensitive information such as social security numbers, take care that the respondents sign all forms, administer additional paper work, and even collect biomarkers

[1] By objective I mean the information that is not adulterated by the subjective evaluation of the respondent; this definition does not necessarily imply that this information is error-free.

like blood spots. As such, the job of an interviewer encompasses a diversity of roles and requires a variety of skills. It is not surprising that we frequently find interviewer effects on all interviewer tasks indicating that there is variation in how interviewers handle their various responsibilities. A substantial body of literature describes interviewer effects on various aspects of the survey process (see Chapter 4.2), but relatively little is known about interviewer effects with regard to innovations in the survey world as described above. The success of such projects depends critically on the respondents' willingness to participate as each refusal reduces the number of cases available for analyses. This first reduces the statistical power and second may result in 'consent biases', if the consenting and refusing respondents differ systematically. As the interviewers are the ones who have to 'sell' this request to the respondents, expanding research on interviewer effects to this new task seems prudent.

Therefore, the main focus of this dissertation is on the effect of interviewers on respondents' consent to the two projects implemented in the German sub-sample of SHARE. In addition, the objective information of the administrative data are used to validate the survey answers in order to learn more about recall error when asking respondents about an autobiographical event.

Guide Through the Thesis

This dissertation is developed from and influenced by my work for SHARE in general and for the record linkage project SHARE-RV in particular. It started with the question of what influences the respondents' decision to consent or refuse to new projects while focusing on interviewers' influence thereupon. The following dissertation consists of four parts:

1. The analysis of determinates influencing respondents' willingness to consent to the linking of their survey data with administrative records,

2. the analysis of the effect of interviewers on respondents' willingness to provide blood spots,

3. and the analysis of recall error in reports of the year of retirement.

4. In addition, a new research infrastructure, namely the interviewer survey, was conceptualized and implemented as part of this dissertation.

Chapter 2 briefly describes SHARE, as all following chapters are based on this survey. Chapter 3 analyzes the determinants of respondents' consent to the record linkage project SHARE-RV. As part of this project, German respondents are asked for consent to linking their survey data with their administrative records at the German Public Pension Fund. In addition to characteristics of the respondents, information about the interviewers as well as the interaction between interviewers and respondents is considered. The results show that the interviewers have important influence on the consent decision of respondents. But given that the information about interviewer characteristics is typically limited to demographics, the results also highlight the need for more information about interviewers. Therefore, we decided to collect detailed information on the interviewers working for SHARE by interviewing them prior to fieldwork. The conceptual framework of the interviewer survey as well as its implementation are discussed in Chapter 4. Chapter 5 is based on the results of Chapter 3 but focusing on the effect of the interviewers on respondents' willingness to provide some drops of blood. As the interviewer survey was implemented in the same wave as the biomarker project, information on the interviewers is used to explain the interviewer effects found. Of special interest are the effects of interviewers' experience and expectations. The variance in consent rates between interviewers can be largely explained by the characteristics collected in the interviewer survey. Chapter 6 targets the another aspect: recall error. It is based on the linked dataset of the project SHARE-RV, which allows the validation of the survey data to learn more about recall error in respondent reports of their year of retirement. The goal is first to quantify the error respondents make and second to identify determinants which influence the error. Comparisons of the self-report and the

true values available from the administrative records show that the majority of respondents report the event correctly. Nevertheless, different determinants such as cognitive abilities, the time-lag since the event, and characteristics of the working history can be identified which increase the likelihood to misreport the event. The last chapter concludes with a future research building on this dissertation.

Publications

Some chapters are already published in reviewed journals, others are still in preparation for submitting. The status of every chapter is as follows:

Chapter 3: Korbmacher J. M. and M. Schroeder (2013): Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer. Survey Research Methods Vol. 7,2 p. 115-131

Chapter 4: Blom A. G. and J. M. Korbmacher (2013): Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework. Survey Methods: Insights from the Field,
http://surveyinsights.org/?p=817

Chapter 5: Korbmacher J. M.: Interviewer Effects on Respondents Willingness to Provide Blood Samples in a Population Based Survey (in preparation)

Chapter 6: Korbmacher J. M.: Recall Error in Reporting the Year of Retirement (in preparation)

2 The Survey of Health, Aging and Retirement in Europe

All chapters of the following dissertation are based on the survey data of the *Survey of Health, Aging and Retirement in Europe* (SHARE). More specifically, they are all based on pilot studies, which had been implemented in the German sub-sample of SHARE. Therefore, this chapter briefly summarizes the survey. A detailed description of the survey can be found in Börsch-Supan et al. (2013). The pilot studies as well as their implementation after the pilot are discussed separately in the second part of this chapter. The relevant aspects of each pilot study will also be discussed in the respective chapters.

2.1 What is SHARE all about?

SHARE is a multidisciplinary panel survey collecting micro data on health, socio-economic status, as well as social and family networks from people aged 50 and older in about twenty European countries and Israel. As SHARE is harmonized with the *Health and Retirement Study* (HRS)[2], which is conducted in the US, and the *English Longitudinal Study of Ageing* (ELSA)[3] a huge research data infrastructure was created. The great strength of these data lies in the comparability of the aging processes and its consequences over different societies. SHARE respondents are asked to be interviewed approximately every two years, starting in 2004 with the first wave of data collection. The number of countries which are included in SHARE differs from wave to wave as new countries join and others drop out. Figure 1 illustrates all countries that ever participated in SHARE.

The Sample

SHARE started in 2004 with its first wave in 11 European countries (Austria, Belgium, Switzerland, Germany, Denmark, Spain, France, Greece, Italy, Netherlands, and Sweden). Israel joined SHARE later and started its first wave in mid-2005. The end of 2006 brought the second wave of data collection with three additional countries (Czechia, Ireland, and Poland). All of these 15 countries participated in the third wave which started at the end of 2008 in most countries. In Wave 4 (2011), four more countries joined (Estonia, Hungaria, Portugal, and Slovenia) but at the same time we also lost two countries (Greece and Ireland). The most recent wave (wave 5) started in 2013 in 15 countries. The difference is composed of one new country (Luxembourg) and three countries which dropped out (Hungaria, Portugal, and Poland) (for an overview see Börsch-Supan et al., 2013). In all countries, SHARE is based on probability samples of the non-institutionalized population aged 50 and older at the time of the sampling. To compensate for panel attrition, and to fill the gap which arises as the respondents gets

[2] `http://hrsonline.isr.umich.edu/`
[3] `http://www.elsa-project.ac.uk/`

Figure 1: SHARE Countries (Börsch-Supan et al., 2013)

older, new refreshment samples are added from time to time. The sampling resources differ between countries but most of them are based on population registers. In addition to the target person, the partner living in the same household is also eligible for the SHARE interview, independent of his or her age.

The Questionnaire

Wave 1, 2, 4, and 5 of SHARE are based on questionnaires that cover the respondents' current living conditions. With the exception of some small changes, the core questionnaire is stable over the waves. Some new modules and measures have been added in some waves. To keep the interview as short as possible, not all respondents have to answer all questions. Information which is valid for the whole household will be collected only once by filtering some modules for respondents in households where both partners participate. This filtering is implemented by assigning specific roles to the respondents of a household: the family respondent (first person in the household participating), the financial respondent (the person who best knows about the financial situation), and the

household respondent (who is also selected at the beginning); they answer the questions to the connected topic for the whole household. In addition, the routing of the questionnaire differs between the respondents of the refreshment and panel sample as some questions will only be asked at the very first interview.

The questionnaire of Wave 3 (named SHARELIFE) differs from all other waves as it covers respondents' life histories retrospectively beginning from childhood until the time of interview (For details on SHARELIFE, see (Schröder, 2011)). The topics covered by the life history calendar are analogue to the standard questionnaire. All questionnaires are developed in English (generic version) and are translated into different languages by the country teams of each country. The generic, as well as the translated questionnaires can be downloaded on the internet: `http://www.share-project.org/data-access-documentation/questionnaires`.

The Fieldwork

The interviews are conducted face-to-face (computer-assisted personal interviewing; CAPI) by trained interviewers mainly at the respondents' homes. Each participating country has a national country team and a national survey agency. Interviewers of all countries are using the same instrument, which is translated into all languages by the country teams. To minimize interviewer or agency effects, SHARE standardized the interviewer training for all countries. The fieldwork is also centrally coordinated so that all countries are in the field during the same time period.

Pilot Studies

Before implementing new methods or techniques, elaborate pilot studies are necessary to test the feasibility of the new procedure and the acceptability by the respondents. Most of the innovations are tested in the German sub-sample of SHARE during one standard wave, before they are implemented for all countries. All projects in SHARE which are the subject of this dissertation have been tested in the German sub-sample, namely: (1) the record-linkage project SHARE-RV, (2) the collection of biomarkers, and (3) the implementation of an interviewer survey. Consequently, only the German sub-sample is used in this dissertation. In the following, the German sub-sample as well as the pilot projects will be described in detail in the order they are used in this work.

2.2 Pilot Studies for Innovations in SHARE

SHARE-Germany

Germany has been participating in SHARE since the very beginning (2004) starting with an initial sample size of 3,008 respondents. In the second wave, a refreshment sample was added which over-sampled the cohorts born in 1955 and 1956 to keep the sample representative of the population 50+. For SHARELIFE (the third wave of data collection) no refreshment samples were drawn in Germany nor in other participating countries. In Wave 4, SHARE-Germany started with the panel sample and a huge refreshment sample. Due to capacity limits during fieldwork, the refreshment sample had not been worked off properly. This sample was excluded for further waves of SHARE and the data which were collected are not publicly released but internally available for methodological research (Kneip, 2013). In Wave 5, a new refreshment sample was drawn and successfully finished. These data will be available in the beginning of 2015.

SHARE-RV

SHARE-RV is a cooperation project of SHARE and the Research Data Center of the German Pension Fund (FDZ-RV) with the goal to link the survey data of the SHARE respondents with administrative records of the same person (for an detailed overview of the project see Czaplicki and Korbmacher, 2010; Korbmacher and Czaplicki, 2013). This project started in 2009 as a pilot study in the third wave of SHARE. The results of that pilot are the basis for Chapter 3. The FDZ-RV provides two different datasets which can be linked with the survey data. First, the longitudinal dataset of the insured population (Versichertenkontenstichprobe, VSKT) which includes information on people's working histories and the state of their pension entitlements, and second the cross-sectional pension data (Versichertenrentenbestand, RTBN) which are available for retirees and include information about all cumulated entitlements used for the pension calculation (Korbmacher and Czaplicki, 2013). The codebooks as well as a user-guide are available on the internet `http://www.share-project.org/data-access-documentation/record-linkage-share-rv.html`.

The challenge of that project lies in linking the data of exactly the same person while ensuring the respondent's anonymity. As the respondent's written consent is requested, a separate consent form is needed which collects the signature and an unique identifier to identify the respondent in the records of the German Pension Fund. The identifier used is the respondent's Social Security Number (SSN). As we assume that not all respondents are able or willing to provide their SSN, we also ask for all information needed to assign the correct SSN. As the collected information is highly sensitive, the interviewer has to send the form directly to the FDZ-RV. They make a huge effort in checking and correcting the SSN to avoid linking data to the wrong person. After the data of the con-

senting SHARE respondents are separated and coarsened into the format of a scientific use file (SUF), the file is sent to the Munich Center of the Economics of Aging (MEA) where the linkage with the SHARE survey data is conducted. To assign the records to the survey data of the same person a number independent from the SHARE-ID and the SSN is used. This number is preprinted on the consent form and has to be entered into the computer by the interviewer during the interview. This step allows the linkage of data while retaining the anonymity of the respondents. Once the data is linked, we compare basic demographics included in the administrative and the survey data such as gender, year of birth, month of birth as well as information on children to check the success of the linkage. If there are any doubts that the data do not refer to the same person, the records will not be published. In this case, the respondents will be asked for consent again in the following wave. This procedure should ensure the direct linkage without allowing for disagreement in any variables used to link the data (deterministic linkage (Calderwood and Lessof, 2009)). Once the respondents consented, data from all waves can be linked and analyzed in parallel until the respondents withdraw their consent.

The goal of that pilot study was to test the whole procedure and the acceptability to the respondents. From a technical point of view, the pilot was successful. All information needed to find the records was available, and the data can be provided for the huge majority of consenting respondents. The acceptability to the respondents is measured using the consent rate. In the pilot study we had two steps of consent: the verbal consent during the interview and a written consent on the consent form after the interview. The two consent rates differ a lot: in the first step 73% of the respondents consented but only 63% of them also sent the signed form to the FDZ-RV. To increase the response rate in further waves, it is important to understand which factors determine the respondent's decision to consent or refuse. This step is the goal of Chapter 3.

The Interviewer Survey

The concept of an interviewer survey was formed in parallel to the analysis of the record linkage project as the important role of the interviewers became apparent. The goal was first to learn more about characteristics of the interviewers and second to link the data collected in the interviewer survey with the SHARE survey data. As a result we can use the information collected in the interviewer survey to explain interviewer effects in the SHARE survey data. The link is done via the interviewer ID which is collected at the beginning of the interviewer survey as well as at the end of each SHARE interview. The conceptual framework, which will be described in detail in Chapter 4, was developed as part of this dissertation. Some questions had been adopted from other surveys, for example the PASS interviewer survey (see Kreuter et al. (2014)), but others

had to be developed specific to SHARE. The first version had been implemented with the interviewers conducting the German sub-sample prior to the fieldwork of Wave 4. It was a paper and pencil questionnaire which was distributed at the national interviewer training session and also collected on the same day. 83% of the German interviewers participated with a negligible amount of item nonresponse.

The Collection of Biomarkers

Another pilot study was implemented for the first time in the fourth wave of SHARE: the collection of new biomarkers. The goal of that project is to collect objective information on respondents' health. This new module consists of four measurements including blood pressure, height, waist circumference, and the collection of some drops of blood. Although the results of the first three measurements can directly be used for analyses, the blood samples have to be analyzed by a special laboratory. For all four measurements, the respondent's written consent is requested which is collected on a separate consent form. As in the record linkage project, the challenge lies in linking the correct information to the survey data of the same person while insuring their anonymity. The procedure is very similar to the one used for the record linkage project. Unique barcode stickers on the filter cards and on the consent form allowed us to make the link with the survey data as the interviewer enter the number of the barcode sticker into the instrument. Here too, the goal of the pilot study was to test whether it is possible to implement such a method in a social survey. An additional study tested the validity of the results isolated from blood spots which had been collected in the same way as in the survey (dried blood spots, see Chapter 5) by comparing them with parameters isolated from venous blood of the same person. The latter is the 'routine' laboratory method, while analyzing dried blood spots is a quite new technique. In addition we sent dried blood spots of the same person to different laboratories to test whether a 'laboratory effect' can influences the results. The results show that the implementation was successful, as the majority of the respondents (59%) consented to allow their blood to be taken and the analysis results obtained from dried blood spots seem to be valid.

The pilot studies SHARE-RV and the biomarker project had both been funded by the "Volkswagenstiftung" within the project 'A new perspective for aging research in Germany: linkages between disciplines (biology, medicine, economics, and social sciences) and linkages between data bases (socio-economic surveys, administrative records, and biomarkers)'[4].

[4]Application by Axel Börsch-Supan, Karsten Hank, Hendrik Jürges, Martin Salm and Mathis Schröder.

From Pilot to Standard Module

After all three pilot studies had been tested successfully, they lost the prefix 'pilot' and were implemented in SHARE.

Until now, SHARE-RV was implemented in Waves 3, 4, 5, and is also planned for Wave 6 of SHARE. Based on the results of the pretest, some changes were made: first, the two consent steps were combined into one by skipping the verbal consent and handing out the form to all respondents. Second, the consent process got completely scripted to reduce interviewer effects. It was first repeated in Germany's Wave 4 refreshment sample. The project was again implemented in the fifth wave with one change: in addition to the refreshment sample, we also asked all respondents for consent who refused in the pilot study. As the linked data are not finalized, a consent rate cannot be reported at this time. But as interviewers stated in the interview whether the respondent consented, was unsure, or refused, we can at least estimate the refusal rate. It is much lower compared to the pilot study. Surprisingly, the rate does not considerably differ between respondents of the refreshment sample and those of the panel sample who refused the first time. These preliminary numbers show that the changes we implemented have been effective. The data of the project SHARE-RV will be made publicly available in April 2015. In addition, SHARE countries other than Germany implemented a record linkage module, too. The projects differ a lot between countries not only in the availability of the data but also in the legal regulation whether the respondents' consent is requested or not.

Also the interviewer survey was successfully expanded to other SHARE countries. With the help of the "Charles Cannell Fund in Survey Methodology"[5] the survey could be programmed as a web survey instead of a paper-and-pencil questionnaire making the implementation and harmonization much easier. For Wave 6, even more countries plan to implement the survey. Beginning with this wave, the survey will be an integral part of SHARE. More information about the international interviewer survey and the questionnaires can be found here: `http://www.share-project.org/methodological-research/interviewer-survey.html`. The data of the interviewer survey will also be made publicly available in April 2015.

The biomarker project was pretested in several SHARE countries in the Pretest of Wave 5 but not in the main survey. As for the record linkage project, the legal regulations differ between countries. The new module is limited to the collection of dried blood spots; all other measures are not longer implemented. The project will be regularly implemented in the majority of all SHARE countries in Wave 6 which will start at the

[5]Application: Julie M. Korbmacher and Ulrich Krieger: Interviewing Interviewers, Feb. 2012; `http://home.isr.umich.edu/education/fellowships-awards/the-charles-cannell/-fund-in-survey-methodology/`

beginning of 2015.

As the data of Wave 5 is not yet available, this dissertation is based on Germany's third and fourth waves of data collection.

3 Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer

Please note: a shorter version of this chapter is already published; Korbmacher J. M. and M. Schroeder (2013): Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer. Survey Research Methods Vol. 7(2), 115-131.

3.1 Introduction

The number of projects linking survey data with administrative records is increasing. At the conference of the European Survey Research Association (ESRA) in July 2011, sixteen papers dealt with challenges of, and findings from, combining survey data with administrative records, a sharp increase from four papers in the 2009 conference. While record linkage is becoming more popular in the social sciences, it is already common in other fields, especially in epidemiology. The enrichment of survey data through administrative records is the primary motivation for the linkage (Calderwood and Lessof, 2009). The data quality in the resulting datasets provides excellent opportunities for research, but the linked data also help to reduce the burden for respondents and interviewers as well as survey cost (Sala et al., 2012; Schnell, 2012). Although administrative data are not primarily generated for research purposes, there are some advantages compared to survey data. For example, they usually cover the universe of the population of interest and are thought to be more accurate than survey data, because problems arising in surveys, such as recall error or misreporting, may not affect the quality of administrative data as severely (Calderwood and Lessof, 2009).[6] On the other hand, administrative data are often collected for a specific purpose and only include standardized information, such as process data for a hospital visit. Unlike surveys, researchers have no, or only limited, influence on what data are collected (Hartmann and Krug, 2009). Thus, using administrative data alone may restrict the selection of control variables. Therefore, researchers benefit if they can combine survey data with administrative records.

There are two common ways to establish the data linkage: one is to use statistical matching procedures based on distance measures, where respondents from a survey are matched to "similar" (in a statistical sense) people in the administrative records. The other way is to ask respondents directly for the permission to link their survey data to their administrative records, building a direct link between the two data sources (see Calderwood and Lessof, 2009, for an overview). There are advantages and disadvantages to both procedures: the first approach heavily depends on the variables identical in both data sources and the smaller this overlap is, the harder it is to establish a match that

[6] As Groen (2012) points out, administrative data could also suffer from errors related to imputation and editing, even though they may exceed survey data in terms of quality in many contexts.

is statistically sound.[7] As long as matching the datasets does not allow the specific and exact identification of respondents, data protection regulations usually do not require the respondents' permission to the matching procedure (Rasner, 2012). For the direct linkage on the other hand, data quality is usually thought to be more promising, but, in most cases, the informed consent by the respondent is necessary (Lessof, 2009; Schnell, 2012).
There are some examples of surveys both in the United States and in Europe asking respondents to consent to linking their data to administrative records. In the US, a well-known example is the Health and Retirement Study (HRS), where respondents were asked for their consent to link their survey data to data from the Social Security Administration (Olson, 1999). In the UK both the English Longitudinal Study of Ageing (Lessof et al., 2004) and the ISMIE-Survey, a sub-sample of the European Community Household Panel (ECHP), link their survey data to administrative records from social security and employer records (Jäckle et al., 2004).

The number of surveys that directly link survey and administrative data in Germany is increasing, mainly because the "Research Data Centers" ("Forschungsdatenzentren", FDZ), providing administrative records for research purposes, were established in 2001 (Gramlich et al., 2010). For example, the Panel Study "Labour Market and Social Security" (PASS) linked survey data with administrative records from the German Federal Employment Agency during its initial wave, conducted in 2006/2007 (Trappmann et al., 2009); the same records were linked with the ALWA survey ("Arbeiten und Lernen im Wandel"; Antoni and Seth, 2012) as well as with the lidA-survey ("leben in der Arbeit"; Tisch and Tophoven, 2011).

Even though the regulatory framework is different in different countries, there is one similarity to all these studies: without the respondents' explicit and informed consent (written or verbal), the linkage of a person's survey data with their administrative records is generally not possible. The consent decision, specifically a refusal to consent, leads to methodological complications, because the sample size of usable combined data decreases and consent bias may be an issue if there are systematic differences between those individuals who consent and those who do not. Consequently, understanding the mechanisms behind the consent decision is important for determining the sources of possible biases and reducing their influence in the future.

This chapter investigates the determinants of consent to record linkage in the German part of the Survey of Health, Ageing and Retirement in Europe, SHARE, where in

[7] Specifically, the distance measure relies on the conditional independence assumption: conditionally on the variables identical in both datasets, the remaining non-overlapping variables have to be independent (e.g. Rässler, 2002; D'Orazio et al., 2006). The fewer variables overlap, the less likely is the assumption to hold.

2008/2009 the pilot study SHARE-RV was conducted to link SHARE with administrative records of the "Deutsche Rentenversicherung" (DRV), the German Pension Fund (see Chapter 2.2). Our results show that while there are effects at the respondent level that determine consent, interviewers are important to the consent decision as well. Section 3.2 presents a brief overview of the previous literature on consent to data linkage, followed by a description of the linked datasets and the linkage procedure in Section 3.3. Section 3.4 develops a model of consent and shows how determinants of consent are measured. The empirical results follow in Section 3.5, while Section 3.6 concludes with a brief summary and a discussion of the findings.

3.2 Previous Research

Systematic research on the differences between consenting and non-consenting respondents is not widespread (Sala, Burton, and Knies, 2012) and is typically found in medical and epidemiological studies (Jenkins et al., 2006). The majority of studies analyze respondent characteristics such as demographics (like age and gender), health status and socioeconomic factors (like education and income), finding some significant differences (e.g. Woolf et al., 2000; Dunn et al., 2004; Kho et al., 2009). Dunn et al. (2004) analyzed data from seven epidemiological mail surveys conducted in the UK, which all contained demographic, disease-specific, and generic items. They considered consent to follow up and/or the review of medical records and found effects of age, gender as well as the symptoms under investigation. In another meta-analysis, Kho et al. (2009) report statistical differences with respect to respondents characteristics (i.e. age, sex, race, income, education and health status) between consenters and non-consenters when using data from 17 unique medical studies, where the influences differ between the studies in direction and magnitude. There is also evidence for an impact of other socio-economic factors such as area effects (e.g. Huang et al., 2007). It is not clear whether the results from these medical studies can be easily transferred to consent questions covering different topics. In addition, the studies mentioned above are all limited to influences of respondents' characteristics.

Early work in the social sciences about potential selection bias in linked datasets is based on the Health and Retirement Study (HRS), where data was linked to administrative records from the Social Security Administration (SSA) in 1992. The analyses provide evidence of a consent bias related to respondent characteristics like age, race, gender, income or education (e.g. Olson, 1999; Gustman and Steinmeier, 2004; Haider and Solon, 2000).

More recent studies also take into account information about the survey design, the behavior of the respondent during the interview and the influence of the interviewers. Jenkins et al. (2006) analyzed two different consent questions in a large survey originally

based on the British part of the ECHP, the ISMIE ("Improving Survey Measurement of Income and Employment"). The authors' findings confirm that there are differences in demographics between consenters and non-consenters, additionally showing that the interview situation is important for the consent decision. For example, respondents with problems during the previous interview are less likely to consent to the linkage with administrative data, while respondents with a longer interview in the previous wave are more likely to consent. When analyzing two different consent questions (consent to record linkage and consent to contact the employer), the authors find that the influences vary depending on the context of the consent question.

In two recent studies, Sakshaug et al. (2012) defined "resistance indicator", that are correlated to the consent decision not just when asking for permission to link the data (Sakshaug et al., 2012), but also when asking for consent to take physical measurements in a survey (Sakshaug et al., 2010). Considering the link of the HRS with administrative data from the SSA, the authors found negative effects corresponding to the number of financial questions the respondents refused to answer both during the current and prior interviews. Respondents who expressed confidentiality concerns in the previous wave are less likely to consent as are those who were rated less cooperative or less attentive by the interviewer. The number of call attempts (current and previous wave) prior to the interview is negatively associated with the consent rate as well.

In addition to these resistance indicators, another important extension of these two studies is the inclusion of interviewer characteristics as well as an estimation of the interviewer level variation. The interviewers' education and experience as an interviewer are both negatively associated with the consent to record linkage. Consent for physical measurements is affected only by the interviewers' race. However, a significant interviewer variance term for both consent questions indicates that additional (not measured) interviewer characteristics are influencing the consent decision.

Sala et al. (2012) investigated the influences on consent in another study based on the British Household Panel Study. Using an interviewer survey, the authors were able to add information on the interviewer level and test its influence on the consent decision for linking survey data to health records and to social security benefit records. Respondents' demographics are not strongly associated with consent, while attitudes toward privacy and community-mindedness seem to be of greater importance. Respondents participating in the panel for a longer time are less likely to give consent, whereas the collected interviewer characteristics (attitudes and personality traits) do not have significant effects. However, the authors found "intra-household dynamics" such that each respondent's decision to consent is "located within the interaction between the individual, the interviewer and the wider household context" (Sala et al., 2012, p. 19).

In Germany, the Institute for Employment Research (IAB) asks respondents for their consent to linkage with administrative records from the German Federal Employment Agency in several surveys. Hartmann and Krug (2009), Beste (2011), and Antoni (2011) analyzed the consent decision in different studies (the so-called "Mainzer Modell", PASS, and ALWA, respectively), reporting the influence of the interviewers. Beyond effects of respondent and interviewer characteristics as well as factors of the interview situation, Antoni (2011) finds a significant interaction effect of respondent and interviewer age: interviewers who are at least 10 years older than their respondents are less successful in obtaining consent.

This overview of studies investigating the determinants of the consent decision shows the growing number of surveys in the social sciences that ask for the respondents' consent to data linkage in various topics. When analyzing determinants of consent, it is important to not only take into account respondent characteristics, but also include indicators of the interview situation as well as interviewer variables, as they are important parameters of the consent decision. There are some general results: in all studies, respondent characteristics turned out to be significant predictors of consent, thus evidencing the existence of a consent bias. Respondents who are more cooperative are also more likely to consent, while problems during the interview (also in previous interviews) reduce the likelihood of consent. Most studies control for interviewer characteristics to some degree, but the results are not definite.[8] Our study adds to the literature by providing additional evidence on the respondents' selectivity in the consent decision. The findings also stress the importance of the interviewer in obtaining consent. Moreover, we advance previous studies by testing different multi-level models to quantify and explain the interviewer's proportion of the variance. In addition, the consideration of interviewer quality and performance measures may help survey agencies in training and selecting interviewers who will increase consent rates.

3.3 Data Linkage in SHARE

The linkage reported in this chapter covers the German sub-sample of SHARE's third wave (SHARELIFE), where the pilot study for other SHARE countries was conducted (see Chapter 2.2; for a project overview also see Czaplicki and Korbmacher, 2010; Korbmacher and Czaplicki, 2013).

The administrative records of the German Pension Fund ("Deutsche Rentenversicherung" , DRV) constitute - for the most part - the universe of all Germans paying into the social security system. People are not included in this database if they have always been self-employed, worked only as civil servants, or have never worked and have not accu-

[8] Antoni (2011) provides a nice tabular overview of the literature, which is replicated in this chapter with the author's permission in Appendix A.

mulated any social security entitlements through other activities. For all others (nearly 90% of the German population, see Rehfeld and Mika, 2006), the data contain monthly information about the respondents' work history beginning at the age of 14. In addition to some basic demographics, detailed information about the employment status (e.g. working, unemployed, in care, disabled) as well as the personal retirement entitlements is included (Mika and Czaplicki, 2010).

All German SHARE respondents were asked for consent to link their survey data with their DRV records. The linking procedure in Germany, conducted via the Social Security Number (SSN), is tied to strict data protection rules. Consent to linking SHARE data with DRV data must be given in written form by each respondent. There are two steps to consent: First, all German SHARE respondents are asked verbally for permission to link their data at the end of the CAPI interview. If the respondents gives their consent, the interviewer provides a consent form to collect the SSN, all information needed to check (and if necessary construct) the SSN, as well as the respondent's signature. The second step is completed by the respondent, who fills out the form and mails it back to the DRV.

Each step presents a hurdle along the way to the data linkage. Conditional on participating in the interview, a respondent may decline consent directly to the interviewer, may fill in an incorrect SSN, omit it, or may not send in the consent letter at all.[9] The analyses of this paper only consider the initial step of consent, as it is similar to the decisions in many other studies in Germany (e.g. PASS, ALWA, LiDA). Further, as we are interested in the influence of the interviewer on the consent decision, the first step is more appropriate to use than the second. Based on release 1.0.0 of SHARELIFE, 1,350 (73%) of the 1,844 respondents with complete interviews gave their verbal consent to link their survey data to the DRV records, 21% (390) refused consent or claimed "don't know", and the remaining 6% (104) stated that they do not have any entitlements from the German Pension Fund.[10] This leads to a consent rate among the eligible of 77.6% (see also Table 1 below), which is lower than in ALWA (91.6%; Antoni and Seth, 2012), but similar to PASS (79.8%; Beste, 2011). In the BHPS, where only written consent is asked for, the rates are between 32 and 41% (Sala et al., 2012).

[9]The DRV checks and if necessary corrects the SSN, and if missing, constructs it if all other information on the consent form is available.

[10]The 6% of respondents who claim that they do not have any entitlements at the DRV is lower than the expected 10% from the general population. However, given that SHARE is representative of the population 50+, it seems very likely that in this specific population the number of people not having paid into the social security system is lower.

3.4 Models and Methods

Groves and Couper assume that only a "few householders have strongly preformed decisions about survey requests" (Groves and Couper, 1998, page 32). The analyses in this chapter base on the assumption that the same holds for the consent decision. Asking the respondents to answer survey questions is different from asking them for consent to data linkage. Although the respondents receive information on the data in their administrative records, they may feel insecure about what exactly they are asked to consent to. In addition, there is no possibility for them to release only a portion of the administrative records. Finally, the respondents have no control about what was collected in the administrative records - they may know the contents, but they cannot change them. These characteristics of the consent decision clearly differentiate the consent question from "regular" survey questions. In fact, the decision to give consent may be viewed as being similar to the decision to participate in a survey, where respondents cooperate without knowing the exact questionnaire.

To model the consent decision, the "conceptual framework for survey cooperation", developed by Groves and Couper (1998), is slightly adapted in Figure 2, depicting aspects influencing the respondent's decision to cooperate or refuse when asking for consent. The respondents' consent decision is the result of several influences channelled through three different groups: social environment (such as the household settings) and respondent characteristics (such as age, gender or personality) in the left column, survey design (such as topic, length or mode) and interviewer characteristics (again age, gender or personality) in the right column, and the interaction between respondent and interviewer in the middle column. The consent decision is the result of influences of some areas on others, as depicted by the arrows. The conceptual framework highlights the fact that the interviewers are an important factor in the process, especially because they are "under researcher control".

Unit nonresponse analysis, a central topic in Groves and Couper (1998), usually lacks sufficient data to test the theoretical hypothesis of what influences participation behavior. SHARELIFE allows for using a full set of control variables from the interview in the third wave as well as from the previous two waves, to investigate the determinants of non-consent. The estimation models in this paper follow the three columns of the framework in Figure 2 in dividing the variables in three areas of influence (described in detail below):

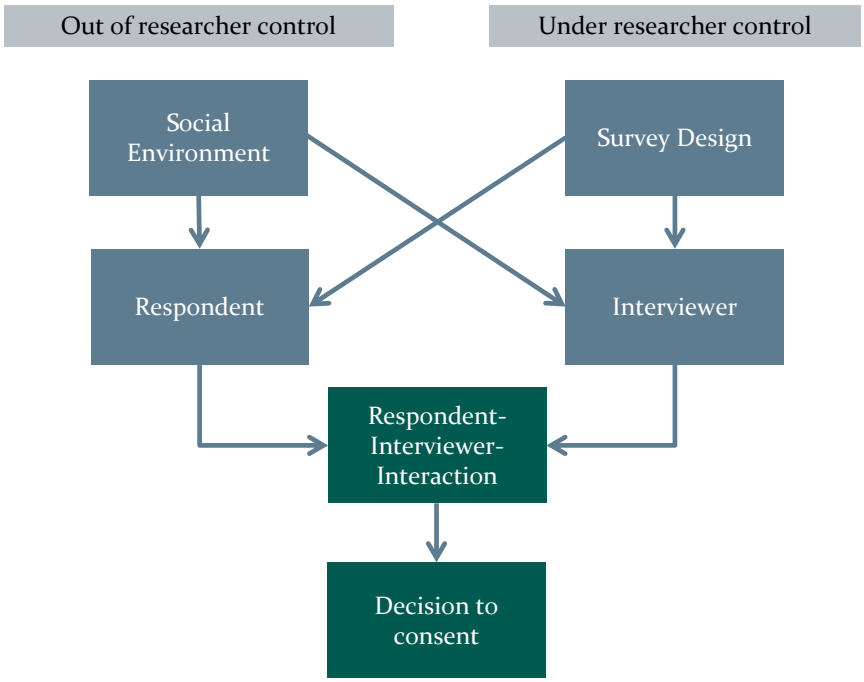


Figure 2: Decision to Consent (Adapted from Figure 2.3 in Groves and Couper, 1998)

- First, variables describing the respondent, including personal characteristics as well as household and environmental determinants.
- Second, as the respondent-interviewer interaction is not directly measurable, aspects of the interview situation including paradata[11] are used as proxies.
- Third, characteristics of the interviewers including both personal information and aspects related to the interviewer's quality. (Given that the survey design is constant for all respondents and interviewers in SHARE, any influence of the survey design on consent cannot be considered in the analyses.)

The choice of variables characterizing the respondents is mainly motivated by findings in other studies. Although contradictory results are reported for gender, age, and years of education, the influences are shown to be significant. Both age and age-square are included in order to control for a u-shaped relationship. The respondents' partner status influences the consent decision in the majority of studies controlling for it and is mea-

[11]There is no exact definition of what the term "paradata" includes, but following the definition of Kreuter (2013) paradata are defined "as additional data that can be captured during the process of producing a survey statistic" (Kreuter, 2013, page 3) For SHARE these data are for example timestamps, interviewers' observations as well as the contact protocol.

sured here with three indicators for currently living with a partner, having ever been married and having ever been divorced. The current work status is influential in three of six studies. The indicator variable used here differentiates between people currently working and all others (retirees, unemployed, housewives). The total number of jobs is added in order to capture the respondents' survey burden in the third wave, as they were asked detailed retrospective questions about each of their past jobs. Respondents with a large number of jobs may see the benefit of consent, as it potentially reduces the time spent on job related questions in the future. Income quartiles of the household's equivalence income (net income divided by the square root of household size), as well as an indicator if household income is not reported are added to capture differences in socio-economic status.[12] An indicator whether the respondent lived in East Germany during communist times is included, as these respondents may feel less confident in their government and hence may be less likely to consent. To control for some environmental influences, information about the household's location (urban or rural), the household's building type (more than two units or a 1- or 2-family home) and whether there is a foreign-born person in the household are used.

To describe the interview situation, some variables originating from paradata are included. Whether the interviewer had been in the household during previous waves can play a role, if knowing the interviewer increases the respondent's trust in the confidentiality of the record linkage. In addition, the interviewer's assessment of how well the respondent understood the questions and needed clarifications is used, where a dummy variable with the value 1 is created if the respondent never asked for clarifications and always (to the interviewer's knowledge) understood the questions. People with problems understanding questions may be less likely to consent because the decision costs are higher. The duration of the interview was found to significantly influence the consent decision in one previous study (Jenkins et al., 2006). To include interview time here, the average time per question was split into a respondent and an interviewer part. The average of how long a respondent takes to answer a question net of the interviewer's average time per question is included to measure the effect of slow or fast respondents independently of the interviewer. Jenkins et al. (2006) found a positive influence of interview time, indicating that using more time (and thus more effort) may be related to more committed respondents. On the other hand, people with more time per question may also be more sceptical and ask more questions, which could reduce the likelihood to consent. Therefore the "net" respondent time per question may influence the consent

[12] At the beginning of the interview a "financial respondent" is determined to answer all questions about the household's financial situation (see Chapter 2.1). Therefore the two variables "equivalent income" and "missing income" are measured on the household level. Note that because the sample is restricted to the first person in each household (see below), using variables at the household level does not introduce a distortion of the variance in these variables.

decision in both directions. The respondent's willingness to answer is measured by the proportion of missing answers (number of missing answers for every hundred questions). This value is split by questions directed at financial values and all other questions, as respondents may view financial questions to be more sensitive and hence more similar to the consent question.[13] To capture possible learning effects, the interviewer's experience with SHARELIFE is included, where a simple counter from the first to the last interview the interviewer conducts was transformed into five categories.

The inclusion of all available interviewer characteristics reduces the number of observations (see Table 1), because the information on interviewer demographics (gender, age, and education) provided by the survey agency was collected after the data collection was finished. Information on age and education is missing for those interviewers, who had left the agency by the time the information was requested. The age of the interviewer enters as a second order polynomial, while education is provided in three mutually exclusive categories: completed high school ($12^{\text{th}}/13^{\text{th}}$ grade), the highest degree; left school after 10^{th} grade; and left school after 9^{th} grade, the lowest category, which is selected as the reference group. For gender and education the findings from other studies are contradictory, while interviewer age shows a significant positive effect in most studies.

In addition to these demographics, two variables were constructed from the current wave for each interviewer. The first is an attempt to measure interviewer "quality": SHARE conducts grip strength measures in every wave, where each respondent is required to conduct two measures of their grip strength with each hand. Interviewers are asked to record these measures ranging from 0 to 100, with the explicit instruction not to round these numbers, because previous waves showed that multiples of 5 and 10 were recorded more than statistically expected. Based on the total number of grip strength measures an interviewer conducted, a 90% confidence interval was constructed around the 20.8%, which are expected if no heaping on multiples of 5 and 10 has occurred.[14] If the actual percentage for an interviewer lies outside this interval, the underlying assumption is that the interviewer is still rounding (if above the upper cutoff point of the confidence interval) or that she is overdoing the nonrounding (if below the lower cutoff point).[15] Two

[13] Differently to Antoni (2011), the missing answer rates combine "refusals" and "don't know"-answers because respondents may use "don't know"-answers to mask a refusal.

[14] The underlying assumption regarding the distribution of digits from 0 to 9 is that it is uniform on the grip strength distribution's support from 0 to 100, which means an expectation of eleven "0"s and ten "5"s, i.e. a fraction of 21 of 101 numbers should be "0" or "5".

[15] As an example: Suppose an interviewer has done 25 interviews with 100 grip strength measures. This creates a confidence interval of: $20.8\% \pm 1.645\sqrt{[20.8\%(1\text{-}20.8\%)]/\sqrt{(100)}}$=[14.3%; 27.7%]. Thus, if the interviewer has 14 or fewer multiples of 5 and 10, the indicator for "too few multiples of 5 and 10" is set to 1. If the interviewer has more than 27 multiples of 5 and 10, the indicator for "too many multiples of 5 and 10" is set to 1. More interviews and thus more grip strength measures reduce the confidence interval. The minimum number of conducted grip strength measures of the interviewers used

dummy variables - one for being above and one for being below the expected cutoffs of the confidence interval are added to the model. An additional variable of interviewer performance is the average time the interviewer took per question over all cases that she interviewed. Although longer does not necessarily mean better in this case, the assumption is that interviewers who have smaller values in this variable are less thorough when reading the question texts. As mentioned before, the inclusion of interviewers' characteristics reduces the number of cases of analysis.

Table 1 gives an overview of the stepwise reduction from the complete into the final sample used in the estimations. The initial sample (SHARELIFE, Release 1.0.0, Germany only) consists of 1,852 interviews, of which eight respondents did not complete the interview. 104 respondents claim that they do not have a record in the DRV data and are therefore excluded. To separate interviewer and "contagion effects" from previous consent decisions within the household, the sample only consists of the first respondent who is asked for consent in a household. Of the 64 interviewers working the sample, 12 (19%) dropped out after the wave, further reducing the sample by 122 (10%) cases. A final reduction by 5 cases is introduced through item-nonresponse. The final sample consists of 1,055 respondents and 51 interviewers, who each interviewed between 7 and 51 respondents in the estimation sample. There are some differences in the consent rates between the cases dropped and the respective remaining sample (shown in parentheses in Table 1), but none are significant.

in the analyses is 34, so even though the measure is less precise for interviewers with fewer interviews, the differences in confidence intervals are not large. Note that the underlying assumption is that the standard errors are not clustered on the respondent level (this would likely increase the standard error). Given that the interviewer's measurement is the variable of interest, this assumption is not far-fetched.

Table 1: Sample Size Development, Consent Rates and Number of Interviewers SHARELIFE 1.0.0, Germany

	Cases Dropped (% Consent Rate)	Reduced Sample (% Consent Rate)	Number of Interviewers
Full Sample		1,852	64
Incomplete Interview	8	1,844	64
	(-)	(-)	
Linkage not applicable	104	1,740	64
	(-)	(77.6)	
Second HH respondent	558	1,182	64
	(76.9)	(77.9)	
Missing interviewer info	122	1,060	52
	(85.3)	(77.1)	
Item nonresponse	5	1,055	51
	(60.0)	(77.2)	
Final sample	797	1,055	51

Notes: The table shows the development from the full sample to estimation sample. The consent rate percentages refer to the dropped or remaining cases only.
No consent rates are provided where some or all observations are not eligible for the consent question.

The dependent variable is verbal consent to record linkage. Interviewers are requested to ask respondents for consent and record their answer in the CAPI program. The dependent variable is 1 if the respondent consents to the linkage of the survey data with administrative records and 0 otherwise.

To take the dichotomous nature of the dependent variable and the hierarchical structure of the SHARE data into account, a multilevel logistic regression is used. Two different levels are distinguished: the respondents (level 1) who are nested within interviewers (level 2). The model is developed from the intercept-only (or "null") model to the final model using all variables. The outcome Y_{ij}^* for the respondent i interviewed by interviewer j is explained as the regression intercept α, the residual at the interviewer level u_j (the random intercept, which is assumed to be normally distributed, $u_j \sim N(0, \sigma^2)$), and the respondent level residual ϵ_{ij} (see Hox, 2010). More information about the assumptions of that model are discussed in Appendix D.1. The intercept-only model can be written as:

$$Y_{ij}^* = \alpha + u_j + \epsilon_{ij} \tag{1}$$

$$Y = \mathbb{1}(Y_{ij}^* \geq 0) \tag{2}$$

As the dependent variable Y_{ij}^* is a dummy variable which can either be 1 or 0, Y equals 1 if the latent variable Y_{ij}^* is greater than or equal to zero. This intercept-only model provides an estimate of the intraclass correlation (ICC) ρ, which is the proportion of variance at the highest level compared to the overall variance. The ICC is calculated as

the variance of the residuals at the interviewer level $\sigma^2_{u_j}$ divided by the total variance $(\sigma^2_{u_j} + \sigma^2_{\epsilon_{ij}})$. Given that the respondent level variance $\sigma^2_{\epsilon_{ij}}$ is not normally distributed but logistically, this term is fixed at $\frac{\pi^2}{3}$. Therefore the intraclass correlation can be written as:

$$\rho = \frac{\sigma^2_{u_j}}{\sigma^2_{u_j} + \frac{\pi^2}{3}} \tag{3}$$

The first set of variables is taken solely from the respondent's side (characteristics of the respondent and of the respondent's household):

$$Y^*_{ij} = \alpha + \gamma_p X_{pij} + u_j + \epsilon_{1ij} \tag{4}$$

where the X_{pij} are the p explanatory variables at the respondent level. Part of the respondent level variance of Model (1) is assumed to be explained by X_{pij} with ϵ_{1ij} being the remaining respondent level residual. At the next step, variables showing the respondent-interviewer interaction are included. As they all vary on the respondent level they are also included in X_{pij}. Finally, q variables describing the interviewer (Z_{qj}) are added. Therefore the final model reads as:

$$Y^*_{ij} = \alpha + \gamma_p X_{pij} + \gamma_q Z_{qj} + u_{1j} + \epsilon_{1ij} \tag{5}$$

Here, part of the interviewer level variance of Model (4) is assumed to be explained by Z_{qj} so that u_{1j} is the remaining interviewer level residual. Via the intraclass correlation this procedure determines how much of the interviewer's proportion of the total variance can be explained in each step toward the full model.

3.5 Estimation Results

Table 2 shows the results from the multilevel estimations, depicting the intra-class correlation for all models along with the corresponding χ^2-statistic from a test of the estimated multilevel model against a regular logistic regression. In all cases, the test rejects the simple logistic regression model. In addition, a χ^2-statistic is provided from a likelihood-ratio test of the current model against the one in the previous column. As mentioned above and shown in Table 1, the sample is reduced by those cases where the interviewer information is missing. To test for sample selectivity and possible bias, all models are re-estimated using the sample that includes the cases with missing interviewer information.[16] The results of these estimations as well as all other robustness checks are referred to in the text and presented in Appendix A.

[16] Of the 122 cases dropped because of missing interviewer information shown in Table 1, five observations need to be removed because of item nonresponse. Hence the sample including those cases without interviewer information amounts to 1,172 observations.

The first column of Table 2 shows the multilevel model without any explanatory variables. The intra-class correlation of 55.2% provides evidence of a very large interviewer influence on the consent decision. An intra-class correlation of 55.2% means, that about 55 % of the variance in the empty model is on the interviewer level. Column 2 of Table 2 shows the model including indicators for fourteen federal states ("Bundesländer") to correct for potential region effects.[17] The intra-class correlation drops slightly to 50.9%, showing that some of the interviewer variation can be attributed to variation at the state level.

The model is then augmented in column 3 by variables that solely depend on the respondent and are not influenced by the interviewer. In this regard, it is not surprising that the intra-class correlation remains almost identical at 50.5%. Including the additional variables is important, as the likelihood-ratio test against the previous model shows. The respondent's age has a significant inversely u-shaped influence on the consent decision. The peak age (from calculating the marginal effects) is at about 65 years, which is right at the official retirement age for the sample under investigation. It is very likely at this age that individuals have obtained most information about their retirement entitlements and the German Pension Fund, while older and younger groups face more uncertainty that reduces the willingness to consent.[18]

[17] Because the states of Bremen and Saarland each only have few observations, they are joined with adjacent states: Bremen and Lower Saxony receive the same state indicator, as do Saarland and Rhineland-Palatinate.

[18] Some of the retirees may also have been in direct contact with the DRV at this age, because the DRV attempts to validate the pension account information directly with the employees ("Kontenklärung") to assure that the pension benefits payments are correct (Rasner, 2012).

Table 2: Multilevel Estimations of the Consent Decision

	(1) Null Modell	(2) State Effects	(3) Resp. Char.	(4) Interview Situation	(5) Interviewer
Respondent Characteristics					
Age			1.401**	1.342**	1.336*
Age^2			0.997**	0.998**	0.998**
Female			1.122	1.127	1.133
Years of Education			1.015	0.998	0.995
Currently employed			0.820	0.859	0.831
Number of jobs			1.122*	1.094	1.104
Lives with Partner			1.875**	1.732**	1.739**
Ever married			1.182	1.352	1.281
Ever divorced			0.495**	0.544**	0.543**
Ever lived in GDR			4.849***	4.003**	3.923**
HH in urban area			0.636	0.554*	0.465**
HH in 1- or 2-family house			1.078	1.096	1.102
Foreigner in household			0.750	0.821	0.779
Income is missing			0.247***	0.486*	0.502*
1st income quartile			0.725	0.784	0.811
2nd income quartile			0.524*	0.507*	0.523*
3rd income quartile			0.728	0.732	0.742
Interview Situation					
Interviewer is known				0.827	0.767
Respondent comprehension				1.810**	1.809**
Seconds per question (net IVer)				1.007	1.014
Missing rate: financials				0.986***	0.987**
Missing rate: non-financials				0.769*	0.794
Interviewer's experience: int. 6-10				0.698	0.712
Interviewer's experience: int. 11-20				0.715	0.740
Interviewer's experience: int. 21-50				0.364***	0.348***
Interviewer's experience: int. 51+				0.230***	0.223***
Interviewer Characteristics					
Interviewer age					0.309***
Interviewer age^2					1.011***
Interviewer Education: high					4.103*
Interviewer Education: middle					4.208
Interviewer is male					1.060
Average seconds per question (I'wer)					1.167
Quality: too few multiples of 5					0.075**
Quality: too many multiples of 5					0.331
State ("Bundesländer") fixed effects	No	Yes	Yes	Yes	Yes
Intra-Class Correlation	0.552	0.509	0.505	0.458	0.352
χ^2 (2) vs. Logistic Regression	266.72***	139.92***	119.61***	64.28***	47.11***
χ^2 of LR-Test against previous model		14.99	74.54***	33.49***	21.33***
(degrees of freedom; p-value of LR-test)		(13; 0.308)	(17; 0.000)	(9; 0.000)	(8; 0.006)

Notes: *, **, *** mark significance on the 10, 5, 1 percent level, respectively
Dependent variable in all models is the dichotomous variable "consent to record linkage".
All models are estimated with 1,055 observations in a multilevel logistic regression with Stata's xtlogit command with a random intercept on the interviewer level. Coefficients are odds ratios.
χ^2-values are the respective test statistics.
Reference categories: Income: 4th income quartile; Interviewer education: low;
Experience: interview 1-5; Quality: rounding within confidence intervals (see text for details)

There are no significant differences between men and women, and neither education nor currently being employed have significant influences on the consent decision. The number of jobs a person had during the working life has a significantly positive influence, which may be related to the survey burden: because (in SHARELIFE) detailed information was asked about each of these jobs, individuals with more jobs may be more likely to see the benefit of record linkage to reduce future survey burden. Partnership and marital status also matter for consent: respondents living with a partner have 88% higher odds to consent, while having ever been divorced has a significantly negative effect of about the same magnitude (calculated as 1/0.495). Ever being married does not show any significant effect. Respondents who have spent some time in East Germany during communist times are much more likely to consent (the odds are increased by 385%). This effect has to be interpreted with the state indicators in mind, which implicitly control for currently living in the East.[19] Almost all who ever lived in the GDR still live in that area (87%, or 258 of 296), so the GDR variable captures the effect of those who moved from the East into the West, showing that this is a selective group compared to those who stayed.

Except for household income, none of the other variables describing the household situation (urban/rural, building size, foreigner) show any significant influence on the consent decision. Compared to the fourth income quartile as the reference category all income groups have a negative effect on consent, where only the middle group (2^{nd} quartile) shows a significant effect on the 10% level. As was expected, those who refuse to report their income (15% of the sample) are far less likely to consent to linking their data with administrative records.

Including those observations that have missing interviewer information does not change the results (see Appendix A, Table 23, column 1): even though some of the significant odds ratios change substantially, qualitatively the results are identical to the model in Table 2 column 3. The test of the indicator for missing interviewer information (bottom in Appendix A Table 23) shows that the sample is not selective with regard to respondent characteristics.

Column 4 of Table 2 shows the effects of the interview situation, which describe the interviewer-respondent interaction. The inclusion of these variables leads to a reduction in the intra-class correlation by five percentage points. Also, the likelihood-ratio test confirms that their inclusion is important. Knowing the interviewer from a previous

[19] With the re-unification five states (the so-called "neue Bundesländer") were joined with the former West Germany while keeping the old states unchanged in their boundaries. As a consistency check, the above estimation was conducted with a simple East/West indicator, which shows that the odds of not consenting are increased by 554% for those (still) living in the East. Leaving out the indicator for the change (ever lived in the GDR) shows that currently living in the East increases the odds of not consenting by 90%, although not significantly so (see Appendix Table 22).

interview is not significantly related to the consent probability. A positive interviewer assessment of the respondent's comprehension increases the odds of consent by 81% compared to those with a negative assessment. The respondent-specific time per question does not play a role in the consent decision, while - as expected - the rate of missing answers is negatively associated with the consent for both financial and non-financial questions. The categories of interviewer's field experience in SHARELIFE show that compared to the first five interviews, the 6^{th} to 10^{th} and 11^{th} to 20^{th} interview is less likely to lead to consent, although not significantly so. From the 21^{st} interview onwards the effect becomes larger and significant. It is likely that the "experience" variables capture two effects working against each other: a "reluctance" effect, which is increasing with the time elapsed in the field work, where respondents who are more reluctant to participate in the survey are also less likely to consent to record linkage. On the other hand, a "learning" effect can be assumed such that the more experience the interviewers have in asking the consent question, the more successful they should become. Here the positive learning effect is not larger than the negative reluctance effect for any measured level of experience.

Using the sample of all interviewers leaves the previous results almost unchanged, and the included indicator for missing interviewer information does also not show a significant influence (see Appendix A, Table 23, Column 2). As a further robustness check, the rates of missing values and the interviewer's assessment of the respondent's comprehension are taken from previous waves to counter the possible endogeneity of using the same waves variables. This reduces the sample by nine observations (0.9%) and leaves the results qualitatively identical, although some coefficients are no longer significant (see column 1 and 2 of Appendix A, Table 23). In addition, the assessment of how willing the respondent was to answer during the previous wave can be used (see column 3 of Appendix A, Table 23). The variable shows a highly significant effect on the consent decision for wave 3: a respondent with a high willingness to answer in the previous wave has 184% higher odds of agreeing to the record linkage.[20]

Turning to the explanatory power that the interviewer level variables provide, column 5 of Table 2 shows that not all considered variables turn out to have a significant effect on the consent decision. Overall, their inclusion is warranted (likelihood-ratio test statistic of 21.3 with eight degrees of freedom). Although the intraclass correlation drops by ten percentage points to 35.2%, the model cannot explain all of the interviewer variance. The age of the interviewer influences consent in a u-shaped way such that older interviewers are more effective in obtaining consent (the turning point calculated from

[20]Note that due to endogeneity, the wave 3 version of this question is not used, as the interviewer's assessment comes after the consent question and is thus not independent of the decision. Variables from previous waves are not included as regular variables in the analyses because of the required panel setting that would limit the use of this study in other contexts.

marginal effects is at 55 years). The education of the interviewer affects consent positively, but the estimates for the indicators are not very precise. The interviewer's gender is not significant. Interaction effects between interviewer and respondent education as well as interviewer and respondent gender do not have significant effects on the consent decision (results not shown). The measure of the average time an interviewer needs per question is not significantly related to consent, while the variables on rounding (included to measure interviewer quality) show a negative and significant effect.

When including the cases missing the interviewer information, the variables on education and age of the interviewer must be dropped. Nevertheless, the remaining effects are similarly estimated, although the interviewer gender effect becomes much larger and significant and the negative effect of the rounding is attenuated (see Appendix A, Table 23, Column 3). The ICC increases, showing that interviewer age and education explain part of the variation on the interviewer level. The effect of missing interviewer information, as shown by the indicator and the likelihood-ratio test, is negative, but not significant. Comparing the ICC across all models in Table 2 shows the importance of the interviewer level variables, as the ICC drops from an initial 55.2% to 35.2% in the final model, a reduction of 36%.[21] This reduction means that 36% of the interviewer level variance can be explained by the characteristics of the interviewer. But still, the unexplained interviewer level variance remains very large in the model, even with the inclusion of variables on that level. This is a strong indication that further unobserved heterogeneity among the interviewers matters in the consent decision.

So far, the results show that interviewers are crucial to obtaining consent. To assess the influences of interviewer performance in the field, some additional variables are now considered, using the same estimation sample. These variables were not included in the analyses before, because they are endogenous to the consent decision to some degree. The considered variables are for each interviewer, (i) the consent decision of the last person visited before coming to the current household; (ii) the consent rate of all previously visited households; (iii) the response rate of all cases previously contacted; and (iv) the overall response rate of this interviewer. The first variable shows the immediate impact of having been successful in the previous household, capturing any boost in motivation to gain consent in the next household. The second variable captures the mix of convincing strategies, perseverance and other interviewer personality traits, which are unobserved but play a role in the consent decision. The third and fourth variables provide a measure of interviewer quality, which is not directly related to the consent decision

[21] As the scale of the outcome variable changes when variables are added to the model in logistic regressions, the comparison of regression coefficients and variance components is difficult (Blom et al., 2011). I rescaled the variance components of the full model to the metric of the empty model as described in Hox (2010) to calculate the ICC, but this does only change the results appreciably. The ICC of the full model is reduced by 3 percentage points. For detail, see Appendix D.2.

but may affect interviewer motivation.[22] All of these variables, especially (i) and (ii), are endogenous, as the error term in the consent decision is likely to be correlated to the variables via unobserved interviewer characteristics. Still we believe that they increase the understanding of the problem at hand.

Table 3 shows the odds ratios of the four variables measuring interviewer performance estimated in four different models while leaving the rest of the model identical to Column 5 in Table 2. An interviewer's experience in the previous household spills over into the next one visited: If the last decision in the previous household was positive, the odds of obtaining consent from the first person in the current household increase by 106%. Interviewers who were more successful up to the current interview are also more likely to gain consent in the present household: a one-percentage-point higher consent rate up to the current interview yields three percent higher odds of obtaining consent. The inclusion of this variable explains almost all of the interviewer variation in the model, as the intra-class correlation drops to 7.7% and is no longer significant (shown by χ^2-statistic). The consent rate picks up otherwise unobserved variation among interviewers, indicating which interviewers are good at obtaining respondent consent. This may be important information for fieldwork agencies, as this rather obvious relationship (high consent rate equals good interviewers) holds up controlling for a whole set of other variables. Assuming that respondent differences are controlled for, this variable allows survey agencies to identify and react to differences in the interviewers' abilities during the fieldwork period. The response rates for individual interviewers lead to different results: both measures (response rate up to the current interview and total response rate) have a negative correlation with consent, where only the total response rate has a significant influence. The ICC in the two models remains significant, such that the inclusion of these variables does not explain much in the unknown interviewer variation determining consent. The coefficients could be taken as an overall performance measure for interviewers, where interviewers with a one-percentage-point higher response rate have odds to obtain consent decreased by almost four percent. However, one has to be careful: interviewers with a high response rates will have convinced more respondents than those with a low response rate. This will also include more reluctant respondents, such that gaining consent is more difficult in such a sample.

[22]The response rate up to the interview is somewhat imprecise, as it is not clear if the interviewer will not contact a household again and "convert" formerly non-cooperating respondents. Nevertheless, interviewers may still be influenced by the success they had prior the current interview.

Table 3: Assessing Interviewer Performance Indicators on the Consent Decision

	Odds Ratio	Intraclass Correlation	n	χ^2
(i) Last person visited before current household gave consent	2.060***	0.334	1004	32.0***
(ii) Consent rate when entering current household	1.031***	0.077	1004	1.0
(iii) Response rate when entering current household	0.994	0.395	1023	66.3***
(iv) Total response rate over all assigned cases	0.963*	0.366	1055	39.5***

Notes: *, **, *** mark significance on the 10, 5, 1 percent level, respectively.
Dependent variable in all models is the dichotomous variable "consent to record linkage".
All models are estimated in a multilevel logistic regression with Stata's xtlogit command with a random intercept on the interviewer level.
Coefficients are odds ratios. χ^2-values refer to the test statistic against a regular logistic model.
Differences in sample size are due to construction of the variables: For variables (i) and (ii), each first observation per interviewer has to be discarded. This is similar for variable (iii) , however, in 19 cases the interviewer had previously been in households not in this estimation sample.

3.6 Summary and Discussion

The analyses in this paper provide insights on what determines respondents to consent to a linkage of their survey data to administrative records. Using a theoretical framework adapted from Groves and Couper (1998), the results show that while some variables at the respondent level are important, the interviewer-respondent interaction and especially the interviewers are a main component in obtaining consent. Using multi-level estimations, the initial proportion the interviewers contribute to the whole variation can be reduced from 55% to 35% by including interviewer level variables such as age, education, and quality indicators. However, a large part of the interviewer variance remains unexplained, which is likely to be related to unobserved interviewer abilities, as additional analyses show.

As far as comparisons are possible, our findings fit well with the existent literature on explaining consent. For most of the respondent variables, the results are similar to the majority of studies, which find significant effects of age, little evidence of a gender or education bias, and a positive effect of being in a relationship. The rate of missing values in financial questions is almost always related to lower consent rates. The effect of the interviewer variables is similar: interviewer gender is not that relevant, while interviewer age is positively related to consent in most studies, which - according to the reported u-shape influence - holds true in the SHARE setting once the interviewer has reached a certain age. The interviewer's experience in the study has a negative effect on consent,

which is similar to Sala et al. (2012). The estimated proportion of the interviewer variance is large in our paper, but similar to other studies which estimate it: Beste (2011) finds an intra-class correlation of 28%, from Sakshaug et al. (2010), a value of 34% can be calculated, while Sakshaug et al. (2012) implies an ICC of 32%.[23]

There are some limitations to this study. One is the lack of an interpenetrated sample, which would be necessary to estimate pure interviewer effects (Bailar, 1983). The multistage clustered sampling in SHARE does not allow for distinguishing interviewer effects from sampling-point effects, because interviewers are not assigned at random to respondents (for details on sampling in SHARE, see Klevmarken et al., 2005). The inclusion of household and respondent characteristics as well as state fixed effects in the analyses minimizes the influence of sampling-point effects as much as possible. Additionally, two studies show that the interviewer-induced variance is greater than the variance component that comes from the different areas (Schnell and Kreuter 2005; O'Muircheartaigh and Campanelli 1999), so this paper's setting may also "benefit" from smaller area and larger interviewer effects.

The results lead to the conclusion that there is some consent bias in the sample, as certain respondent characteristics are important determinants of consent. With the focus on the consent bias, this paper addresses only one part of the total survey error, and does not relate it to other sources of error. A comparison with the attrition bias in SHARE's previous waves would in principle be possible, but the setup is not easily adapted to an attrition analysis, because, due to the construction of SHARELIFE, most of the variables used here are not available in the previous waves. Investigating the size and direction of consent bias in relation to attrition bias is clearly an important path for future research. The study by Sakshaug and Kreuter (2012) suggests that nonresponse biases and measurement errors are generally larger than nonconsent biases, while the direction of the bias is ambiguous.

The attrition process may also have influenced the sample composition, which could question how well the analyses extend to other studies in different contexts or different consent questions. But even though the selection could be problematic on the respondent level, the interviewer's importance for the consent decision is unlikely to vary. The German SHARE sample consists of two parts: those who are interviewed since 2004 and those from a refreshment sample drawn in 2006. Additional tests do not show a difference in consent when considering the respondents' participation time in the SHARE panel.

This paper focuses on "first consenters" in a household and, unlike Sala et al. (2012),

[23] Ideally, such a comparison would be done for the respective "empty" models, to compare the original degree of interviewer variation. However, neither of the papers provides such information, hence the full models have to be used for the comparison here.

does not consider intrahousehold dynamics. This restriction was applied to avoid contagion effects and measure the "pure" interviewer effect on the first consent decision in a household. Although the SHARE interviewer instructions call for interviews without additional persons present, there could be communication among household members during the first interview that influences the decision to consent. In such a case, the distinction between interviewer and household effects may not have been perfect. However, with the current data available, any such communication cannot be detected. Future research should thus expand the multilevel approach to the context of intra-household dynamics and investigate how both interact in their effect on the consent decision.

The analyses show that the interviewers are a main source of differences in the consent decision, which highlights the importance of interviewer training in general. Future research should investigate how training could reduce the effect of the interviewer in such a setting. In an ideal world (from a researcher's perspective), all interviewers would be trained such that there are no detectable interviewer effects. As this state will never be achieved, both researchers and survey institutes need information about interviewers' abilities in order to be able to identify important drivers of not only the participation decision but also the consent to link data sources. Future research should thus focus on obtaining this information - possibly through interviewer questionnaires - to use resources more effectively in increasing consent rates, reducing consent bias and improving the overall quality of survey data.

4 Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework

The results of Chapter 3 show the importance of the interviewer but also the need for more information on them. To close that gap, we developed and implemented an interviewer survey which will be described in detail in the following chapter.

This chapter is already published; Blom A. G. and J. M. Korbmacher (2013): Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework. Survey Methods: Insights from the Field, http://surveyinsights.org/?p=817

4.1 Introduction

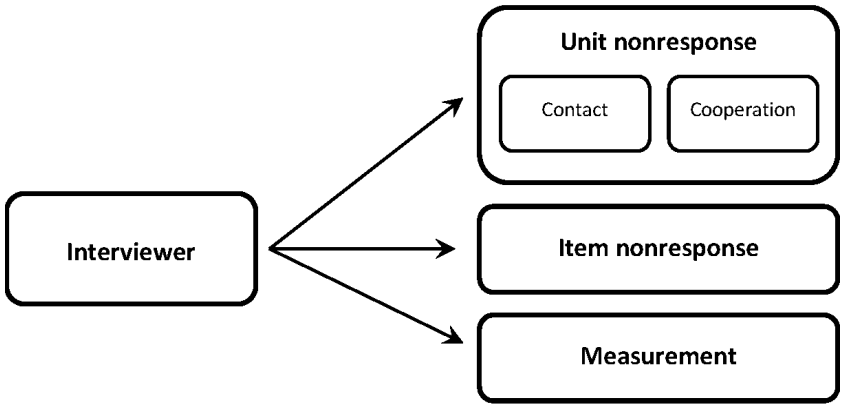
In all interviewer-mediated surveys interviewers play a crucial role during the entire data collection process. They make contact with and gain cooperation from the sample unit, ask survey questions, conduct measurements, record answers and measures, and maintain respondents' motivation throughout the interview (Schaeffer et al., 2010). As such, the job of an interviewer encompasses a diversity of roles and requires a variety of skills. Especially with the rise of computer-assisted interviewing, which permits the collection of even more complex data, a well-trained staff of interviewers has become indispensable.

When examining survey data we frequently find interviewer effects on all of these interviewer survey tasks indicating that there is variation in how interviewers handle their various responsibilities. Yet often, researchers are far removed from the interviewers and the actual survey operations (Koch et al., 2009) and have little or no information about what determines these interviewer effects. In fact, for the majority of survey data collection is contracted out and thus researchers have no influence on which interviewers work on their study and on how they were trained.

The literature describing interviewer effects on various aspects of the survey process is substantial (for an overview see Schaeffer et al., 2010, chapter 13). However, only few studies have succeeded in explaining the interviewer effects found (cf. Jäckle et al., 2013). One possible reason for this research gap is the lack of information on the interviewer level, which is necessary for identifying determinants of interviewer effects. In the past years, paradata (Couper and Lyberg, 2005) have been increasingly used to explain interviewer effects. Another potentially powerful source of auxiliary data is interviewer characteristics collected through an interviewer survey.

This chapter presents the conceptual framework of a new international interviewer questionnaire to explain interviewer effects. We specifically focus on interviewer effects other than interviewer falsification, since we believe that the latter cannot be explained by

Figure 3: Types of Interviewer Effects in Surveys



means of interviewer surveys. Furthermore, we developed an interviewer questionnaire for researchers who contract out fieldwork. Survey agencies aiming to identify suitable interviewers through an assessment might find a different questionnaire more appropriate. The questionnaire was developed in cooperation with researchers across various survey projects and will thus be relevant to survey projects across countries and disciplines.

This chapter consists of three parts. First, a theoretical background and literature review outlines the main aspects of the data collection process affected by interviewer effects (Section 4.2). Section 4.3 provides the subsequent conceptual framework, which constitutes the core of the chapter. Finally, the last section (4.4) presents findings on the variation of interviewer characteristics collected with the new interviewer questionnaire and implemented on the Survey of Health, Ageing, and Retirement in Europe (SHARE) in Germany in 2011. These results show that the survey well discriminates between interviewers, which is a prerequisite for explaining interviewer effects in survey data.

4.2 Theoretical Background and Literature Overview

Exposing interviewer effects implies that outcomes of sample units assigned to the same interviewer are more similar than would be expected if variation were random.[24] Three main types of interviewer effects can be distinguished: interviewer effects on the unit nonresponse process, on item nonresponse and on the actual measurement (Figure 3).

[24] An interviewer effect is typically estimated by an intraclass correlation coefficient (ICC), i.e. the ratio of the interviewer variance to the sum of all variances in the model (e.g. Anderson and Aitkin, 1985; Groves and Magilavy, 1986). The ICC allows us to estimate to which extent the variation across respondents in the survey estimate is clustered within the interviewers conducting the survey.

Interviewer Effects on Unit Nonresponse

When considering the unit nonresponse process we find that interviewers are differentially successful at recruiting sample units leading to differential unit response rates. A growing literature has examined the role of the interviewer in the nonresponse process and attention has been paid to interviewer attributes, such as experience (Durbin and Stuart, 1951; Couper and Groves, 1992; Singer et al., 1983; Snijkers et al., 1999; Olson and Peytchev, 2007; Lipps and Pollien, 2011; Jäckle et al., 2013), interviewer skills (Morton-Williams, 1993; Campanelli et al., 1997), as well as survey design characteristics, such as interviewer burden (Japec, 2007) and interviewer payment (De Heer, 1999; Durrant, Groves, Staetsky, and Steele, 2010). To explain differential response rates across interviewers survey methodologists have examined interviewer attitudes and motivation (Campanelli, Sturgis, and Purdon, 1997; Groves and Couper, 1998; Hox and de Leeuw, 2002; Durrant, Groves, Staetsky, and Steele, 2010; Blom, de Leeuw, and Hox, 2011; Jäckle, Lynn, Sinibaldi, and Tipping, 2013). This strand of research was inspired by the work of Lehtonen (1996), who developed a short interviewer attitudes scale and showed that attitudes correlate with attained response rate. Another line of studies focuses on interviewer behavior and interviewer-respondent interaction (Couper and Groves, 1992; Campanelli et al., 1997; Groves and Couper, 1998; Snijkers et al., 1999; Conrad et al., 2013; Broome, 2014). This started with the pioneering work of Morton-Williams (1993), who analysed tape recordings of survey introductions and identified successful interviewer strategies, such as, using professional and social skills, and adapting these to the doorstep situation.

Interviewer Effects on Item Nonresponse

In addition, interviewers have an influence on item nonresponse, i.e. on the respondents' willingness to answer each question in the survey and on their consent to providing additional information. The consent to the collection of additional information can be diverse; typical examples are consent to record linkage (Lessof, 2009; Calderwood and Lessof, 2009; Sakshaug, Couper, Ofstedal, and Weir, 2012; Sala, Burton, and Knies, 2012; Korbmacher and Schröder, 2013; Sakshaug, Tutz, and Kreuter, 2013) and consent to the collection of biomarkers in health surveys (Sakshaug, Couper, and Ofstedal, 2010).

Traditionally, the literature on interviewer effects on item response rates describes a clustering effect of item nonresponse within interviewers and tries to model these interviewer effects by demographic characteristics of the interviewer (Singer et al., 1983). Another strand of research looks into collecting additional information about the interviewers, for example on their expectations, by means of interviewer questionnaires (Singer and Kohnke-Aguirre, 1979; Singer et al., 1983).

Interviewer Effects on Measurement

Finally, interviewers can through their observable characteristics and their actions influence the measurement itself, i.e. which answer a respondent provides. Theory related to this third type of interviewer effect typically stems from the literature on respondents' cognitive processes when answering survey questions (Tourangeau et al., 2000). This process is complex and iterates through various stages, which may be influenced by the interviewers (Cannell et al., 1981; Tourangeau et al., 2000). Since survey questions differ widely in content and structure and since interviewer effects are estimate-specific, they can be different for different questions and topics (Schaeffer et al., 2010) and cannot be generalized for all measurements within a survey. Covering all of these different types of interviewer effects on measurement goes beyond the scope of the conceptual framework developed in this chapter. Instead we focus on identifying interviewer characteristics potentially associated with interviewer effects on unit and item nonresponse. As described, there have been several previous attempts at explaining interviewer effects in survey data by means of interviewer surveys. However, the studies found that the predictive power of the variables collected on the interviewer questionnaires was low and explained only part of the observed variance (e.g. Hox and de Leeuw, 2002; Durrant et al., 2010; Blom et al., 2011). The conceptual framework of the interviewer questionnaire presented in this paper ties in with previous work with an important extension. Instead of focusing on interviewer demographics, which seldom prove significant in explaining interviewer effects (c.f. Singer et al., 1983), and avowed doorstep behavior, the questionnaire covers four dimensions of interviewer characteristics: Interviewers' attitudes towards the survey process, their own behavior regarding data collection requests, experiences with conducting certain types of surveys and measurements, and their expectations regarding the survey outcome.

4.3 Conceptual Framework

The goal of the new questionnaire is to implement an instrument measuring a wide range of interviewer characteristics, which have been shown relevant in previous studies (see the literature review in Chapter 4.2). In particular, we aim to find correlates of interviewer effects on various types of unit and item nonresponse.

The questionnaire covers all four dimensions of interviewer characteristics: interviewer attitudes towards the survey process, interviewers' own behavior regarding data collection requests, interviewers' experience with measurements, and interviewers' expectations regarding the survey outcome in terms of response rates. It consists of two parts. First, a battery of general items assumed to be associated with general unit and/or item nonresponse relevant across a variety of social surveys is considered. Second, various blocks of questions aim at explaining interviewer effects that were specific to the

fourth wave of SHARE Germany. These blocks may or may not apply to other surveys, which have a different survey design and focus on different research questions. The full questionnaire collects information on interviewer characteristics to explain five groups of interviewer effects: on unit nonresponse in general, on income nonresponse (as an example of item nonresponse), on unit nonresponse across different incentives groups of an experiment, on consent to the collection of four types of biomarkers, and on consent to record linkage. It is obvious that some items in this questionnaire focus on the SHARE survey, but it is not restricted to it. Segmenting the questionnaire along the five groups of interviewer effects also allows other surveys to implement the questionnaire by adopting the relevant elements. The conceptual framework is based on our own experiences at interviewer trainings on a diversity of studies, from findings in previous analyses of interviewer effects, and from consultations with survey methodologists on various European and US surveys. When aiming to explain interviewer effects by means of characteristics collected in an interviewer survey, the underlying assumption is that interviewers differentially impact on the data collection process, that this differential impact is related to their - conscious and subconscious - appearance and actions, and that these can be explained by characteristics collected in an interviewer survey. Table 4 displays the four dimensions measured in the interviewer questionnaire (rows) and the interviewer effects they aim to explain (columns). We expect the first three dimensions - attitudes towards the survey process, own behavior with regards to data collection requests, and experience with relevant types of measurements, to independently impact on the survey outcomes. The fourth dimension - interviewers' expectations regarding the survey outcome - is expected to be influenced by attitudes, behaviors, and experiences. The concepts covered by these four dimensions are described in the following. In addition, the interviewer survey collects general interviewer demographics and measures of interviewing experience. The question numbers cited in the following refer to the questions in the SHARE interviewer questionnaire (see Appendix B).

Interviewer Attitudes Towards the Survey Process

Interviewers that are good at making contact and gaining cooperation from the sample unit are usually good at tailoring their approach to the situation they find at the visited address (Morton-Williams, 1993). However, tailoring takes more effort and skills than repeating the same routine with each sample unit. The extent to which interviewers make the effort of tailoring their approach might be related to their general attitudes towards their job as interviewers and towards life in general. In addition, interviewers' own concerns about data protection and their trust in other people might shape the way they approach sample units and ask their respondents for sensitive information.

Table 4: Conceptual Framework of the Interviewer Questionnaire

	General part		**SHARE-DE specific part**		
	Unit non response	Item non-response (income)	Unit non-response (incentives)	Consent to biomarker collection	Consent to record linkage
Attitudes	Q3: reasons for being an interviewer Q4: how to conduct standardized interviews Q5: how to achieve response Q6, Q11, Q12: trust, data protection concerns	Q4: how to conduct standardized interviews Q6, Q11, Q12: trust, data protection concerns		Q6, Q11, Q12: trust, data protection concerns	Q6, Q11, Q12: trust, data protection concerns
Own behavior	Q8, Q9: own survey participation Q27, Q28: use of internet social networks / online banking	Q27: use of internet social networks / online banking Q34: income response	Q10: incentives received	Q22: consent to biomarkers, hypothetical Q24: blood donation	Q13: data disclosure, hypothetical Q14, Q16: data linkage, hypothetical Q17: "pension records cleared" Q27, Q28: use of internet social networks / online banking
Experience with measure-ments	Q1, Q2: experience working as an interviewer Q18: SHARE experience	Q1, Q2: experience working as an interviewer Q18: SHARE experience	Q1, Q2: experience working as an interviewer Q18: SHARE experience	Q23: experience with collecting bloodspots	
Expecta-tions	Q19: effect of incentives on unit response	Q20: income response	Q19: effect of incentives on unit response	Q21: consent to biomarker	Q15: consent to data linkage

Note: The question numbering refers to the questions in the SHARE Germany interviewer survey (see Appendix B). Questions on the interviewers' demographic background are not displayed in the framework

This first dimension of general interviewer attitudes in the conceptual framework covers these aspects. Some of the attitudes collected in the interviewer questionnaire are related to the questions on previous interviewer questionnaires (e.g. De Leeuw and Hox, 2009). However, in addition to questions on the contacting and cooperation process, i.e. unit nonresponse, the SHARE interviewer questionnaire also collects information that might be related to item nonresponse and non-consent. The attitudes addressed are reasons for being an interviewer (Q3), attitudes towards under which circumstances it is legitimate to deviate from the standard interviewing protocols (Q4), how to best achieve unit response (Q5), and general questions regarding trust and data protection concerns (Q6, Q11 and Q12) that might be particularly effective in explaining non-consent and item nonresponse on income.

Interviewers' Own Behavior Regarding Data Collection Requests

The maxim "do as you would be done by" runs as a common theme through many cultures. Therefore, it is not difficult to imagine that survey requests, which interviewers themselves would not answer to, are difficult for them to sell to respondents. The second dimension of the conceptual framework thus assumes that the way interviewers behave or would behave, if faced with a similar situation as the respondent, influences the way they interact with the respondent. If interviewers participate in surveys themselves and supply all of the information asked from them, they are likely to be better at eliciting such information from their respondents.

A series of questions in the interviewer questionnaire covers interviewers' own behavior. These questions for example cover whether interviewers have taken part in surveys and, if so, what kind of surveys these were and whether they received any incentives (Q8, Q9 and Q10). Along a more general line, we examine how easily interviewers divulge information about themselves in their daily lives by asking about their membership in social networks like Facebook, Myspace or Twitter and their use of online banking (Q27, Q28). The questionnaire also asks about their income (Q34), to see whether item nonresponse on income on the interviewer questionnaire is correlated with item nonresponse among respondents to the SHARE survey. For measures of consent to the collection of biomarkers and consent to record linkage we inspect interviewers' actions in similar situations. The questionnaire asks whether the interviewer donates blood (Q24) and whether they have cleared their pension records ("Kontenklärung"), a process German citizens are asked to go through to ensure that the pension records that the state holds are correct (Q17). Finally, the questionnaire contains hypothetical questions on whether interviewers would disclose sensitive information (Q13), consent to record linkage (Q14 and Q16) and consent to the collection of biomarkers (Q22) if asked in an interview situation.

Interviewers' Experience with Measurements

Interviewers' familiarity with different types of surveys and measurements may influence their confidence in conducting these. This, in turn, may shape the professionalism with which they interact with the respondents. Interviewer training levels out some of the differences in experience with measurements; however, only up to a certain degree. If interviewers, for example, have previously worked on SHARE, they have more background knowledge about the content of the study, which is knowledge they may employ in their introduction. Likewise, if interviewers have experience with pricking a small needle into someone's finger for collecting blood spots in blood sugar tests, they are likely to feel more confident about collecting dried blood spots for biomarkers and to portray this confidence during the interview. The SHARE interviewers are diverse in the experiences that they have gathered on their job and in their life in general. Some Wave 4 SHARE interviewers have worked on all of the previous SHARE waves and are well used to the type of sample and the instrument. Others have conducted surveys that cover similar aspects as SHARE does.

The third dimension of the interviewer questionnaire, therefore, investigates interviewers' experiences with working as an interviewer (Q1 and Q2), with SHARE (Q18), and with conducting blood sugar tests for diabetics (Q23).

Interviewers' Expectations Regarding Survey Outcomes

Anecdotal evidence from interviewer trainings suggests that interviewers' perceptions about the viability of a survey are related to fieldwork outcomes. While implying a causal effect of interviewers' expectations on fieldwork outcomes would be far-fetched, in the context of explaining interviewer effects empirically testing whether interviewers who are confident about the success of a survey are also more likely to reach high response rates is informative.

The final dimension in the conceptual framework covers interviewers' expectations of unit nonresponse rates, consent rates and item nonresponse rates. The survey asks interviewers what response and consent rates they expect for the different incentives groups (Q19), for the various biomarker measurements (Q21), for consent to record linkage (Q15), and for the survey questions on income (Q20).

Alternative Conceptualization

When developing the interviewer questionnaire we opted for a general conceptualization of just four dimensions. We believe that dimensions one to three influence both the expectations interviewers' hold about their performance as well as their actual performance. As depicted in Table 4, we expect certain items within each dimension to be

correlated with only one of the survey outcomes, unit nonresponse, item nonresponse or biomarker / record-linkage non-consent, while others are expected to be associated with all types of nonresponse.

Our framework for explaining interviewer effects is just one of many possible conceptualizations. One recent interesting conceptualization, while not directly comparable to our approach, can be found in Jäckle et al. (2013). In their complex framework they model interviewers' influence on a sample persons' likelihood of cooperation in a survey as the interplay of household psychological predisposition, interviewer observable attributes and interviewer behavior. All of these are in turn influenced by a complex system of personality traits, interpersonal skills, expectations, experience, and socio-demographic characteristics. An alternative conceptualization of our framework might also go into more detail on the interrelatedness of interviewers' demographic characteristics, psychological predispositions, social environment, survey design, and the dimensions measured in the interviewer questionnaire. However, unlike other researchers involved with interviewer questionnaires previously, we consider various types of interviewer effects together. Through the complexity of a more detailed conceptual framework one might miss the wood for the trees. Nonetheless, when analysing processes leading to unit nonresponse, item nonresponse, non-consent to biomarkers or non-consent to record-linkage and considering interviewer effects thereupon, we recommend developing a specific and detailed conceptual framework for each process.

4.4 Variation Across Interviewers: Results From the 2011 SHARE Interviewer Survey

In early 2011 an interviewer questionnaire based on the conceptual framework described in this paper was implemented at the end of the interviewer training sessions for SHARE Germany. In total, 197 interviewers were trained. Participation in the interviewer survey was voluntary and interviewers did not receive any incentive for participating. 163 interviewers completed the questionnaire, yielding an 83% response rate. There was a negligible amount of item nonresponse and answers that were not codeable.

In addition, other large-scale social surveys implemented this or a similar interviewer questionnaire. Having presented and further developed the conceptual framework at the 2010 International Workshop on Household Survey Nonresponse in Nuremberg, Germany, several other studies showed interest fostering cooperation with survey methodologists across surveys and countries. At the end of 2010 the German PASS study (Panel Arbeitsmarkt und soziale Sicherung) at the Institute for Employment Research (IAB)[25] provided the first version of the questionnaire and implemented it online, with a 10 Euro conditional incentive and well before their interviewer trainings (see Kreuter et al.

[25] http://www.iab.de/780/section.aspx

(2014).

By 2012 the core of this interviewer survey has been implemented in at least three further large data collections: (1) a survey aimed at measuring the methodological effect of filter questions at the IAB, (2) the German part of the Programme for the International Assessment of Adult Competencies (PIAAC) at GESIS[26] and (3) the recruitment interview of the German Internet Panel (GIP), a longitudinal internet survey based on a face-to-face recruited probability sample of the general population conducted by Mannheim University[27]. In addition, several other studies have shown an interest in implementing interviewer questionnaires based on the conceptual framework in this paper.

Variation in the interviewer data is a prerequisite for explaining interviewer effects in survey data. The following section shows that there is considerable variation in key variables in the 2011 SHARE Germany interviewer survey. We focus on variables from the core of our conceptual framework, i.e. those related to item and unit nonresponse.

Variation in Attitudes Towards the Survey Process

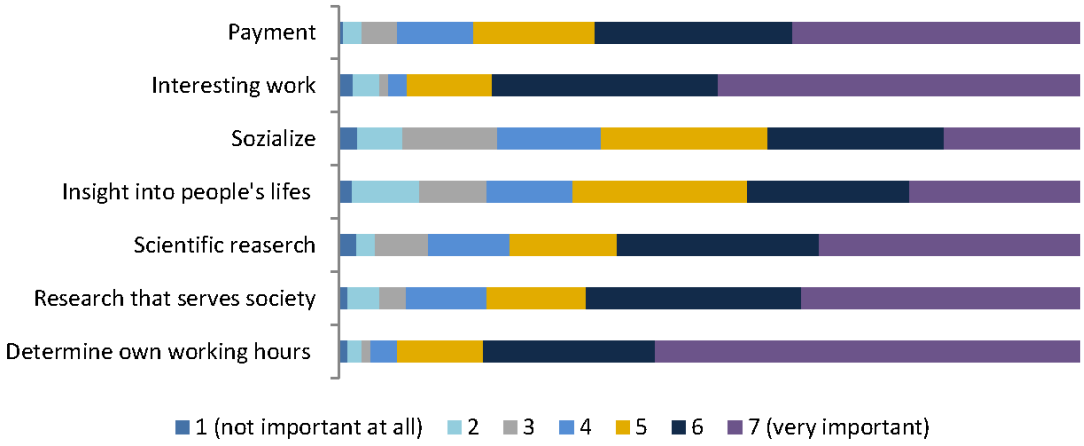
The first dimension of the conceptual framework is the attitudes that interviewers hold regarding survey interviews. In question 3 interviewers were asked for their reasons for working as an interviewer. Figure 4 shows that while many interviewers gave importance to most of the reasons presented, there was considerable variation. For example, while the opportunity of interacting with people (socialize) and gaining insight into other people's social circumstances were given importance scores of six and seven by about 45% of interviewers, about 80% of interviewers mentioned that the possibility to determine their own working hours and interesting work was this important to them.

The survey also contains an item battery inquiring interviewers' attitudes towards sticking to the prescribed interviewing protocols. Since interviewers are regularly trained and know what they are supposed to do, we were concerned that interviewers' attitudes towards the protocols would only reflect their training. Therefore, all items were phrased such that it would be legitimate for interviewers to admit that they deviate from the protocols. As Figure 5 portrays, there is large variation across items and interviewers. For example, interviewers widely differed in their answers to the statement "If the respondent doesn't understand a question, I explain what is actually meant by the question". Approximately 30% of interviewers answered that this statement does not at all apply to them, while almost 40% said that it perfectly applied to them. Similarly, there is great variation across interviewers as to whether they speak faster, if they notice that the respondent is in a hurry. Regarding other statements interviewers answered more homogenously. Almost all interviewers stated that they "always exactly stick to the

[26] http://www.gesis.org/en/piaac/piaac-home/

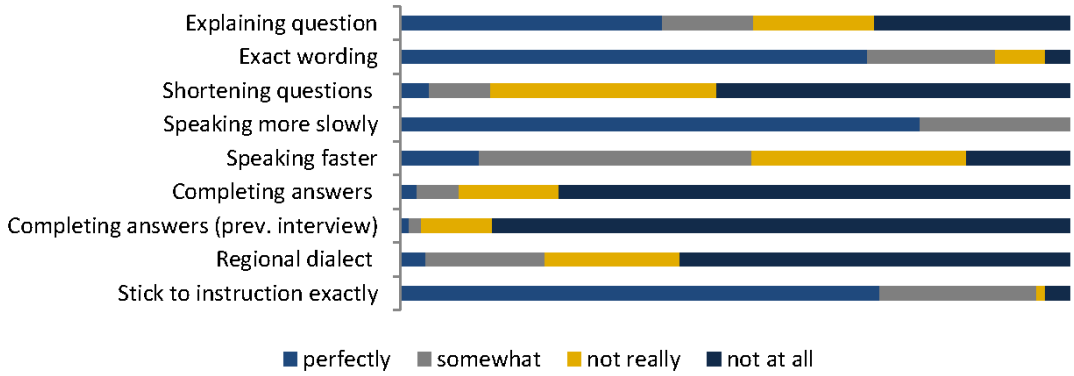
[27] http://reforms.uni-mannheim.de/english/internet_panel/home/index.html

Figure 4: Attitudes: Reasons for Working as an Interviewer



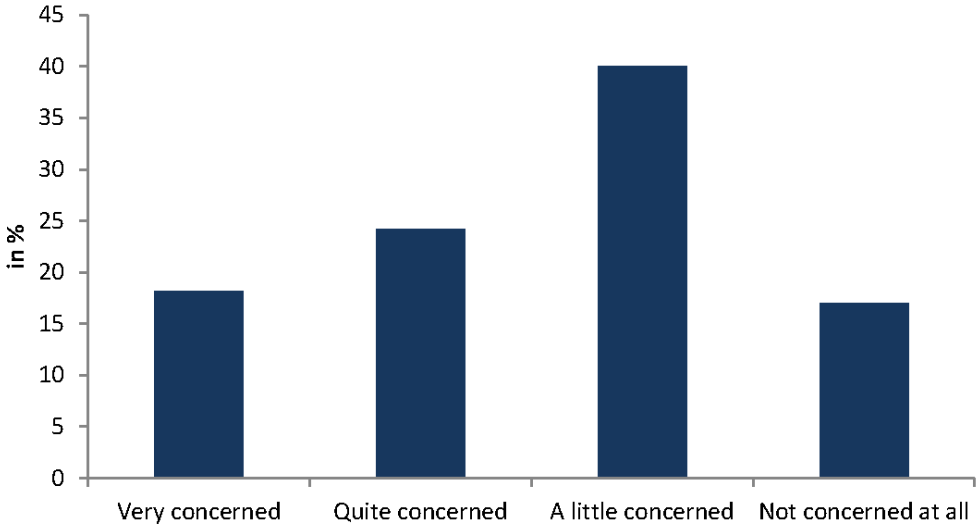
interviewer instructions, even if [they] don't consider them sensible" and all agreed that if they "notice that the respondent has difficulties understanding the question, [they] speak more slowly".

Figure 5: Attitudes: Following the Standardized Interview Protocols



We researched interviewers' attitudes towards data protection concerns and asked them how concerned they were about the safety of their personal data. As described above, we assume that this might be an indicator of how much trust in data protection they can instill in the respondent during the interview. Again, the results from the survey demonstrate variation in data protection concerns across interviewers (Figure 6) with between 17% and 40% of answers in each of the four categories.

Figure 6: Attitudes: Data Protection Concerns



"How concerned are you about the safety of your personal data?" (Q11)

Variation in Interviewer Behavior Regarding Data Collection Requests

The second dimension of the conceptual framework measures interviewers' own behavior in survey situations or similar contexts. The items displayed in Figure 7 indirectly look at whether interviewers are concerned about their private data, as we asked them if they used social networks and online banking. The figure illustrates that interviewers by no means are a homogenous group of people when it comes to their behavior on the Internet. While about 35% of interviewers use social networks, 63% have sufficient trust in the safety of the Internet to use it for online banking.

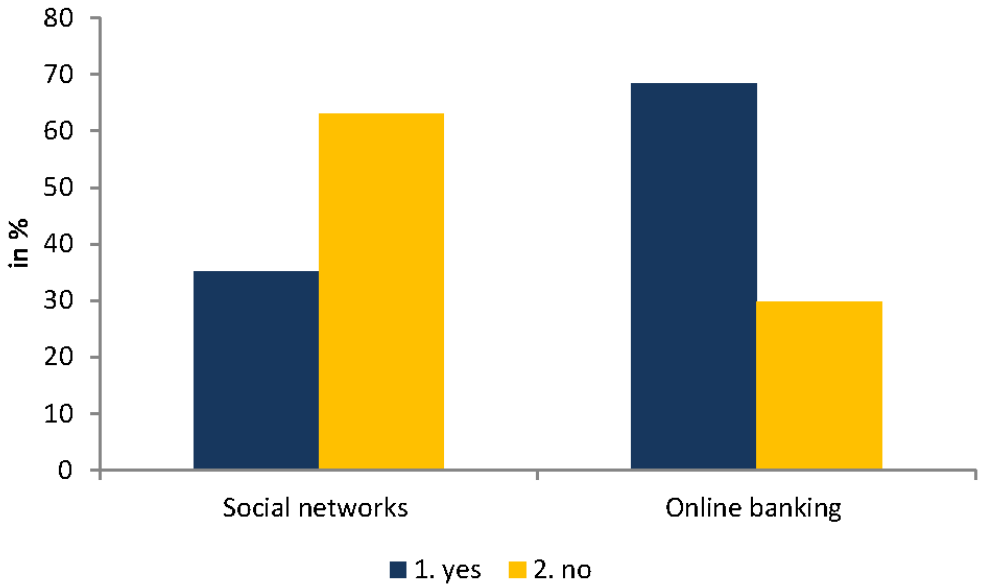
Variation in Experience

The interviewer survey contains several items measuring interviewers' experience with various measurements including their experience working as an interviewer, working for previous waves of SHARE, and collecting bloodspots. Figure 8 displays their general experience working as an interviewer. The results show that the interviewers working on the fourth SHARE wave varied in their experience: While 23% had less than one year of experience, 27% had been doing this work for more than 10 years.

Variation in Expectations

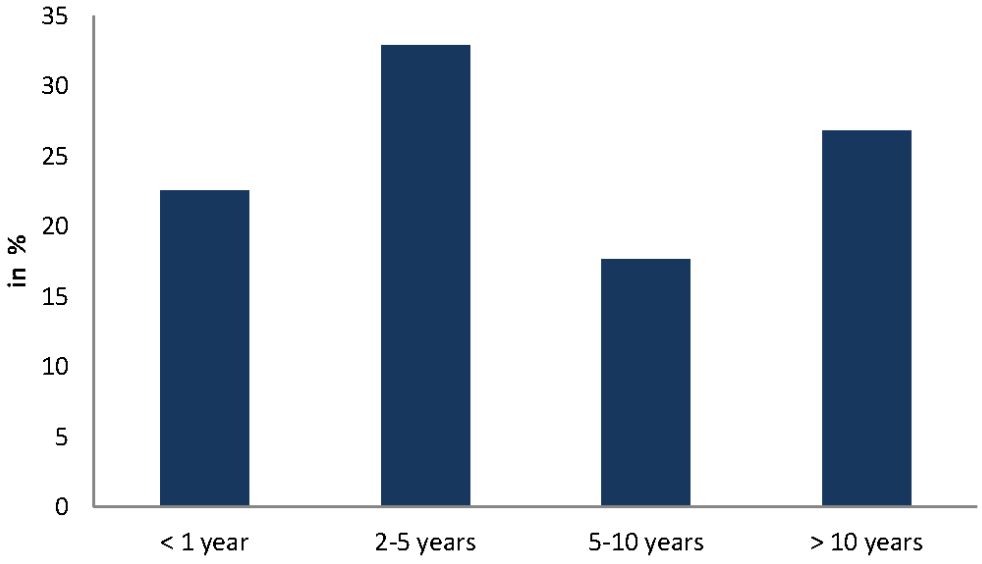
In 2011 the refresher sample of the SHARE survey in Germany was allocated to an incentives experiment with four treatment groups of unconditional incentives (€0, €10,

Figure 7: Own Behavior: Social Networks and Online Banking



"Do you use social networks in the internet like Facebook, Myspace or Twitter?" (Q27)
"Do you use the internet for online banking?" (Q28)

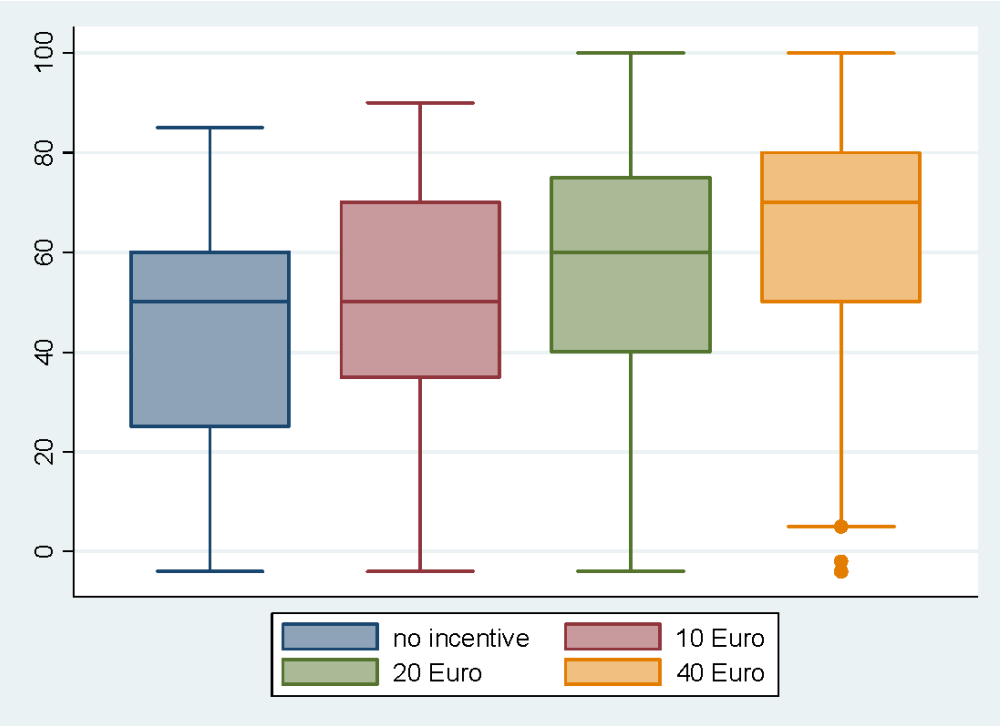
Figure 8: Experience with Measurements: Working as an Interviewer



"How long in total have you been working as an interviewer?" (Q1)

€20, €40). In addition, all respondents are always promised 10€ for the completion of the interview. In the interviewer survey we asked about interviewers' expectations regarding their unit response rate for each of experimental conditions. The results show that interviewers differed substantially in their confidence in achieving high response rates (Figure 7). When no unconditional incentive is sent with the advance letter, the SHARE interviewers on average expected unit response rates of 43%. However, as the boxplots in Figure 9 illustrate, the variation around the median is great. Furthermore, interviewers were confident that the higher the value of the incentive the more successful they would be in recruiting respondents. According to the interviewers' expectations the 40€ unconditional household incentive paired with a 10€ conditional individual incentive would on average yield a 23% increase in the unit response rate compared to a setting where no unconditional incentive is sent.

Figure 9: Expectations: Response Rates at Different Incentives Levels



"Studies vary as to whether they reward respondents for their survey participation and how much respondents receive. Please imagine that your respondents receive the following incentives. What do you expect, which percentage of your sample persons will agree to the interview, if..." (Q19)

4.5 Summary and Discussion

This works aimed to propose a conceptual framework of measuring interviewer characteristics for explaining interviewer effects on unit and item response, including consent to the collection of biomarkers, consent to record linkage, and item response on income measures. The conceptual framework encompasses four dimensions of interviewer characteristics:

- *Interviewer attitudes* towards the survey process that might shape the way interviewers approach sample units and ask their respondents for sensitive information, such as attitudes towards their job as interviewers, concerns about data protection and trust in other people.
- *Interviewers' own behavior* regarding data collection requests and hypothetical behavior when faced with survey requests or similar measurements.
- *Interviewers' experience with measurements*, for example, experience with conducting specific surveys or the collection of specific measurements like biomarkers or consent to record linkage.
- *Interviewers' expectations* about the unit and item response rates they will achieve on a given survey.

This conceptual framework formed the basis of an interviewer questionnaire implemented during the interviewer trainings in the fourth wave of SHARE Germany in early 2011. Exploratory analyses show that the survey well distinguishes between interviewers on the measures implemented along these four dimensions. This is a prerequisite for explaining interviewer effects. The theory, conceptual framework, and findings presented in this paper are merely a starting point for analyses of interviewer effects. The interviewer data can also be linked with paradata and survey data allowing a multitude of analyses into interviewer effects in SHARE Germany. Furthermore, parts of the interviewer questionnaire were also implemented in other surveys. Cross-survey analyses will allow investigating, whether findings are survey specific or hold generally across large-scale social surveys. This paper aims to contribute to the literature on interviewer effects by stimulating the development, collection, and analysis of new measures of interviewer characteristics to explain and ultimately adjust for interviewer effects in survey data. We make our conceptual framework and the interviewer questionnaire available to the public to encourage the continuous development of both and to conduct analyses of interviewer effects across surveys and countries. We hope to thereby foster research and insights in the area of interviewer effects in interviewer-mediated data collections.

5 Interviewer Effects on the Willingness to Provide Blood Samples in SHARE

The following chapter deals with the effect of the interviewers on respondents' willingness to consent to the collection of dried blood spots within the biomarker project. The availability of the data collected in the interviewer survey (Chapter 4) allows focussing on characteristics of the interviewers.

5.1 Introduction

The second example of innovations, discussed in Chapter 1, is a relatively new strand of research: the combination of medical studies and social surveys. Over the past few years, more and more studies have started the collection of biomeasures in social surveys as objective measurements of the respondent's health. A very promising new biomeasure is the collection of dried blood spots, as this new technology allows analyzing meaningful and objective blood parameters from only a few drops of blood. I discuss this project as an example of an extension of survey data with additional data sources as the blood has to be analyzed in a laboratory to then link the resulting data with the survey data.

As discussed in Chapter 1, this new module has a high influence on the job of the interviewers. Also here respondents' informed consent is usually necessary, so that the list of tasks interviewers have to preform has been extended to obtaining this consent and conducting the measurements. For all survey requests which require the respondent's consent, the consent process itself is a very important element. First, if not all respondents consent, this decreases the number of cases and therewith the statistical power. Second, systematic differences between the respondents who consent and those who do not can lead to bias (e.g. Korbmacher and Schröder, 2013; Sakshaug, 2013; Sakshaug et al., 2010).

Referring to the results of Chapter 3 we know that the interviewers are of special interest as they have an high influence on respondents' decision to consent or refuse. In addition, interviewers are under the researchers' control (see Groves and Couper, 1998), meaning that the characteristics of the interviewer can be influenced either by the selection process or by training. Therefore, understanding the mechanisms behind interviewer effects is essential for increasing the usefulness of the data and decreasing any potential consent bias. Different to Chapter 3, where only limited information about the interviewers had been available, the combination of the survey data with the data collected in the interviewer survey allows to focus on characteristics of the interviewers and their influence on the respondents' consent decision.

The goal of this chapter is to quantify interviewer effects on the respondent's decision to consent to the collection of dried blood spots in SHARE. In a second step, the data of the

interviewer survey will be used to explain the interviewer effects by the characteristics of the interviewers. Two characteristics of the interviewer are of special interest: their experience and their expectations in the consent rates they will reach.
The following chapter summarizes the idea of biomeasures in social surveys and the role of the interviewer (Section 5.2). Section 5.3 summarizes previous research. The research question and a description of the data are provided in Sections 5.4 and 5.5, respectively. Section 5.6 and 5.7 discuss the method and the results, and the chapter closes with a brief discussion of the results in Section 5.8.

5.2 The Collection of Biomeasures in Social Surveys

The integration of biology and the social sciences has become increasingly important in recent years, so that social surveys have started to collect biological characteristics and measurements (Sakshaug et al., 2014). The advantage of this development compared to classical medical surveys is that social surveys are typically based on probability samples with a large number of observations. Medical surveys are mostly based on small and non-random samples and miss collecting contextual information which is important for social scientists (Schnell, 2009). Collecting biological information in a social survey is a fruitful way to combine these two important disciplines.

The specific measures implemented in different surveys vary, as do the terms used for that class of measurements. Schnell (2009) differentiates between biometric attributes, biological attributes, and biomarkers, whereas Jaszczak et al. (2009) use the term *biomeasure* to summarize the "biological, anthropometric, functional, and sensory measurement" (Jaszczak et al., 2009, p.5) which can be collected in a survey. The following study adapts Jaszczak's terminology of *biomeasures* to summarize all physical measurements (such as grip strength and lung power test), measurements of the body (such as height, weight, or waist circumference), as well as the collection of bodily fluids. The advantage of these biomeasures compared to respondents' self-reports is first that they allow also for detecting undiagnosed diseases, for example diabetes, and second, that they can provide objective information about the health status without the measurement error due to respondents' misreporting (Jaszczak et al., 2009). Which biomeasures can be collected in a survey is not only a question for the usefulness for the research questions, but also of feasibility. One has to bear in mind that these measurements are typically collected by interviewers without medical training.

Two of the pioneer studies in implementing biomeasures within social surveys are the *Health and Retirement Study* (HRS) and the *National Social Life, Health, and Aging Project* (NSHAP).[28] HRS is a face-to-face panel survey in the U.S. conducted since 1992. It collects data from people aged 50 and older every two years (Sakshaug et al.,

[28]For a detailed overview of surveys collecting biomeasures, see Sakshaug (2013).

2010). It first piloted biomeasure collection in 2001, and has fully implemented these measurements since 2006 (e.g., physical measures, blood pressure, blood spots, and the collection of saliva) (Weir, 2008). The NSHAP started in 2005 and collects data from U.S. adults aged 57 to 85 face-to-face with the goal of studying "the links between health and sexuality in the lives of older American" (O'Muircheartaigh et al., 2009, p.12). The NSHAP collects a battery of twelve measures (including blood spots, saliva, vaginal swabs, and measures of weight, waist, and blood pressure) and, as in the HRS, all are conducted by trained interviewers without medical degrees (Jaszczak et al., 2009; Weir, 2008). In recent years, European social surveys have also implemented the collection of biomeasures in interviewer-mediated surveys. The British survey *Understanding Society* started collecting biomeasures such as height, weight, waist circumference, blood pressure, and blood in its second wave (2010), employing nurses (McFall et al., 2012). It switched to trained interviewers within a sub-sample in 2011 (McFall et al., 2014), demonstrating the pros and cons of using nurses vs. interviewers. In 2011, the *Survey of Health, Ageing and Retirement in Europe (SHARE)* tested new biomeasures in the German sub-sample (see Chaper 2.2). The following study will be based on the results of that study.

Biomeasures in the Survey of Health, Ageing and Retirement in Europe

The *Survey of Health, Ageing and Retirement in Europe* (SHARE) is a multidisciplinary and longitudinal survey collecting micro-data on the health, socio-economic status, and social and familial networks to learn more about the process of population aging (see Chapter 2 and Börsch-Supan et al., 2013). As health is a key aspect of aging, SHARE collects subjective and objective health measurements, such as self-perceived health status, grip strength, walking speed, and lung power, since the first wave of data collection (for an overview see Sakshaug et al., 2014; Hank et al., 2009). In 2011, SHARE implemented a pilot study of collecting new biomeasures in the fourth wave within the German sub-sample (Schaan, 2013). This pilot study tested the feasibility of collecting additional biomeasures with non-medical interviewers at the respondents' homes. This new module consists of four measurements: 1) height, 2) waist circumference, 3) blood pressure, and 4) the collection of blood spots. For these four measurements, the respondents' written consent is required, which is collected by the interviewer on a separate paper form. Respondents have to tick a box for each measurement they agree to, so that agreeing to one of these measurements is separate from agreeing to any of the other three measurements. This paper will only focus on the collection of dried blood spots, as this is a new and very sensitive procedure in social surveys. Neither German survey agencies nor German interviewers have had any experience applying this technique.

The Role of the Interviewer and Potential Interviewer Effects

The tasks for the interviewers when collecting dried blood spots in SHARE are manifold, with a high potential for interviewer effects. They not only have to ask for the respondent's consent, but also have to conduct the measurement and administer the process. All these tasks are assumed to influence the respondent's willingness to participate.

- **Asking for consent and answering all the respondents' questions**. Obtaining the respondent's informed consent is necessary when collecting dried blood spots (Sakshaug, 2013). This includes informing respondents about how the procedure works, the potential risks involved, which parameters will be analyzed from the blood, how consent can be withdrawn, and so on. In addition, respondents can restrict the parameters to be analyzed. As a result, the consent form is very long, consisting of four pages with very detailed information. Compared to other survey questions, this task requires much more skills on the part of the interviewer, as this request can hardly be completely scripted. This is of particular importance if respondents have questions or doubts, as the interviewers have to react spontaneously. One could assume that interviewers differ in their reactions, so that this request is quite prone to interviewer effects.

- **Conducting the measurements**. The collection of dried blood spots, that is, letting an interviewer prick a small needle into one's finger, probably requires that the respondent place more trust in the interviewer than does answering survey questions or participating in other physical measurements. The interviewer–respondent relationship during the interview thus far could play an important role in the decision of the respondent to participate.
  Interviewers might differ in how successful they are in building a trustful situation with the respondent, which could then affect the respondent's willigness to consent.

- **Administration of the process.** This task includes the handling of the materials in preparation of the blood collection. The blood is collected on special filter cards, which have to be prepared with a unique barcode sticker and sent to the laboratory in a special envelope. The number on the sticker has to be entered into the computer system so that the results can be linked to the correct respondent. This last step is also assumed to affect the respondent's decision to consent, as the way interviewers handle the materials prior to the consent question can affect the respondent's assessment of the interviewer's experience in that measurement.

This brief overview shows that the role of the interviewer in collecting dried blood spots in SHARE is much more prominent than for other questions or measurements in a survey. Interviewers who feel uncomfortable with the measurement and the whole procedure are assumed to be not as successful in 'selling' this request to respondents than are

interviewers who do not have any concerns or fears. These systematic differences would then result in interviewer effects in the consent question.

5.3 Previous Research

The number of studies analyzing consent requests in general is increasing but most studies focus on respondents' characteristics as determinants of consent. Only a small number of studies take the effect of the interviewer into account. Four recent studies on consent to record linkage show that interviewers are important as they have an influence on the respondent's decision to consent (Chapter 3: Korbmacher and Schröder, 2013; Sakshaug et al., 2013; Sala et al., 2012; Sakshaug et al., 2012). All studies analyzed the interviewers' demographics and only Korbmacher and Schröder (2013) found a significant effect of age. All four studies analyzed the effect of the experience of an interviewer, but the results are not clear. In addition, the studies differ in which aspect of experience they measure. Korbmacher and Schröder (2013) analyzed the experience within the current wave of data collection and found a statistically significant negative effect which is also found in Sakshaug et al. (2012) but not statistically significant. Sakshaug et al. (2013) included overall job experience as a dummy to compare interviewers working 37 months as an interviewer and those reporting a shorter period, and also found a statistically negative effect. Sala et al. (2012) included both job experience in years and the number of previous interviews, and found a positive effect of both but only the effect of the number of previous interviews is statistically significant. Sala et al. (2012) and Sakshaug et al. (2013) used additional data on the interviewers coming from an interviewer survey to analyze interviewer effects on consent to record linkage. Beside experience, they controlled for additional characteristics, such as attitudes and personality traits, interviewers' income, hypothetical own-consent to a different consent request, membership in social networks, and the expected consent rate. Only the interviewer's own-willingness to consent to a series of consent requests showed a significant (and positive) effect (Sakshaug et al., 2013).

Even less is known about interviewers' influence on consent to the collection of dried blood spots. Previous studies show that interviewers vary a lot in the consent rates they obtain and that these consent rates also vary between different biomeasures (Sakshaug, 2013; McFall et al., 2014; Jaszczak et al., 2009). To my knowledge there is only one study which systematically analyzes the consent to the collection of dried blood spots taking the interviewer into account.[29] Sakshaug and colleagues (Sakshaug et al., 2010) used the 2006 wave of the HRS and analyzed the differences between consenters and non-consenters to the collection of dried blood spots. In addition to the respondents'

[29]The authors analyzed consent to a set of three biomeasures, including dried blood spots, together as the dependent variable. Separating the regressions by biomeasure did not change the results, so I will use the phrase "consent to dried blood spots" in the remainder of the present chapter.

demographics and widespread measures of the respondents' health status, they included a set of variables measuring general survey resistance indicators coming from paradata as well as information on the interviewer. They controlled for the interviewer's age, gender, race, educational level, Hispanicity, and experience being an HRS interviewer. At the interviewer level, only the interviewer's race shows a significant effect on the respondents' consent. A significant interviewer variance term suggests that interviewer characteristics (other than demographics) have an influence on consent. This variance term shows that more information about the interviewers is needed to get the full picture. The implementation of an interviewer survey allows filling that gap by collecting that specific information on the interviewer which is assumed to have an influence.

5.4 Research Question

As interviewer effects on different consent questions have been proven to exist in several surveys (Korbmacher and Schröder, 2013; Sakshaug et al., 2013; Sala et al., 2012; Sakshaug et al., 2012, 2010), I hypothesize that interviewer effects also occur in SHARE when asking for consent to the collection of dried blood spots. Therefore, the first step of this analysis is to test that assumption and quantify the effect of the interviewer.
If this assumption can be confirmed, the next step will be to analyze the effect of interviewer characteristics, focusing on experience and expectations, as these are two characteristics of the interviewer which can be manipulated via selection and/or training of the interviewer.

Interviewers' Experience

The interviewers' experience seems to matter in their success at getting respondents' consent to record linkage (Korbmacher and Schröder, 2013; Sakshaug et al., 2013; Sala et al., 2012; Sakshaug et al., 2012). I hypothesize that interviewers' experience also influences the consent to the collection of dried blood spots. Three different aspects of an interviewer's experience will be distinguished: job experience, experience in collecting dried blood spots, and experience in measuring blood sugar.

- *Job experience*: This is measured as the number of years working as an interviewer. In contrast to Sala et al. (2012), I do not expect the effect of job experience to be linear. I hypothesize that an increase in experience is mainly effective at the very beginning of the career. In addition, being on the job for a very long time also implies that the interviewer's job and therewith the required tasks has changed substantially. I suspect that interviewers who started working as an interviewer a long time ago are less successful than interviewers who started more recently.
- *Experience in collecting dried blood spots in the actual wave of data collection* is not related to the first experience measurement. As SHARE is the first survey

in Germany to collect dried blood spots by interviewers, they all start without any experience in asking respondents for consent to that measurement. But one could assume that interviewers learn how to persuade respondents from interview to interview. I hypothesize that interviewers are less successful in getting consent at their very first interviews.

- *Prior experience in the technique of collecting blood spots*: The procedure is almost identical to measuring blood sugar levels for people who have diabetes. I hypothesize that interviewers who are experienced in that measurement (independently from their job as interviewers) are better at getting the respondent's consent, as they are less fearful about the procedure.

Interviewers' Expectations of the Consent Rate They Will Reach

Even if the effect of an interviewer's expectations when asking for consent to record linkage was not significant in the work of Sakshaug (2013), I hypothesize that expectations are important in this specific consent request. It will be tested whether interviewers who expect to achieve a higher consent rate also reach higher consent rates. The theoretical assumption behind this is the theory of self-fulfilling prophecies, which should affect all interviewers in the same way: expectations influence the behavior of the interviewer and therewith the respondents' reactions to the request.

5.5 Data

SHARE: Survey Data

This paper is based on release 1.0.0 of the German Wave 4 panel sample, in which the collection of dried blood spots was implemented for the first time.[30] The target population of SHARE consists of persons aged 50 or older at the time the sample was drawn, including partners living in the same household regardless of their age (Börsch-Supan et al., 2013). A total of 1,570 respondents were asked for consent to the collection of dried blood spots during their personal interview.

Interviewer Survey

Since the information about the interviewers delivered by the survey agency is limited to a few demographical characteristics, a separate interviewer survey was conducted with the interviewers working for the fourth wave of SHARE in Germany. The questionnaire asked for information about the interviewers' experiences and expectations related to different features of the fourth wave of SHARE-Germany, including the collection of dried

[30] For a detailed overview of SHARE's cooperation rates, see Kneip (2013).

blood spots (see Chapter 4). The interviewers were asked to complete the survey at the end of the training session, ensuring that their attitudes were measured independently from their first experiences in the field. Out of 197 interviewers attending the training, 165 completed that survey (a response rate of 83.8%).

Combining Both: Linking the SHARE Survey with the Interviewer Survey Data

The data of the two surveys could be linked via the interviewer ID which was requested in both surveys: in the CAPI instrument at the end of each completed interview and at the beginning of the interviewer survey. Despite the high response rate of the interviewer survey, several causes limited the number of cases for which respondent survey data could be linked to the survey data of the interviewer who conducted the interview.

1. Not all interviewers who had been trained for SHARE decided to complete their job as a SHARE interviewer: 40 of them quit before the fieldwork started.

2. Additional interviewers were hired during the fieldwork and were trained at separate training sessions. As the interviewer survey was implemented at the regular training session, these new interviewers were not asked to participate in the survey. These interviewers were mainly deployed for the refreshment sample, which is not included in the following analysis. Only three new interviewers worked for the panel sample and conducted 27 interviews.[31]

3. Unit- and item-nonresponse in the interviewer survey are responsible for an additional reduction of the sample size. The question regarding the interviewer ID in the interviewer survey suffers from item-nonresponse, so that these data could not be linked to the survey data. The survey data of 26 interviewers who conducted 555 (36%) interviews suffer from unit- or item-nonresponse.

The survey data of 988 respondents could be linked successfully with the data of the interviewer survey. This corresponds to 63% of the completed panel sample ($N = 1,570$). As the selection into the final sample is not random but depends on the interviewer, one cannot rule out that the sample is selective. However, a t-test of the interviewer sociodemographics of those interviewers who are included and those who are excluded from the final sample shows no significant differences in the characteristics which are available for all interviewers as they are provided by the agency (see Table 5). In addition, differences in the interviewer specific consent rate and the total number of interviews were tested and also show no significant differences. With respect to the seven variables which are available for all interviewers, the sample is not selective.

[31] This corresponds to 1.7% of the sample.

Table 5: Comparison of Interviewers who are Excluded with the Final Sample

	Excluded		Final sample	
Age	57.2	(1.96)	58.5	(1.24)
Men	48.3%	(0.94)	53.4%	(0.07)
Experience	5.4	(0.90)	5.2	(0.50)
Years of education	11.3	(0.39)	11.9	(0.21)
Having SHARE experience	41.4%	(0.09)	37.9%	(0.06)
Consent rate	50.5%	(4.95)	54.2%	(3.9)
Total no. of interviews	20.1	(2.96)	17.1	(1.9)
Number of interviewers	29		58	

Notes: *, **, *** mark significance on the 10, 5, 1 percent level, respectively

Standard errors in parentheses

5.6 Methods and Models

To analyze the effect of the interviewer on the respondent's consent requires a multilevel model to take the hierarchical data structure into account, as the respondents (first level) are nested within the interviewers (second level). The dependent variable in this model is the consent to the collection of dried blood spots. The consent form was handed to the respondent after the interviewer explained the procedure. This form collects consent for all four biomeasures separately. Only if the respondent signs this form is the interviewer allowed to conduct the measurements. At the end of the biomeasure module, the interviewer answered a question in the CAPI instrument indicating which measurements were completed. That final result is the dependent variable which is coded as a dummy, being 1 if the interviewer states that he or she conducted the measurement and 0 if not. The model used here is identical to the model of Chapter 3 but nevertheless the description of the model will be repeated here.

Intercept-Only Model

As a first step, an intercept-only model is calculated which does not include any explanatory variables. The outcome Y_{ij}^* for respondent i interviewed by interviewer j is explained as the regression intercept α, the residual at the interviewer level u_j, and the respondent level residual ϵ_{ij} (see Hox, 2010).

$$Y_{ij}^* = \alpha + u_j + \epsilon_{ij} \tag{6}$$

$$Y = \mathbb{1}(Y_{ij}^* \geq 0) \tag{7}$$

As the dependent variable is a dummy variable which can either be 1 or 0, Y equals 1 if the latent variable Y_{ij}^* is greater than or equal to zero. This intercept-only model provides an estimate of the intraclass correlation (ICC) ρ, which is the proportion of variance at the highest level compared to the overall variance. The ICC is calculated as the variance of the residuals at the interviewer level $\sigma_{u_j}^2$ divided by the total variance $(\sigma_{u_j}^2 + \sigma_{\epsilon_{ij}}^2)$. Given that the respondent level variance $\sigma_{\epsilon_{ij}}^2$ is not distributed normally but logistically, this term is fixed at $\frac{\pi^2}{3}$.

$$\rho = \frac{\sigma_{u_j}^2}{\sigma_{u_j}^2 + \frac{\pi^2}{3}} \tag{8}$$

Full Model

The interviewer–respondent assignment in SHARE Germany is not random (no interpenetrated sample) but by region, which implies that all respondents interviewed by the same interviewer also live in the same region. If the respondents in one region differ systematically in some characteristics that also influence their consent, this would result in a high ICC. In such a case, the interpretation of the ICC as interviewer effects would be misleading, as these are in fact area effects. To take into account such potential area effects, the respondents' basic demographics and some health related parameters which showed significant influences on consent in other studies are controlled for. In the next step, the characteristics of the interviewer as well as of the respondent will be included in the model, where the X_{pij} are the p explanatory variables at the respondent level and the Z_{qj} are the q explanatory variables at the interviewer level. The slopes of the X_{pij} are assumed not to vary at the interviewer level (fixed slope model) in the final model. Part of the interviewer level variance of Model (6) is assumed to be explained by Z_{qj}, with u_{1j} being the remaining interviewer level residual. Simultaneously, part of the respondent level variance of Model (6) is assumed to be explained by X_{pij} with ϵ_{1ij} being the remaining respondent level residual.

$$Y_{ij}^* = \alpha + \gamma_p X_{pij} + \gamma_q Z_{qj} + u_{1j} + \epsilon_{1ij} \tag{9}$$

In both models (1) and (4) the term u_j is included, which is the residual at the interviewer level (random intercept). In multilevel models these residuals are assumed to be normally distributed (Hox, 2010; Snijders and Bosker, 2012; Rabe-Hesketh and Skrondal, 2008). For further discussion see Appendix D.1.

Explanatory Variables

Interviewers' Experience and Expectations

Three different measures of the interviewers' experience are included: The experience of working as an interviewer was measured in years[32] and is included as a continuous variable. In addition, the quadratic term of years of experience is included to test the assumption of an inversely u-shaped effect.

Experience in collecting dried blood spots in the actual wave of data collection is included to take learning effects into account. The dummy variable is 1 if the actual interview is within the first five interviews of that interviewer. As this variable is not a fixed interviewer characteristic which is stable over all respondents interviewed by the same interviewer, this variable should (from a multilevel point of view) be categorized as a respondent level characteristic. As I am interested in the learning effect of the interviewer, I will interpret this variable at the interviewer level. In addition, we asked interviewers about their experience in that measurement[33] and include that measure of experience as a dummy variable (1= familiarity, 0= otherwise).

Interviewers expectations with regard to the collection of biomeasures are asked in the interviewer questionnaire separately for each of the four new biomeasures.[34] These expectations are included as a continuous variable. The theoretical assumption behind the expected effect of the interviewer's expectations is the theory of self-fulfilling prophecies, which should affect all interviewers in the same way: the expectations influence the behavior and thereby the respondent's reaction to the request. Another explanation of a potential correlation between an interviewer's expectations and the respondents' willingness to consent could be based on the interviewer's experience. Experienced interviewers could be assumed to be more realistic in the assessment of their own abilities, meaning that they are more realistic in their expectations. If this mechanism is the one driving the effect, interviewers' expectations should be more important for experienced interviewers. To test whether the effect of interviewers' expectations differs by the interviewer's experience, an interaction of the experience and the expectations is included in the model[35] to differentiate between the two potential mechanisms.

[32] Q1: How long in total have you been working as an interviewer?

[33] Q23: Do you personally have experience with measuring blood sugar levels, either because you or someone you know has diabetes?

[34] Q21: What percentage of your respondents do you think will consent to [...] the collection of small blood spots?

[35] Both variables are centered around their mean.

Control Variables at the Interviewer Level

In addition to these explanatory variables, some control variables at the interviewer level are included. Interviewers *age* in years (included as a continuous variable), *gender* (1=male, 0=female), and *education* (as three dummies for low, medium, and high educational levels) are controlled for. We asked interviewers whether, if they themselves were SHARE respondents, would have consented to the collection of dried blood spots. This variable is included as a dummy. The variable *social networks* is a dummy variable taking the value of 1 if the interviewer is active in online social networks like Facebook, Myspace, or Twitter, and 0 otherwise. It is used as an indirect indicator of how open-minded an interviewer is about new technologies and the disclosure of personal information.

There are different reasons why people decide to work as interviewers. The measurement of their motivation is more complex, as the question includes a battery of different reasons for working as an interviewer, which should be rated in their importance to the interviewer separately.[36] Two reasons are included in the model: first, to "be involved in research that serves society," expecting these people to be intrinsically motivated, and second, "to have the opportunity to interact with other people," assuming that this reason does not reflect an intrinsic motivation. The two variables included in the model do not reflect the value of the rating (1–7) but they are coded as a dummy which is 1 if the corresponding aspect was rated more highly than the other aspects. Table 6 summarizes the distribution of these interviewer characteristics.

Control Variables at the Respondent Level

To control for systematic differences between respondents living in the same region, the respondents' demographics and characteristics which showed an effect in a preliminary unpublished analysis (see Weiss, 2013) of differences between SHARE respondents who consent and those who refuse are included in the model. As at the interviewer level, the respondents' gender, age, and education are included. Respondents who grew up in the former German Democratic Republic seem to be more willing to consent to record linkage (Korbmacher and Schröder, 2013; Lamla and Coppola, 2013) and to the collection of dried blood spots in SHARE (Weiss, 2013) than those who had not.[37] As respondents' health status could be influenced by their residential area and also affect their willingness to consent to the collection of dried blood spots (Weiss, 2013; Sakshaug et al., 2010), the three health measurements Weiss (2013) used in her study are included. These are whether respondents had been diagnosed with high blood cholesterol or diabetes (both

[36] There are different reasons for working as an interviewer. How important are the following aspects to you? 1=not at all, 7=very important.

[37] This effect could not be found when taking the interviewer into account.

Table 6: Sample Statistics

Interviewer characteristics	Mean/%		Min	Max
Years of job experience	7.7	(9.1)	0	41.5
Experience in measuring blood sugar	36.4			
Expected consent rate	59.6	(18.5)	4	90
Hypothetical own consent	78.2	-	-	-
Age	57.2	(9.4)	36	76
Male	54.6	-	-	-
Low educational level	7.2	-	-	-
Medium educational level	56.4	-	-	-
High educational level	36.4	-	-	-
Motivation "socialize"	27.3	-	-	-
Motivation "research"	47.3	-	-	-
Member of social networks	36.4	-	-	-
Number of interviewers	55			

Notes: Standard deviation of means in parentheses

have a significant positive effect) and the number of difficulties with everyday activities due to health problems (which influences consent negatively).[38] We know from previous research on consent to record linkage that the willingness to provide income information is a strong predictor of the probability of consenting to record linkage (Korbmacher and Schröder, 2013; Lamla and Coppola, 2013; Sala et al., 2012; Beste, 2011), and consenting to the collection of dried blood spots (Weiss, 2013). Therefore a dummy variable is included which is 1 if income was not reported and 0 otherwise. The variable "urban" is a characteristic of the area the respondent lives in and is 1 if the interviewer coded the area as a big city, suburbs or outskirts of a big city or a large town, and 0 otherwise.

5.7 Results

To answer the first research question, whether interviewer effects occur, we turn to an explorative approach. Figure 10 displays the interviewer specific consent rate, where each circle represents one interviewer and the size of the circle corresponds to the number of interviews that interviewer conducted.[39] This pattern shows that there is a large variation between interviewers in how successful they are in getting respondents' consent

[38]Included are activities such as dressing, preparing a meal, eating, getting in or out of bed, and so on.

[39]Included are only those interviewers who are in the final sample. The graph which refers to all interviewers looks very similar.

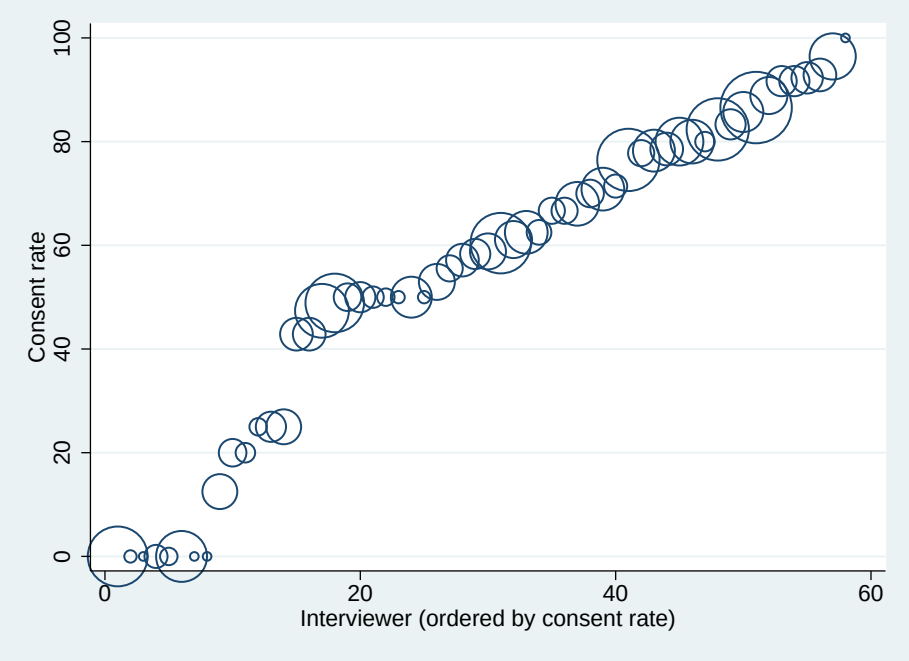


Figure 10: Interviewer Specific Consent Rate

to the collection of dried blood spots, and is a first hint of the existence of interviewer effects. The intercept-only model is used to confirm the interpretation of Figure 10: there are sizeable interviewer effects on respondents' decisions to consent to the collection of dried blood spots. The ICC of the null model is 0.36 and is statistically significant, meaning that 36% of the total variance is at the interviewer level (see Table 7, Model 1).

The next step is to analyze the effect of the interviewer's experience and expectations. Model 2 of Table 7 shows the results of the interviewer characteristics from the full model. As the respondents' characteristics are only control variables, the results are not discussed here but displayed in Appendix C, Tabel 25. Two of the three experience measures show significant effects on the respondents' willingness to consent: the experience within the actual wave of SHARE and the years of job experience. There is a positive learning effect within the field period, as the significantly negative effect of being within the first five interviews shows. Respondents who are not one of the first being interviewed by that interviewer are more likely to consent to the collection of dried blood spots. The coefficient of the job experience (years working as an interviewer) as well as the coefficient of the quadratic term are both negative and statistically significant, indicating that the effect of job experience is not linear. As the interpretation of the effect gets rather complex when transformations of a variable and interactions with

other variables are included, a graphical display of the relationship helps to understand the effect of experience on the dependent variable.

Table 7: Multilevel Estimation: Consent to the Collection of Dried Blood Spots

	Model 1	Model 2	
Age		1.03**	(0.02)
Male		0.62	(0.19)
Low educational level		0.16***	(0.10)
Medium educational level		1.30	(0.42)
Member of social networks		1.48	(0.50)
Hypothetical own consent to dried bs		1.16	(0.45)
Motivation: "socialize"		0.51*	(0.19)
Motivation: "research"		0.96	(0.34)
Experience in measuring blood sugar		0.83	(0.27)
1–5. interview		0.56***	(0.11)
Years of experience		0.96*	(0.02)
Years of experience^2		0.99***	(0.00)
Expected consent rate		1.04***	(0.01)
Years*Expectations		1.01***	(0.00)
ICC	0.36	0.09	
Number of interviewers	55	55	
Number of cases	843	843	
χ^2 against logistic regression	174.79***	5.93 ***	
χ^2 of LR test against previous model		99.08***	
(degrees of freedom; p-value of LR test)		(27; 0.000)	

Notes: *,**,*** mark significance on the 10, 5, 1 percent level, respectively

Dependent variable in all models is the dichotomous variable "consent to dbs collection"

All models are estimated in a multilevel logistic regression with Stata's xtlogit command with a random intercept on the interviewer level. Coefficients are odds ratios.

χ^2 are the respective test statistics; Model also controls for respondent characteristics.

Standard errors in parentheses.

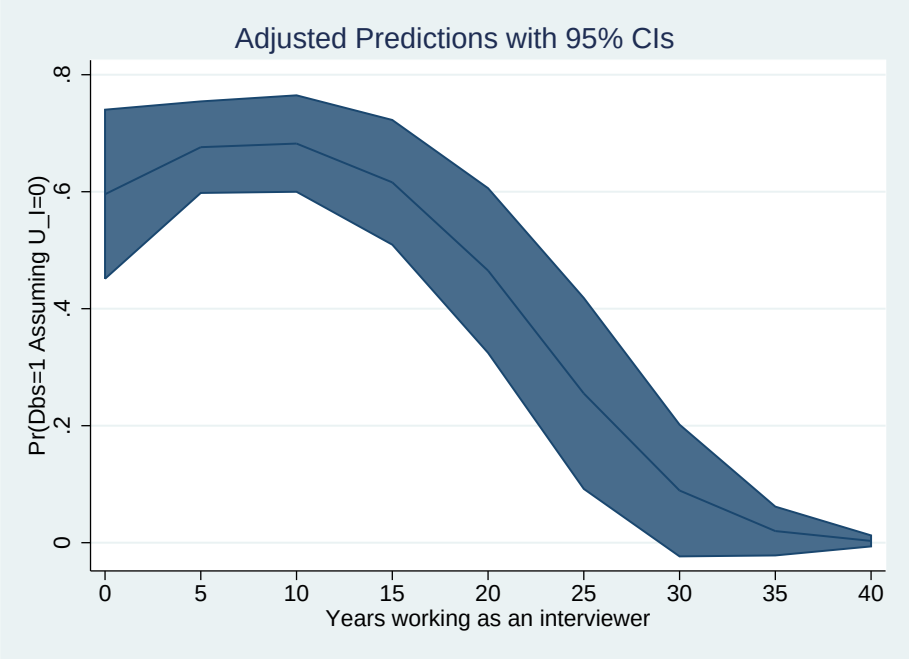


Figure 11: Predicted Probabilities of Consenting by Experience

Figure 11 shows the predicted probabilities for different levels of experience. All other variables are fixed at their mean and the random intercept is fixed at zero, meaning that the predicted probability refers to the average interviewer. The curve of the predicted probabilities shows that experience has a positive effect at the beginning of a career, reaching its peak at 7 years. After 7 years, the effect of experience is negative, meaning that the predicted probability of consenting decreases with each additional year of being an interviewer. But as the confidence interval shows, the increase in the first years is not statistically significant. The third experience measure, whether interviewers have experience in measuring blood sugar levels, shows a negative effect on consent, but this effect is statistically not significant.

Interviewers' expectations regarding the consent rate that they will reach show a positive effect on respondents' consent, as does the interaction term of expectations with experience. Similar to the results of experience, the effect of an interviewer's expectations will be discussed with the help of the predicted probabilities. As the interaction term is significant, the predicted probabilities will be presented for three different levels of experience: the lowest 10%, the average level of experience, and the highest 10%. Figure 12 shows that interpreting the positive coefficient of expectations independently from experience would be misleading. The curves of the predicted probabilities for expectations differ substantially depending on the interviewer's experience. In contrast

to inexperienced interviewers, who show a negative but statistically non-significant effect of expectations on the predicted probability of consenting, the effect is positive for interviewers with about 10 years of experience. Very experienced interviewers show a different pattern, as the increase in the predicted probabilities starts later with a larger slope. The share of unexplained variance at the interviewer level after controlling for interviewer characteristics decreased to 9%: a reduction of 27 percentage points.[40]

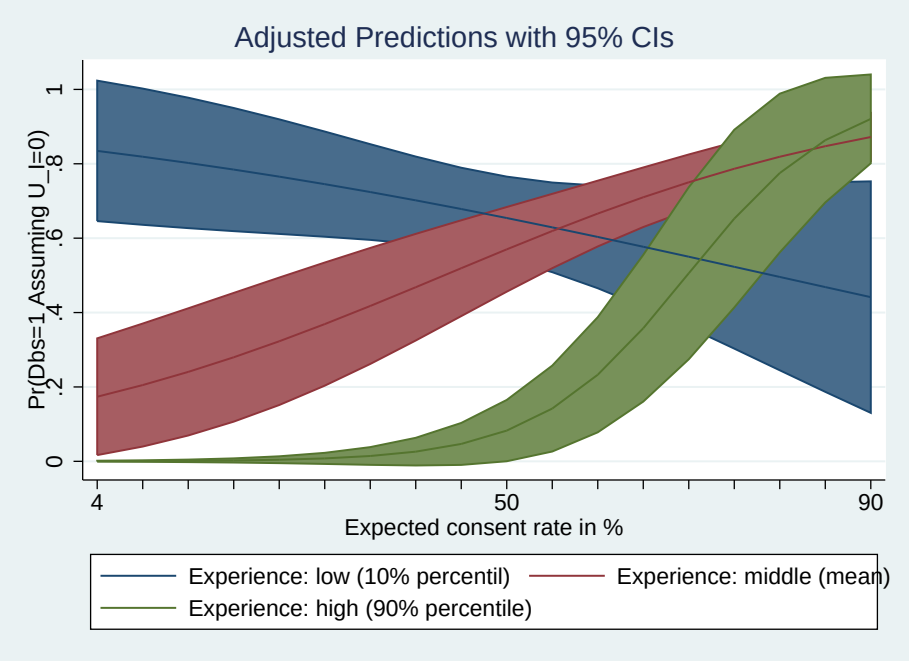


Figure 12: Predicted Probabilities of Consenting: Expectations by Experience

Regarding interviewers' demographics, both education and age have significant effects on respondents' consent. Less educated interviewers are significantly less successful than more educated interviewers in obtaining respondents' consent. Age is also positively correlated with consent. Of the two reasons for working as an interviewer, only the motivation "socialize" shows a significant effect: interviewers who report that the opportunity to interact with other people is most important are less successful in obtaining consent than are interviewers who rated other motivations higher.

A replication of the analysis after excluding two conspicuous interviewers who both

[40] As the scale of the outcome variable changes when variables are added to the model in logistic regressions, the comparison of regression coefficients and variance components is difficult (Blom et al., 2011). I rescaled the variance components of the full model to the metric of the empty model as described in Hox (2010) to calculate the ICC, but this does not change the results. For details, see Appendix D.2.

interviewed a high number of respondents who all refused to consent, shows that the effect of interviewers' experience is not stable, which could violate the external validity of the results (see Appendix C.2). It is questionable whether these interviewers should be included in the sample or not. Following a suggestion in Matschinger et al. (2005), I decided to not exclude them for two different reasons: First, it is hard to decide which interviewers should be excluded as there is no criterion available (Matschinger et al., 2005). So the decision is in some way arbitrary. And second, by removing interviewers based on certain characteristics, the assumption of a random set of interviewers (which should be representative of the whole sample) is systematically violated.

5.8 Summary and Discussion

The goal of this paper was to examine whether interviewers have an effect on respondents' decision to consent to the collection of dried blood spots and whether the interviewer effects can be explained by characteristics of the interviewer. There are three main findings: First, interviewers have a large effect on respondents' consent decision, as the ICC of 0.36 in the empty model shows. Second, the information collected in the interviewer survey is very useful in explaining the interviewer effects. The interviewer variance in the full model was decreased substantially. Third, interviewer's experience and expectations regarding their success prior to the first interview shows significant effects, as does their interaction term.

This study shows that the effect of experience is not linear, but curvilinear, being positive at the beginning and negative after seven years of experience. Comparing the effect of interviewers' experience over different studies reveals ambiguous results, making it hard to come to a final conclusion about how experience affects survey outcomes. One possible explanation of this contradictory and unclear result is that different aspects overlap when controlling for the number of years a person has worked as an interviewer. Durrant et al. (2010) demonstrate that the positive effect of experience on cooperation changed after controlling for interviewers' pay grade, which reflects skill level but also increases over time. In addition, the authors suggest that self-selection processes can complicate the interpretation of experience effects. The authors assume that interviewers stay longer in a job if they are successful, whereas interviewers with a lower level of performance quit earlier. From my point of view, it is debatable whether these two mechanisms can be generalized over surveys and countries, as the organization of survey agencies and the arrangement of how interviewers are paid varies a lot between countries. As interviewer experience is an important aspect in the process of interviewer recruitment prior to a survey, learning more about this effect is very important. This is an issue where survey agencies and survey researches should better cooperate.

Another aspect discussed in Chapter 5.6 is related to the changing demands of the interviewer's job over the last decades. This affects particularly those interviewers who have been on this job for a very long time. Due to the implementation of computer-assisted personal interviews, the abilities needed to conduct an interview have changed substantially. So the question is whether all interviewers are able to acquire the new skills at the same speed as the demands change. Further research is needed to disentangle these different aspects and learn more about the effect of experience.

The effect of interviewers' expectations regarding their own success at getting consent seems to be positive at first glance. But the results change when taking interviewers' experience into account. For interviewers with an average level of experience and for those with much experience, a positive trend is observable. The relationship of experience and predicted success is the reverse for inexperienced interviewers. As discussed in Chapter 5.6, the significant effect of the interaction term could be a hint that self-fulfilling prophecies are not the underlying mechanism that explains the correlation of experience and outcome. It rather seems to be the case that experienced interviewers are better at assessing their own abilities, and therefore they perform better at forecasting their consent rates. Two limitations of this study make it difficult to differentiate between "self-fulfilling prophecies" and the ability to forecast consent rates. First, due to the problems noted in Chapter 5.5, the number of cases which are included in the calculation of the predicted probabilities is very low, not allowing more differentiated analyses. In addition, expectations are only measured at one point in time (after the interviewer training) but could be assumed to change during fieldwork as interviewers obtain their initial experiences. Therefore, the effect of expectations prior to the first interview could affect the first interviews differently. The low number of cases does not allow limiting the analysis to the first interviews. Further research is needed to analyze the effect of expectations. The collection of dried blood spots will be repeated in the sixth wave of SHARE with a new refreshment sample, which will quadruple the sample size and thereby also increase the number of interviewers.

6 Recall Error in the Year of Retirement

The following chapter is an example of using the linked dataset SHARE-RV, which allows the validation of survey answers. Information about the year of retirement is included in both datasets and will be compared to learn more about recall error in survey data.

6.1 Introduction

As discussed in Chapter 1, measurement error is one of the errors within the total survey error paradigm which influence data quality. Measurement error is defined as "a departure from the true value of the measurement as applied to a sample unit and the value provided" (Groves et al., 2009, page 52). This definition covers various sources for deviations between the true value and the one which is measured. The books *Measurement Errors in Surveys* (Biemer et al., 2004) and *Survey Error and Survey Costs* (Groves, 1989) give an overview about, and are structured along the different sources of error, which could be the interviewers, the respondents, the questionnaire, or the mode of data collection (Groves, 1989; Biemer et al., 2004). The following chapter focusses on the respondent as the source of error. In this context, one often meets the term 'response error' as a subtype of measurement error. 'Response bias' might result if the measurement error is systematic, meaning that there is a consistent direction of the error (Groves et al., 2009).

The term response error often provokes the (negative) association of 'lying respondents' who are aware of the true answer but not willing to provide it in the interview; an explanation which is often used in the context of personally sensitive questions. This strand of the literature deals with measurement error as a result of social desirable answering behavior (for example Esser, 1991; Stocké, 2004; Stocké and Hunkler, 2007). It is well documented that the error can go in two directions: overreporting as well as underreporting, depending on whether the survey question is about socially desirable or undesirable behavior and attitudes (Bound et al., 2001).

Another strand of the literature treats the cognitive processes which occur when respondents are interviewed (Bound et al., 2001). Tourangeau et al. (2000) propose a 'Model of the Response Process' which is based on four main components of the response process, which are: comprehension of the question, retrieval of the information, judgement of the information, and the final response with the information. Unlike the first example of social desirable answer behaviour, measurement error is not discussed as a conscious decision of not reporting the truth, but as a result of errors in one or more steps of the cognitive process. In the following, the term 'recall error' is used when referring to an error which is based on cognitive processes to delimit this source of error from measurement error in general.

One challenge when analyzing measurement error is the question of how to assess it. With one single measurement one can detect implausible values but this does not allow assessing the error, as no information as to the true value is available. Therefore, at least two measures of the same construct are needed. These could be multiple indicators of the variable or validation data (Bound et al., 2001). Dex (1995) uses the terms 'reliability' and 'validity' of the data, to distinguish between these two constructs: the first refers to differences between repeated measures of the same construct under equal conditions, and the second to differences from almost error-free external records.
Administrative data which could be linked on the micro-level to the respondent's answers are often discussed as a promising source of validation data (Bound et al., 2001; Calderwood and Lessof, 2009; Couper, 2013; Korbmacher and Schröder, 2013). In doing so, one should not ignore the fact that other factors as measurement error can lead to differences between the value reported by the respondent and the one included in the administrative data (Bound et al., 2001). Whether a comparison of the survey and administrative data is a valid way to assess the measurement error depends on both, the survey question and the administrative data being used as validation data.

As more and more surveys have started to link survey data and administrative data (see Chapter 3), an increasing number of validation studies are based on the possibility to validate survey responses by comparing them with administrative records (for example: Pyy-Martikainen and Rendtel (2009); Mathiowetz and Duncan (1988) (unemployment spells), Kreuter et al. (2010) (welfare benefit recipients, employment status, age, citizenship), Bingley and Martinello (2014) (education, income, employment)).
Bound et al. (2001) provide a detailed overview of validation studies analyzing labor related phenomena such as: *(1) earnings, (2) transfer program income, (3) assets, (4) working hours, (5) unemployment, (6) labor force status, and transition to and from unemployment (7) occupation*, as well as health related variables such as: *(1) health care utilization, health insurance, and expenditures, (2) health conditions or education.*

Unlike the topics mentioned above, the goal of this chapter is to validate a variable which is assumed to be unaffected by socially desirable answering behavior, to learn more about recall error in survey data. In addition, the selection of an adequate variable is limited to information for which external validation data are available. One variable within SHARE which fulfils both conditions (not socially desirable and the availability of external validation data) is the year of retirement. This variable seems to be especially suitable for a validation as it is (1) not personally sensitive, (2) an event which takes place in most people's lives, (3) of special interest for SHARE, as retirement is one key aspect of the survey, (4) an event which already took place for a large fraction of the SHARE population (50+), and (5) retrospectively collected with a huge variance in how long that event dates back over respondents.

Transition into Retirement

The transition into retirement is an important life event for most people, not only because active working life stops but also because a new episode in peoples' lives, the so called 'sunset years,' starts. Researchers of different disciplines and with different focuses are using that event either as a dependent or independent variable. Some authors analyze the factors and circumstances which can influence people's decisions to retire, for example, their health status (Dwyer and Mitchell, 1999), a women's own reproductive history (Hank and Korbmacher, 2011, 2013), informal caregiving (Dentingen and Clarkberg, 2002) or the economic crisis (Meschi et al., 2013). Another strand of research explores the consequences of retirement, for example with regard to cognitive functions (Mazzonna and Peracchi, 2012; Börsch-Supan and Schuth, 2013), health (van Solinge, 2007), social networks (Börsch-Supan and Schuth, 2013) or even aspects such as smoking cessation (Lang et al., 2007).

In Germany, as well as in many other European countries, different political reforms changing the retirement age require research on how people react to these reforms. To analyze peoples' behavior it is important to know how valid the self reports are. It is well known that survey data suffer from measurement error, but most models assume a classical error which implies that the error one variable is independent of the true value, independent of the other variables which are in the model as well as their respective measurement errors, and independent of the stochastic disturbance (Bound et al., 2001). A violation of these assumptions can have far-reaching consequences. In the worst case, it exists a systematic error which is correlated with the other variables in the model. If, for example, women have a tendency to report their year of retirement earlier than it took place, the mean retirement age of women would be underestimated and wrong conclusions could be drawn.

As far as I know, nothing is yet known about how good respondents are in reporting the year they retired. The project SHARE-RV, which combines survey data of the German sub-sample of SHARE with administrative records of the German Pension Fund, provides a unique possibility to validate respondents' answers with external and very reliable data. This comparison should help in answering the question whether recall error is an issue also for such key events as the year of retirement.

The chapter is structured as follows: Section 6.2 describes the validation of the year of retirement based on the comparison of survey and administrative data. Sections 6.3 and 6.4 focus on the psychological model of the response process and the aspects which are hypothesized to be relevant to explain recall error in the year of retirement. Section 6.5 provides the model and results whereby Section 6.6 closes with some final conclusions.

6.2 Validating the Year of Retirement Using SHARE-RV

The project SHARE-RV, which combines SHARE survey data with administrative records of the same person (see Chapter 2.2), allows analyzing the error respondents make when reporting their year of retirement, as this information is included in both datasets. The data used for this chapter is based on the German sub-sample of the fourth wave of data collection. This sample consists of respondents of the panel sample (Release 2-0-0) which participated for at least two waves of data collection and respondents of a refreshment sample (unpublished internal data, see Chapter 2.2) which participated for the first time. To link respondents' survey data with their administrative records requires respondents' written consent. For the respondents of the panel sample, consent was collected in the third wave of SHARE, whereas respondents of the refreshment sample were asked for consent in the fourth wave.[41] Unlike to Chapter 3, I'll not report the consent rate but combine consent, availability, and the 'linkability' of the administrative records into a linkage rate (see Korbmacher and Czaplicki, 2014). The linkage rate is 48.5% for the panel sample and 34.3% for the refreshment sample.[42] These linkage rates include both the data of the employment histories (Versichertenkontenstichprobe: VSKT) as well as pension data (Rentenbestand: RTBN); in other words, respondents are counted as linked if either the VSKT or RTBN data is available and linkable. The sample for the following analyses is based on cases which could be linked with the RTBN, as this dataset includes the variable of interest. The sample consists of 851 respondents who receive some kind of old age pension (based on the administrative records, see Table 8).

Table 8: Overview: Linked Cases by Sample

	Panel	Refreshment	Both
Number of cases	1,572	1,463	3,035
Number of linked cases	559	292	851

The most recent version of the RTBN records refer to the calendar year 2012 and had been made available in autumn 2013 by the German Pension Fund. The fieldwork of Germany's fourth wave of SHARE took place from the beginning of 2011 until spring 2012. As a consequence, the reporting year of the administrative records and the survey data are not completely overlapping. For the validation of the year of retirement, this would lead to discrepancies for respondents who retired between 2011 and 2012, more

[41] By mistake, some interviewers also asked panel respondents for consent in wave four so that the consent rate reported in Chapter 3 is a bit lower.

[42] Compared to the panel sample, the linkage rate for the refreshment sample is much lower. This is due to the fact that only 80% of the refreshment sample should have been asked for consent. In addition, some problems during fieldwork make it impossible to link all records, so that a consent rate cannot be calculated for the refreshment sample.

precisely: after the SHARE interview but before the end of 2012 (the release version of the RTBN). For these cases, the administrative data and the survey data would not match with regard to the employment status. This holds for 57 cases, which are dropped for the following analyses.

In SHARE, respondents are asked about their current employment status by choosing **one** of the following categories *(1) Retired, (2) Employed or self-employed (including working for family business), (3) Unemployed, (4) Permanently sick or disabled, (5) Homemaker, (97) Other (Rentier, Living off own property, Student, Doing voluntary work).*[43] Only if the respondents declare that they are retired, are they asked about the year in which they retired.[44] Respondents for whom the status is not unique (for example, working part-time and also being retired) have to decide which status best describes their current job situation. As Table 9 shows, 88% of the respondents who are officially retired (based on the administrative records) also declare themselves as retired. The columns highlighted in red show the respondents with differences between their self-reported and their official employment status. Within the 12% of the respondents who deviate in their answer from the records, it exists a clear difference between male and female respondents. Overall, the agreement between the administrative data and the self-reports is much higher for men than for women (92.6% vs. 83.2%). Male retirees who do not declare themselves as retired declare themselves as either employed or sick.[45] In contrast, the majority of female retirees with deviations are homemakers.

The administrative records provided by the German Pension Fund include two variables about the year of retirement: the starting year of the first benefit period and the starting year of the actual benefit period. For most cases (83%) these two dates are the same. Differences between the two values indicate that the kind of benefit they receive had changed. This occurs for example for respondents who receive(d) a disability pension (*Erwerbsminderungsrente*): the year of the beginning of this status is reported in the first variable, and the year the respondent reaches the official retirement age is reported in the second variable. A difference between the beginning of the first benefit period and the beginning of the actual benefit period exist only for 114 respondents. The majority of the respective respondents (N=60) reported in the survey the year of the first beginning, six respondents reported the year of the actual period. For 48 cases, neither the first nor the actual benefit period matches exactly with the self report. I generated one variable which combines this information by using the year with the smallest deviation.

[43] Question ep005: Please look at card 18. In general, which of the following best describes your current employment situation?

[44] the month is only asked if respondents retired after 2008 and will therefore not be validated

[45] People being permanently sick or disabled can receive a "*Erwerbsminderungsrente*" which is coded as pension benefit in the administrative data. The respondents declaring themselfes as sick are all receiving this kind of benefit.

Table 9: Self-reported Job Situation for Respondents who are Retired (Based on Administrative Records) by Gender

			Male		Female	
Self-reportd job situation	Freq.	%	Freq.	%	Freq.	%
Retired	699	88.0	377	92.6	322	83.2
Employed or self-employed	20	2.5	10	2.5	10	2.6
Unemployed	0	0	0	0	0	0
Permanently sick or disabled	36	4.5	17	4.2	19	4.9
Homemaker	31	3.9	0	0	31	8.0
Other (specify)	3	0.4	1	0.3	2	0.5
Missing	5	0.6	2	0.5	3	0.8
Total	794	100	407	100	387	100

In the following, I refer to the difference between the value provided in the administrative records and the reports of the respondents in the interview (see Bound et al., 1994). The underlying assumption is that the administrative records provide the "true" value, and are error free. This is of course a strong assumption which can be doubted, as recent work about measurement error in administrative data shows (see Groen, 2012). Administrative data are defined as data that are not primarily generated as a research source and are routinely collected by agencies (Calderwood and Lessof, 2009). Therefore the term 'administrative data' covers a diversity of data sources, which can greatly differ not only in their content and the purpose they are collected for but also in the methods of their production, and consequently also in their quality. From my point of view, whether the administrative data should be used as a 'gold standard' to validate the survey data should be evaluated for each variable separately. For the variable discussed here (the year of retirement), the administrative data are assumed to be of very good quality, as they are first-hand information from the institution regulating and paying the benefits.

Recall error is here defined as the difference between the survey response and the true value and is calculated as

$$dif_{year_{abs}} = |year_{reported} - year_{admin}| \tag{10}$$

$dif_{year_{abs}}$ is the absolute deviation between the report of the respondent ($year_{reported}$) and the value provided by the administrative data ($year_{admin}$). Figure 13 illustrates the differences between the year of retirement reported by the respondent and the year of retirement provided by the German Pension Fund. All respondents who provide the same answer in the survey as is stored in their records are marked on the diagonal.

As a random noise is added to the graph (by the command jitter (1)in Stata), a small deviation from the diagonal is not a real misreporting but due to the jittering. The figure shows that most of the respondents are on the diagonal, so in general respondents are accurate in reporting their year of retirement.

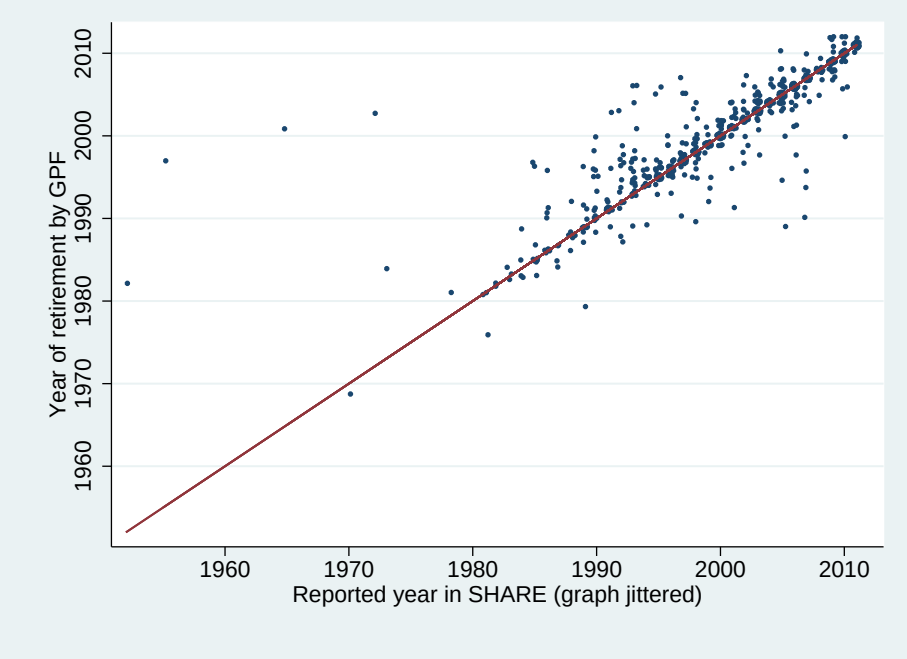


Figure 13: Difference in Reported and True Values of the Year of Retirement

To provide a better impression of the errors' extent, Figure 14 reports the distribution of the absolute difference in years between the two data sources. Deviations of more than 10 years are combined into the last category (10 years). The histogram confirms the impression from Figure 13: more than 60% of the respondents report the year correctly. Conversely, this means that about 40% of the respondents misreport the event, mainly within a range of three years.

As a first descriptive result, we see that even a very important event in a respondent's life, the year of retirement, is affected by recall error. Based on this result, the question arises whether the determinants increasing the likelihood of an error are identifiable. In the following, I use the 'Model of the Response Process' described by Tourangeau et al. (2000) to identify determinants that are assumed to affect the correctness of the respondent's reports.

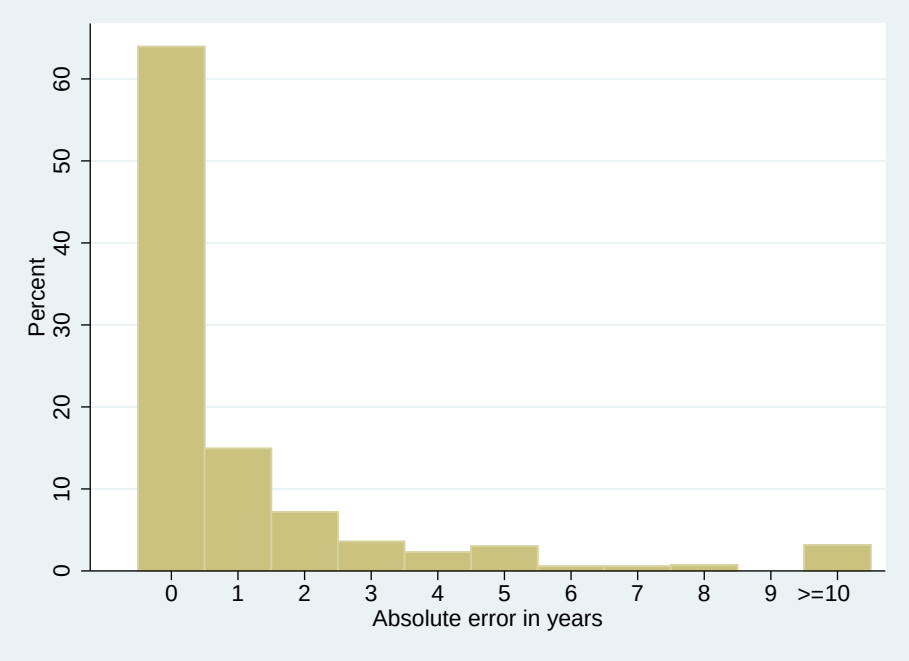


Figure 14: Distribution of Absolute Error

6.3 A Psychological Model of the Response Process

There is a long history of psychological research on the processes which occur when answering survey questions. Most models agree on the fact and the content of several tasks, which are necessary to come to an answer (Sudman et al., 1996). I focus on the model proposed by Tourangeau et al. (2000), as this is one of the most recent models, taking previous research into account. The model described by Tourangeau et al. (2000) is based on four major components of the survey response process. Each of the components is allocated to specific processes, as displayed in Table 10. In more detail, the steps entail the following:

- **Comprehension of the question**: This step is essential, as if the respondent misunderstands the question, the construct the researcher intends to measure and the construct the respondent's answer refers to, are not the same. Therefore the wording of a question is very important. Such aspects as grammar, ambiguous or vague words, and complex formulations can affect the comprehension of the question.

- **Retrieval**: If it is clear what the question is about, the respondents recall the relevant information from memory in this step. "Retrieval refers to the process of

bringing information held in long-term storage to an active state, in which it can be used" (Tourangeau et al., 2000, page 77). This process differs for factual and attitudinal questions, as for the latter there is the possibility that the respondent never thought about the issue before. In the following, only autobiographical facts are considered. What is retrieved from memory is not the experience itself, but a representation of it. The demands are very different for different questions. For example, questions can refer to stable characteristics, meaning that the answer is independent of the point of time the question is asked (as, e.g., the year of birth), or they can be dependent on the time the question is asked (e.g., the age).

- **Judgement**: If the result of the last step (retrieval of information) is not an explicit answer to the question, the step of judgement combines or supplements the information retrieved from memory to assemble an adequate answer.
- **Response**: This is the final step in the process, which is selecting and reporting the answer. The respondents have to adapt their result to the response options of the question. In addition, they can also decide to not provide the answer by answering 'don't know' or refusing to answer.

Table 10: Components of the Response Process (Tourangeau et al., 2000)

Component	Specific Processes
Comprehension	Attend to questions and instructions
	Represent logical form of question
	Identify question focus (information sought)
	Link key terms to relevant concepts
Retrieval	Generate retrieval strategy and cues
	Retrieve specific, generic memories
	Fill in missing details
Judgment	Assess completeness and relevance of memories
	Draw inferences based on accessibility
	Integrate material retrieved
	Make estimate based on partial retrieval
Response	Map judgment onto response category
	Edit response

The authors also state that it cannot be ruled out that some steps are overlapping or indistinct, or that respondents jump back to an earlier step within the process. The model also allows for skipping single steps, if for example respondents' are unwilling to answer and, hence, say 'don't know' even before the very first step. Factors as respondents' motivation to answer accurately or the time they have to answer can influence which steps are skipped.

The Response Process when Asking About the Year of Retirement

The model of Tourangeau et al. (2000) describes the processes when answering a survey question in a general way. This model will now be adapted to the specific autobiographical event, the year of retirement, which is asked in SHARE as well as many other surveys. Following Tourangeau et al. (2000), the question is categorized as a 'time of occurrence' question, as it asks about the date an event happened. Beginning with the first step (comprehension of the question), the exact wording of the question should be considered. The generic English question reads as follows:

- "In which year did you retire?"

At first glance, the question is not complex, and does not include any ambiguous words or terminologies, so that one could assume that the comprehension of the question is not problematic. Nevertheless, a closer look at the wording of the question shows that there is a potential for misunderstanding: Based on the "Longman Online Dicitionary," the definition of 'to retire' is as follows: "to stop working, usually because you have reached a certain age[46]". The focus of the generic wording is not on beginning the period of retirement but rather on stopping the working period. This impression is also confirmed by Rust (1990), who discusses the ambiguity of the English term 'to retire.' He provides some interpretations that respondents may have in mind when declaring themselves as retired. They all refer to quitting the career job. His example shows that respondents can define themselves as retired even if they are working full-time but quit their career job (see Rust, 1990). The meaning of that phrase is different in German, where an equivalent verb does not exist. The German translation is:

- "*In welchem Jahr sind Sie in Rente gegangen?*"

"*Rente*" is defined as "*regelmäßiger, monatlich zu zahlender Geldbetrag, der jemandem als Einkommen aufgrund einer [gesetzlichen] Versicherung bei Erreichen einer bestimmten Altersgrenze, bei Erwerbsunfähigkeit o.Ä. zusteht*[47]" which is a regular, monthly payment a person receives when reaching a given age because of a [legal] insurance [...]. The focus of the German wording is rather on entering into retirement than on leaving

[46] www.http://www.ldoceonline.com/dictionary/retire
[47] www.http://www.duden.de/rechtschreibung/Rente

the workforce. Even if the German wording seems to match the administrative data, one could not rule out that respondents differ in their interpretation of the question. To better understand how German respondents interpret the question, I used the fact that the respondents of the refreshment sample are asked three different questions: first, the year they retired, second, the year they stopped working, and third, the year they received a pension for the first time. A comparison of these three answers shows that most of the respondents link the question with the concept of receiving a pension (see Appendix E.1 for more details).

The second step of the response process is the retrieval of the requested information: the year the respondent retired. The most obvious determinant here is how much time passed since the requested event occurred. Respondents who recently retired should remember the exact year better than respondents who retired a long time ago. As no reference period is given in the question[48], the answer can refer to a great range of years. It is generally recognized that the longer the timelag between the event and the interview, the less likely it is that people remember it correctly. One explanation of that effect is that with passing of the time, the chance that the same event occurs again increases. This makes it harder for the respondents to distinguish between the events (Tourangeau et al., 2000). For the example discussed here (the year of retirement) it is very unlikely that the same event takes place twice, as for most respondents this is a non-repeating event. However, there are exceptions, as the next section will show. In addition, the salience and importance of an event influences how well it is remembered (Eisenhower et al., 2004).

Once the event is recalled, it has to be adapted to the correct format of the question. People may differ in whether they remember the exact year, a range of plausible years, or their age when they retired. In the latter case, this form of representation requires that respondents convert their answer from age into calendar time. Depending on the respondent's cognitive abilities, this step could be seen as another source of error. If they are not sure about the exact date, they have to decide whether they answer with an approximation, answer that they do not know the date, or use a typical date, such as the legal retirement age.

The last step, reporting the answer is expected to be rather easy, as the question clearly indicates that a year is requested. The answer does not have to be allocated to a response category or formulated as for an open ended question.

To sum up, respondents' cognitive abilities, as well as the characteristics of the event, are assumed to influence the response process and therewith the accuracy of the reporting.

[48] Some questions refer to a given time period as '*during the past 12 months...*' or '*since our last interview...*'

When the Process Fails

In the best case scenario, respondents are asked about the year they retired, they retrieve the event which is stored in memory with the exact date, and they report that. In the second best case scenario, the information of the year is not available immediately but as the respondents make some effort they do remember the year. In both of these cases, the difference between the self-reported year and the year provided by the German Pension Fund is zero. If the worst comes to the worst, respondents do not remember the year, they do make some effort to come up with a plausible value, but it is not the correct one. This last case is of interest here: people who misreport the year they retired. The goal here is to learn more about the mechanism behind that error. The focus is on the question of whether the respondents' cognitive abilities and/or the characteristics of the event can help to explain the errors the respondents make. In addition, two other aspects are discussed: rounding to prominent years as well as respondents' gender. I'll first discuss these two additional aspects, and then focus on cognitive abilities and employment history. All aspects, their operationalization as well as some bivariate results, will be discussed in the following. The results of the multivariate analyses will be discussed in Chapter 6.5.

6.4 Predictors of Recall Error

Rounding and Heaping

One source of the error which often occurs when asking respondents retrospectively about the calender year an event took place is rounding (e.g. Torelli and Trivellato, 1993; Bar and Lillard, 2012). The consequence of rounding to specific values is the heaping effect, which is "an abnormal concentration of responses at certain [...] dates (for questions asking when an event took place), where 'abnormality' results with respect to external validation data or reasonable *a priori* expectations about the smoothness of the frequency distribution." (Torelli and Trivellato, 1993, page 189). The years I define as *prominent years* are those which are decades or multiples of five-year spans (for example, the years 1970, 1975, 1980, 1985, 1990, and so on). The distribution of the reported years is shown in Figure 15. The red lines indicate the years which would occur disproportionately if the respondents round. The results show no clear hint for heaping at these prominent years in comparison to the other years.[49] In addition, if respondents round, the share of prominent years would be higher in the self reports than in the administrative data. To compare these two shares, I generated two dummy variables, one for the reported year and one for the true year which are one if the year is a multiple of five. The result are displayed in Table 11. At first glance, the share of prominent years

[49] The same graph based on the administrative data can be found in Appendix E.2; the comparison of the two does not show a clear pattern indicating that respondents round.

is slightly higher in SHARE than in the administrative data. But the paired t-test shows that the H-0 (the difference between the two means equals zero) cannot be rejected.[50] Therefore, the difference between reports in the administrative data and the SHARE data is not statistically significant.

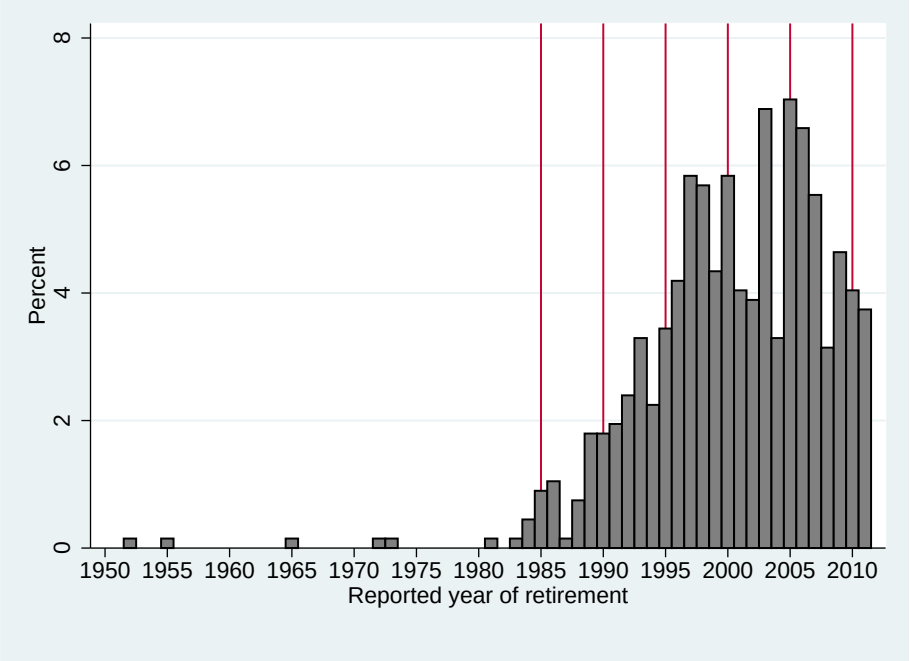


Figure 15: Distribution of Reported Years

Table 11: Comparison of Prominent Years in Self-reports and Administrative Data

	prominent years	non-prominent years
SHARE	156 (23,35 %)	512 (76,65 %)
Admin data	143 (21,41 %)	525 (78,59%)

[50]The corresponding two-tailed p-value is 0.2004

Gender

Following general stereotypes about the differences between men and women would lead to the assumption that women are better at remembering the dates of events. Men are often depicted as the ones who forget birthdays, anniversaries, or other events (Skowronski and Thompson, 1990). Interestingly, there is also empirical evidence that men and women differ in how good they are in reporting the date of autobiographical events. For example, Skowronski and Thompson (1990) found that female students are better at remembering the dates of events they recorded in diaries than are male students. Based on these results, Auriat (1993) compared reports of residential moves with register data and found that female respondents are better at dating the moves than are male respondents. If the result of Skowronski and Thompson (1990) is valid in general, females should be the more accurate daters and recall error in reporting retirement should be less likely for female respondents. A bivariate consideration of the absolute error and respondents' gender cannot confirm the results cited above. Men and women do not significantly differ in how well they remember their year of retirement (see Table 12). Nevertheless, respondents' gender will be included in the multivariate model, as a control variable. In addition, respondents' gender could be especially important in the context of working history. To test whether the effects of respondents' working history differ between men and women, I also include interaction terms of gender and some aspects of the working history.

Table 12: Mean Absolute Error by Gender

Gender	Mean error	Std. error	Frequency
Male	1.12	0.11	370
Female	1.07	0.13	298
Combined	1.10	0.85	668

Cognitive Abilities

As cognitive abilities are a fundamental aspect of aging (Mazzonna and Peracchi, 2012), SHARE implemented a module of questions which measures respondents' cognitive abilities in different ways. This module consists of items about self-rated skills of reading, writing, and memory, and some objective tests which measure orientation in time, memory, verbal fluency, and numeracy. Not all respondents have to answer all questions, as the routing differs for the refreshment and the panel sample. Therefore, only those questions can be considered, which are asked of all respondents. These questions are described in the following.

- Serial numeracy: Respondents are asked to subtract the number 7 five times, starting from 100. The interviewer notes the respondents' answer without commenting on whether or not the result is correct. The exercise stops if the respondent refuses or answers "don't know" for the first time, or after five subtractions at the latest. Therefore, the number of correct answers can vary between 0 and 5. In addition to mistakes the respondents can make, this variable is prone to errors the interviewers make while entering the numbers. I cleaned the variable by correcting for obvious typos as transposed digits. I decided to allow for subsequent mistakes when counting the number of correct answers, as otherwise the ability to subtract seven would be underestimated. The counter of correct answers adds one if the result of substraction is seven less than the result answered before, independently of the correctness of the result answered before.
  As Table 13 shows, there is little variation in respondents' calculation ability when referring to the German Wave 4 sample.[51] The majority of respondents (67%) made no mistakes and 19% made only one mistake.

Table 13: Serial Numeracy: Number of Correct Answers

Correct answers	Frequency	%
Refused	68	2.24
0	2	0.07
1	43	1.42
2	58	1.91
3	184	6.06
4	569	18.75
5	2,042	67.28
Not applicable	69	2.27
Total	3,035	100

Another dimension of cognitive abilities, discussed by Mazzonna and Peracchi (2012), is respondents' processing speed. The authors argue that it is important to also consider the time respondents took to arrive at an answer. Respondents who answer all the questions correctly but took a long time should be rated with less cognitive skills than a respondents who gave the same number of correct answers in a very short time. Using the keystroke variables collected during the SHARE

[51] The category 'not applicable' results from the fact that SHARE allows for proxy interviews for most of the modules. The cognitive functions module is excluded, so that all questions of that module are skipped. I excluded all interviews where a proxy was included, to ensure that the respondent answered all questions herself/himself.

interview allows considering the time respondents needed to arrive at an answer. According to Mazzonna and Peracchi (2012), I first grouped respondents by their number of correct answers (0 - 5) and within each group by the time they needed to answer per question. But as the time recorded by the instrument is also influenced by the interviewer (Mazzonna and Peracchi, 2012), I also take the interviewer into account. To do so, I calculated the time the respondent took net of the interviewer average (exclusive of the current interview) and grouped it into terciles. The variable now consists of 16 categories: one for respondents with zero correct answers, and the 3 terciles for each number of correct answers. Table 13 gives an example of how the outcomes are categorized.

Table 14: Example: Number of Correct Answers Including Response Time

Correct answers	Tercile	Category
0	-	0
1	third	1
1	second	2
1	first	3
2	third	4
2	second	5
2	first	6

- Verbal Fluency: Respondents are asked to name as many animals as possible within one minute of time. The instrument is programmed in a way that with confirming that the respondent understood the question, a one-minute countdown starts. The interviewer is instructed to note all animals on a separate paper. When the minute is over, the interviewer enters the total number of valid answers into the CAPI. On average, respondents named about 21 animals with a minimum of 1 and a maximum of 49 animals.

- Ten-word learning list: This is a test of verbal learning and memory which is based on Rey's Auditory Verbal Learning Test (RAVLT) (Dal Bianco et al., 2013). Respondents are randomly assigned to one of four different lists of ten common words.[52] To minimize interviewer effects, the words which should be read out by the interviewers always appear on the screen in the same time interval. When the interviewer has read out all words, the respondents are asked to repeat those they remember (immediate recall). At the end of the same module, they are asked again which of the words they still remember (delayed recall). The result of the

[52]To minimize learning effects, respondents of the panel sample will not get the same list as in the last interview.

so called 'ten-words learning test' is the sum of correctly remembered words from the immediate and the delayed recall. The final variable varies between 0 and 20 correct answers. On average, respondents recall 9.7 words over both questions. As one would expect, the mean of the immediate recall (5.5 words) is higher than the mean of the delayed recall, which is 4.2.

Of course, the different measurements of cognitive abilities refer to different aspects of the memory. It is unclear how these aspects are connected and correlated with those cognitive abilities which are beneficial to the recall of the year of retirement. The correlation matrix of the three measurements shows that verbal fluency and word recalling are highly correlated (0.48), whereas the correlation of the numeracy score with the wording test as well as with the verbal fluency test are only weakly correlated (0.25 each). Therefore, I combined the two highly correlated variables by adding their standardized values into one new variable.

Cognitive Abilities and Recall Error

The scatterplot in Figure 16 shows the correlation of the absolute error in reporting the year of retirement with the combined measure of cognitive functions. There is a clear negative relation between cognitive functions and the errors respondents make, illustrated using the red line, which is the prediction from a linear regression. This negative coefficient of cognitive functions is statistically significant at the 0.01 significance level. In contrast, there is no effect of the third measure of cognitive functions (numeracy) on the absolute error.[53] Therefore I take the numeracy score not into account for the multivariate analysis. The negative effect of cognitive functions is no longer significant when controlling for the respondent's age. This is not surprising, as cognitive abilities are known to decline as people get older. The respondent's age has a positive and statistically significant effect, meaning that the probability of misreporting the year of retirement increases for older respondents. But given that the event itself depends on the respondent's age, the time elapsed since the event took place and respondents' actual age are highly correlated. Therefore the respondent's age is no longer considered but replaced by the number of years between the event and the report[54].

[53] I tested all versions of that variable, namely, (1) the raw number of correct answers, (2) the raw number when considering subsequent faults, (3) a combination of (2) plus the time the respondent needed to answer

[54] Comparing the AIC and BIC of the three models (including cognitive functions and (1) age, (2) elapsed time (3) age and elapsed time) also shows that model (2) has the best fit.

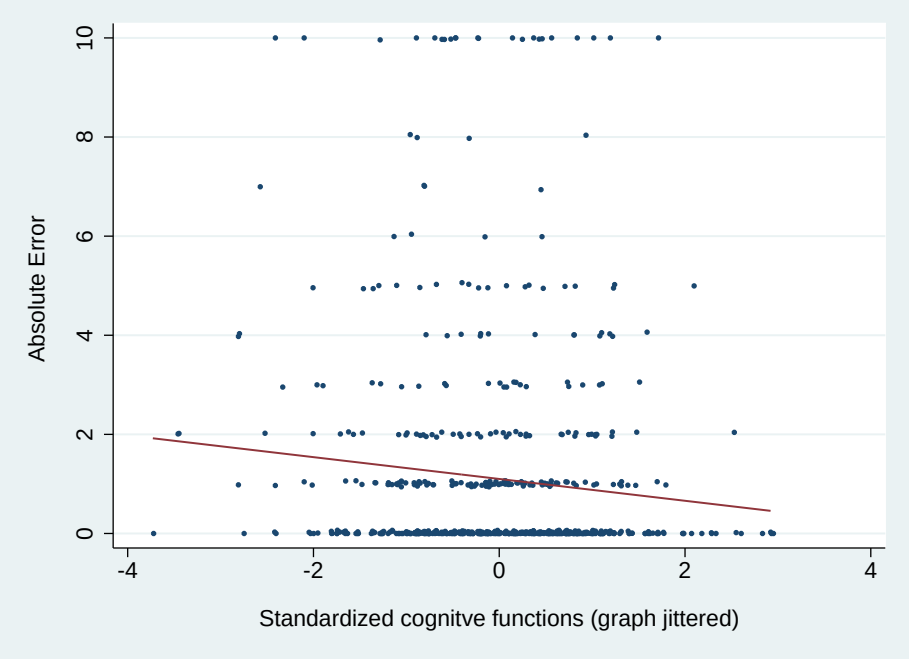


Figure 16: Cognitive Functions and Measurement Error

Characteristics of the Event

Four different aspects are considered with regard to the event which should be remembered. First, the time elapsed since the event took place; second, characteristics of respondents' employment history; third, typical vs. atypical retirement behavior; and fourth whether the true event is close to the turn of the year.

Elapsed Time

There is evidence for a relation between the time elapsed since an event and the difficulty of remembering it (Sudman, 1980; Sudman et al., 1996; Auriat, 1993). But there does not seem to be a general forgetting curve which is the same for all events (Sudman et al., 1996). In addition, as mentioned before, this event typically takes place in later life within the same time span for most people. A descriptive consideration of the correlation of years elapsed since the event with the error is shown in Figure 17. The negative effect of elapsed time is highly significant in this bivariate consideration.
The effects of the two variables, cognitive functions and time-lag, are assumed to be linear. To test whether this assumption holds, a generalized additive model (gam) is calculated. The advantage of this semi-parametric model is that no a priori assumption of the functional form of the effect influences the output. The results of the *gam* confirm the

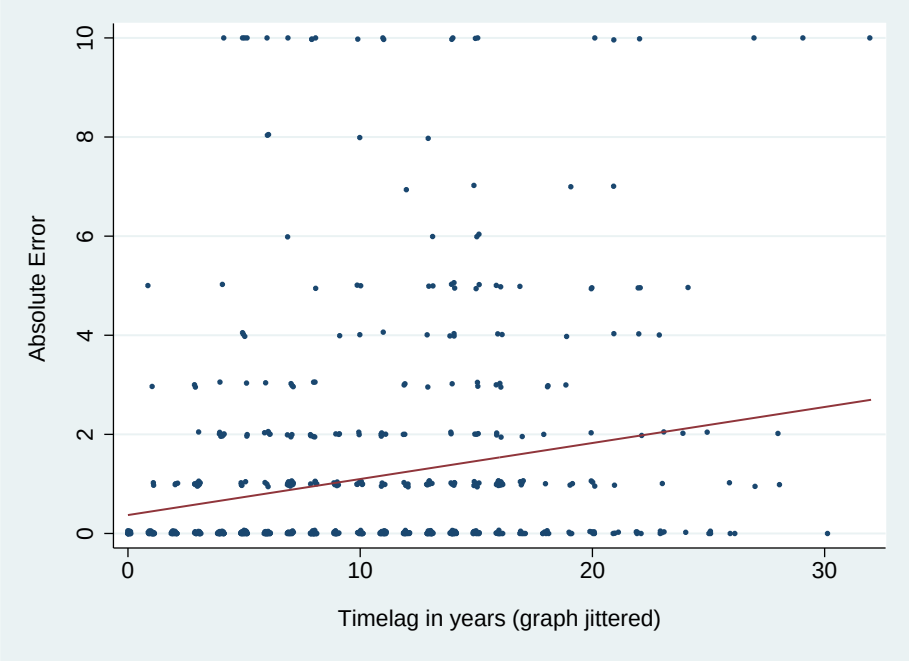


Figure 17: Time-lag and Recall Error

linearity of the effect (results not shown) for cognitive abilities. The results for elapsed time are not that clear. Figure 18 shows the result of the *gam* regression of elapsed time on absolute error controlling for cognitive functions. The red line corresponds to the coefficient of the linear regression. For 98% of the cases the linear effect is within the confidence interval of the effect of the generalized additive model. Strong differences between the two effects are only visible for respondents with more than 24 years between the event and the reporting of the event. As only 14 respondents have a gap of more than 24 years, interpreting the effect as linear seems to be valid. The green line refers to the linear effect when excluding these 14 respondents, to test whether these cases influence the coefficient of the linear regression. As the two lines are very close to each other, the 14 respondents with a very high time-lag do hardly influence the slope of the estimation.

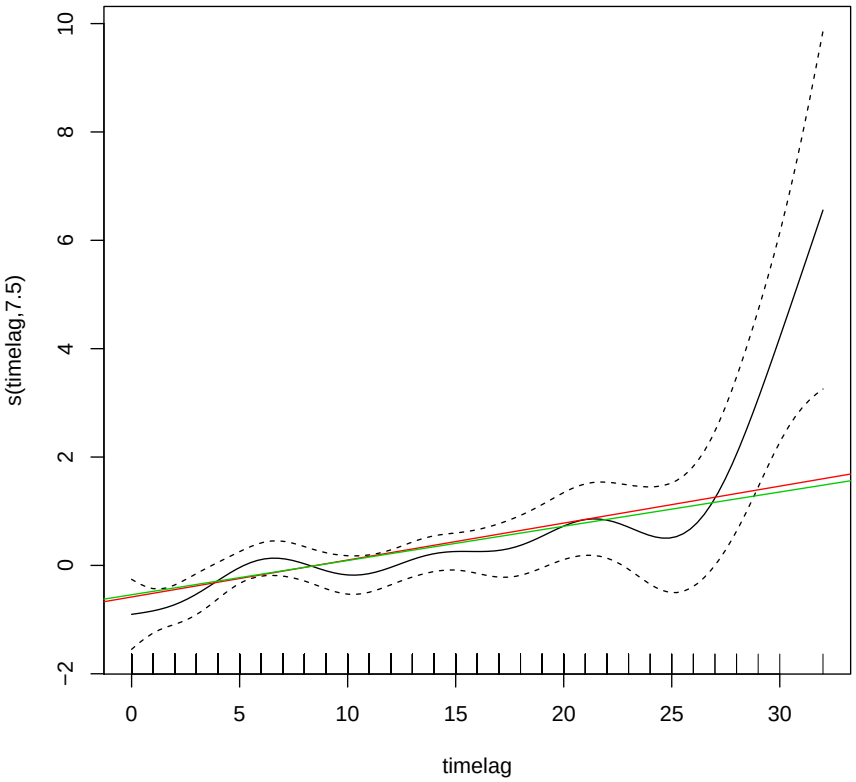


Figure 18: gam Regression: Time-lag and Measurement Error

Employment History

There is evidence that events which are important and salient may be remembered more accurately than less important events (Sudman, 1980). Sudman (1980) names three dimensions to distinguish between more and less salient events: (1) the uniqueness of the event, (2) the economic and social cost or benefits, and (3) continuing consequences. To get an idea of how salient the event of their retirement was for the respondents, I use the last employment episode as a reference point. For people entering into retirement from active employment, the consequences are obvious: first they have much more time at home and second, they have less money, as pensions are lower than salaries. Even if this pattern (employment - retirement) is the one people have in mind when thinking about the transition into retirement, other scenarios are also possible. For example, housewives who worked earlier in their career or accumulated contributions due to education or care-giving enter into retirement as they reach their retirement age. For them, the consequences are less obvious as the daily routines are assumed not to change. The same holds for people who were unemployed. Therefore, I hypothesize that the last employment status matters for the recall error in the reported year in a way that respondents who enter into retirement from active employment are hypothesized to have a better memory of the year this event took place.

The employment histories of the respondents are provided in the administrative dataset of the German Pension Fund. The variable "*Soziale Erwerbssituation*" (*"social employment status"*) differentiates between 15 different statuses (see Table 15). I consider the last status of the employment history as the final one. Some of the categories are not used, for example, education or military service, as these events typically take place earlier in a respondent's life. I add one category to the list: if retirement was not a non-recurring event[55]. This could be the case if respondents receive a disability pension and start working again before they get old age pensions. Only a small proportion of the sample shows this pattern (about 6%), but due to the fact that more than one event could be remembered when asking about the year of retirement, there is an increased chance of a mismatch between the event reported by the respondent and the administrative data. To differentiate between respondents who worked in the last spell before retirement and those who didn't, I summarized the statuses: categories 10-13 are combined as 'working.' Respondents of category 0 (no information is available) are under the summary heading 'not working' [56]), as well as those of category 3 (unpaid care), category 5 (disabled), and categories 6-8 (unemployment). The variable "*Soziale Erwerbssituation*" of the ad-

[55]I defined single spells by working status and kept the spells one before the status was retired. Respondents who have more than one retirement spell are in the category of several retirement spells

[56]Even if is not clear what these people are doing, I label them as not working. The great majority of these people are housewives/househusbands. As being a housewife does not accumulate pension benefits there is no incentive to report this activity to the Pension Fund.

ministrative data contains a surprisingly high number of missing values (about 11%), which would decrease the number of cases for analysis. To not lose these cases, I added the dummy "missing" which is one if the longitudinal employment biographies are not available.

To summarize, four different dummy variables related to the last employment status are included in the model. (1) a dummy which is one if the respondent's last employment status was working (0 otherwise), (2) one if he or she was not working, (3) a dummy for several retirement spells, and (4) a variable indicating that the administrative dataset is not available (see Table 16). Surprisingly, there are only small differences between men and women when considering their last employment status.

Table 15: Social Employment Status

Code	Last employment status	Combined in dummy	Number of cases	%
0	no information	not working	126	18.86
1	Education (school)	-	0	0
2	Education (voc. training)	-	0	0
3	Care (not paid)	not working	8	1.20
4	Childcare and homemaker	-	0	0
5	Disabled	not working	21	3.14
6	Unemployed & "ALGII"	not working	19	2.84
7	Unemployed & "ALG"	not working	131	19.61
8	Unemployed "*Anrechungszeit*"	not working	37	5.54
9	Military/ civilian service	-	0	0
10	"*Geringfügig Beschäftigt*"	working	14	2.10
11	Self-employed	working	1	0.15
12	Other	working	2	0.30
13	Employed	working	248	37.13
14	"*Zurechnungszeit*"	-	0	0
15	Pension receipt	-	0	0
17	additional cat.: Several spells	Several spells	38	5.69
.	Missing	Missing	23	3.44
		Total	668	100

As the end of the employment history does not reveal information about the whole working history, I added the number of full months for which contributions were paid ('*Vollwertige Beitragszeiten*') as an indicator of a continuous working history . The idea behind this variable is analogous to the previous: the event of leaving the employment market is assumed to be a more influential event for people who have been on the

Table 16: Social Employment Status: as Four Dummy Variables

Dummy	Frequency	%	Male (%)	Female (%)
Working	265	39.67	40.81	38.26
Not working	342	51.20	50.00	52.68
Several ret. spells	38	5.69	6.76	4.36
Missing info	23	3.44	2.43	4.70
N	668			

employment market for a long time. The variable is truncated after 48 years (=576 months) and provided in months. Respondents have on average 373.5 full contribution months, which corresponds to 31 years. As this variable refers to the whole employment history, differences between men and women are bigger than for their last employment status. Women have on average 136 months less than men, which corresponds to more than 11 years.

Another characteristic which is related to the event is the respondent's age at retirement. Here, I do not include the age at retirement but whether it differs from 'typical behavior.' As the legal retirement age changed over time, I count as typical those years with a clear peak in the distribution. Again, differences between men and women are considered. The calculation of the age at retirement is based on the information of the administrative data. For men, three different peaks are visible: at ages 60, 63, and 65; for women at ages 60 and 65. This information is summarized into one dummy variable, which is 1 if the respondent's retired at one of these peak ages, and 0 otherwise. The majority of cases (65%) are classified as 'typical', i.e., the dummy takes the value 1.

The administrative data not only provide the year of retirement but also the month. The month could be especially important for respondents who retired close to the turn of a year. For them to be out by just one month can result in a difference of one year. Therefore, I hypothesize that respondents who retire close to the turn of the year (this is defined as within +/- 2 months around the turn) have a higher chance of misreporting the year they retired. A dummy variable is included in the model, which is one if the respondent retired in November, December, January, or February.

6.5 Model and Results

The Sample

The sample of the following analysis consists of 668 cases. Table 17 gives an overview of the stepwise reduction of the sample of linked cases reported in Table 8. Even if 851 cases could be linked successfully, not all can be used for the analysis. As mentioned before, some of them retired after the SHARE interview, and others didn't declare themselves as retired. In three cases there are hints that the interview was answered by or with help of a proxy. These interviews are dropped, as I cannot rule out that the proxy also answered the question of the year of retirement. The last 28 cases cannot be included as they suffer from item nonresponse on any of the explanatory variables. After excluding all the cases, the final sample consists of 668 respondents.

Table 17: Sample Selection

Number of linked cases	851
Retired after Interview	57
Retirement not reported	95
Proxy interviews	3
Item nonresponse	28
Final Sample	668

The following section is divided into two three parts: the first refers to the absolute error (which is the difference in years of self-reports and the administrative data, independently of the direction of the difference) and the second focusses on the question of systematic error (which also takes the direction of the error into account).

The Absolute Error

Referring to Equation 10 on page 76, the absolute error is considered. The distribution is truncated at a difference of 10 years so that the dependent variable ranges from 0 (no error) to 10 (a maximum difference of 10 years). Figure 14 on page 78 illustrates the distribution of this variable: a very high share of the outcome 0 compared to the alternative outcomes of 1 to 10 (zero-inflation). Per definition, the outcomes can never be negative, but are integer values between 0 and 10. These characteristics are often found in count data and it is well known that using a classical linear regression is mostly inappropriate in that case (Loeys et al., 2012). Therefore, I chose a model which is recommended for count data. In addition, to take into account the possibility that the processes of committing an error at all can differ from the process determining how big the error is, a hurdle regression model is used. It consists of two steps: first a

binary model to predict the zero outcomes, and second a zero-truncated model to predict the non-zero outcomes (Mullahy, 1986). Setting the hurdle to zero can also solve the problem of excess zeros (Farbmacher, 2013). The two separate steps will be described in the following (see Long and Freese, 2006). Step I is a logistic regression to predict the zero outcomes, which refers to making no error. It can be written as:

$$Pr(y_i = 0|x_i) = \frac{exp(x_i\gamma)}{1 + exp(x_i\gamma)} = \pi_i \tag{11}$$

For the second step, I use a zero-truncated negative binomial model. As positive outcomes can only occur if the zero hurdle is passed, the conditional probability is weighted:

$$Pr(y_i|x_i) = (1 - \pi_i)Pr(y_i|y_i > 0, x_i) \text{ for } y > 0 \tag{12}$$

The unconditional rate combines the mean rate for those with $y = 0$ (which is 0) and the mean rate for those with positive outcomes:

$$\mu_i = E(y_i|x_i) = [\pi_i \times 0] + (1 - \pi_i) \times E(y_i|y_i > 0, x_i) \tag{13}$$

In the zero-truncated binomial regression, the conditional mean $E(y_i|y_i > 0, x_i)$ equals:

$$E(y_i|y_i > 0, x_i) = \frac{\mu_i}{1 - (1 + \alpha\mu_i)^{-1/\alpha}} \tag{14}$$

Unlike the Poisson regression, where the conditional mean and the conditional variance are assumed to be equal (equidispersion) (Cameron and Trivedi, 1986), this assumption can be relaxed for the negative binomial regression by adding the α parameter that reflects unobserved heterogeneity among observations (Long and Freese, 2006; Greene, 2008). Different variance–mean relations can be used, two of them are discussed by Cameron and Trivedi (1986): Negbin I and Negbin II. When using truncated models, the assumption of the variance–mean relation is even more important than for non-truncated models, as here not only the standard errors can be biased but also the estimated βs. As mentioned before, the assumed variance–mean relation of the Poisson model is

$$Var(y_i|x_i) = E(y_i|x_i) = \mu_i = exp(x_i\beta) \tag{15}$$

The Negbin I model implies a constant variance–mean ratio and can be written as[57]

$$Var(y_i|x_i) = \mu_i + \alpha\mu_i \tag{16}$$

The Negbin II model implies a variance–mean ration which is linear in the mean:

$$Var(y_i|x_i) = \mu_i + \alpha\mu_i^2 \tag{17}$$

[57]The following formulas refer to the normal negative binomial regression model. When referring to the zero-truncated model, $Var(y_i|x_i)$ has to be replaced by $Var(y_i|y_i > 0, x_i)$.

An even more flexible way to model the variance–mean relation is the Negbin P model introduced by Greene (2008). In this model, the exponent of the term $\alpha\mu_i$ is replaced by P, which is also estimated. Consequently, $P = 0$ refers to the Poisson regession model, $P = 1$ refers to Negbin I, and $P = 2$ refers to Negbin II.

$$Var(y_i|x_i) = \mu_i + \alpha\mu_i^P \tag{18}$$

Following Farbmacher (2013), I calculated three different Negbin versions (I, II, and P) to find the adequate model with the best model fit.[58] Table 18 shows the results of the hurdle regression. Column 1 refers to the first step (a logistic regression of passing the hurdle), columns 2 to 4 refer to the Negbin I, Negbin II, and Negbin P model, respectively.

Model (1) is the logistic regression with a dependent variable which is 1 if the respondents make no error and 0 otherwise. The interpretation of the signs of the effects is the following: a negative coefficient represents a smaller chance of making **no** error, the reverse represents a higher chance of making an error.

Unlike the findings that women are better at remembering events, the effect of gender goes in the opposite direction but is not statistically significant. Significant influences can be found for the interaction terms of gender and employment history. As expected, respondents' cognitive abilities significantly influence the chance of making an error at all. The better the respondents perform in the two cognitive functions tests, the higher the likelihood that they do not make an error in reporting the year of retirement. The time-lag between the year the event occurred and the year the question was asked also shows the expected effect. The longer the event dates back, the higher the chance respondents misreport the year of the event. Respondents who didn't work before they retired, as well as respondents who had several retirement spells, have a significant higher chance of misreporting the year of retirement. These effects do not significantly differ for men and women, as the interaction terms show ('Male*not work.' and 'Male*several'). The result is different for the effect of the number of full contribution months: the main effect does not show a significant effect but the interaction with respondent's gender ('Male*month') does. When predicting the marginal effect for men and women separately, the effect is negative but not significant for women and positive and significant at the 5% significance level for men. Therefore, the interpretation of the effect is that the more contribution months men have, the more likely it is that they report the year correctly. A significant interaction term can also be found for the effect of the dummy variable that indicates whether the respondent retired at a typical age. This effect is positive and highly significant for women but close to zero and not significant for men.

[58] Negin I and II are implemented in Stata's command '*ztnb*' for zero- truncated negative binomial models by changing the parametrization of the dispersion (mean is the default); to calculate the Negbin P, I used the ado '*ztnbp*' which was programmed by Helmut Farbmacher (see Farbmacher (2013)).

Table 18: Hurdle Regression Model of Absolute Error

	(1) Logit	(2) NegBin-I	(3) NegBin-II	(4) NegBin-P
Male	0.40	-1.32	-0.33	-1.10
	(0.35)	(1.57)	(0.47)	(0.92)
Cognitive Functions	0.20**	0.08	0.07	0.08
	(0.09)	(0.12)	(0.08)	(0.12)
Time-lag (years)	-0.07***	0.04**	0.04**	0.04**
	(0.01)	(0.02)	(0.02)	(0.02)
Not working	-0.83**	0.12	-0.06	0.09
	(0.34)	(0.50)	(0.29)	(0.46)
Several spells	-1.77**	1.39***	1.04***	1.32**
	(0.75)	(0.50)	(0.35)	(0.56)
Contribution months	-0.07	0.10	0.05	0.10
	(0.16)	(0.19)	(0.16)	(0.19)
Typical ret. age	1.25***	0.67*	0.42*	0.62
	(0.29)	(0.40)	(0.24)	(0.40)
Turn of year	-0.35*	-0.35	-0.37*	-0.40
	(0.19)	(0.33)	(0.19)	(0.35)
Interactions				
Male*not work.	-0.11	-0.31	-0.23	-0.34
	(0.39)	(0.61)	(0.44)	(0.60)
Male*several	1.37	-0.86	-0.87	-0.70
	(0.87)	(0.81)	(0.60)	(0.78)
Male*months	0.36*	-0.56**	-0.44**	-0.58**
	(0.21)	(0.25)	(0.18)	(0.23)
Male*typical	-1.24***	1.36	0.29	1.14
	(0.38)	(1.55)	(0.45)	(0.90)
Miss data	-0.58	0.53	-0.14	0.45
	(0.50)	(0.34)	(0.34)	(0.41)
Constant	1.25***	-0.16	0.17	-0.12
	(0.30)	(0.60)	(0.36)	(0.55)
δ		2.02		
		(0.42)		
α			1.19	1.98
			(0.47)	(0.51)
P		1.00	2.00	1.16
		(fixed)	(fixed)	(0.29)
N	668	243	243	243
ll	-389.63	-434.73	-438.01	-434.37

*, **, *** mark significance on the 10, 5, 1 percent level, respectively

Dependent variables: making no error (1); years of difference (2)-(4) if error > 0

Robust standard errors in parentheses, clustered by interviewer

The dummy variable indicating whether the event was close to the turn of the year also shows the expected effect: respondents who retire +/− 2 month around the turn of the year have a significantly higher chance of misreporting the year.
Models (2) to (4) refer to the second step of the hurdle regression model: the zero-truncated negative binomial model for those respondents who misreport the year of retirement. Excluded are all respondents who reported the event correctly. As discussed above, the three models differ in the assumption of the variance–mean relationship. In all three models, a positive Alpha (or Delta) indicates that the data is overdispersed, so that a Poisson regression model would not only (downward) bias the standard errors but given that the model is truncated, also bias the estimated βs (Long and Freese, 2006; Farbmacher, 2013). When comparing the log likelihoods of the Negbin I (model (2)) and the Negbin II (model (3)) regression model, the first has the better model fit. The log likelihood of the Negbin P model is very close to that of the Negbin I, which is not surprising as the estimated P is 1.16 and therewith very close to the Negbin I model. The confidence interval of the P also shows that 1.16 is not significantly different from 1. As the Negbin-P model has a slightly better fit, I use that one to interpret the results. The interpretation of the signs of the coefficients is different from the first model. Here a positive coefficient shows that the variable increases the error (Long and Freese, 2006, page 389). Unlike the first step, respondents' cognitive abilities no longer show a significant effect, indicating that good cognitive abilities decrease the chance of making an error, but given that there is an error, they do not significantly influence how big the difference in years is. That's different for the time effect. The number of years between the event and the report influence both, the chance of making an error and the amount of the error. The more years have passed between the two points, the higher the error the respondents make. The same pattern occurs for the effect of several retirement spells and the number of contribution months (for men): a significant and consistent effect can also be found in the second step. Two variables which have been significant in the first step are no longer significant in the second step, namely the effect of retirement age (typical vs. not) and whether the event occurred close to the turn of the year.

Recall Error and Bias

The prior section aimed at finding the determinants of reporting errors, whereat the error is defined as the absolute deviation between the date reported by the respondents and the administrative data. This approach allows us to learn more about whether respondents do report the year of retirement correctly and which characteristics can influence the correctness of the answer. The results show that there is some error in respondents' reports which can partly be explained. To assess the consequences of these errors for empirical analyses it is important to know whether or not these errors are systematic. This means that (depending on other variables), the error goes in one specific direction.

A hypothetical example for such a systematic error would be if men in general give more recent dates for the event and women do not. To learn more about a potential systematic of the error, in the following the absolute error is replaced by the normal error expressed as:

$$dif_{year_{total}} = year_{reported} - year_{admin} \tag{19}$$

A positive value indicates that respondents report the event later than it took place, whereas a negative value indicates the respondent reported the event earlier than it took place. One phenomenon often discussed when referring to the dating of autobiographical events is 'telescoping'[59] (Sudman et al., 1996; Rubin and Baddeley, 1989; Huttenlocher et al., 1988). That is "the report of a too recent date for an even" (Huttenlocher et al., 1988, page 471), which would be a positive error in terms of Formula (19). Huttenlocher et al. (1988) and Rubin and Baddeley (1989) analyzed this effect by assuming that the events are not stored incorrectly, but errors occur within the retrieval process (Sudman et al., 1996). The effect of 'telescoping' is based on three factors: (1) retention is greater for events which took place more recently, (2) errors that occur when remembering events increase with time since the event, (3) time boundaries in questions can affect 'telescoping' as events which took place before the requested period can be remembered as being within the period. This is not possible in the other direction, which would mean reporting events which will take place in the future. Point (3) is not of importance here as the question does not refer to a specific time period (as for example the last five years) so that boundary effects cannot occur. The same holds for point (1) as retirement is a much more important event than the events typically used in these studies (e.g., participation in talks, watching a movie) so that remembering whether the event occured or not seems not to be a problem. The effect of point (2) can be confirmed (see Tabel 18) when analyzing the absolute error. Whether the time-lag also influences the direction of the error, will be analyzed in the following. Figure 19 shows the distribution of the total error, which can be positive or negative. If telescoping were to occur, the distribution would be negatively skewed, which cannot be confirmed by Figure 19.

As the variable now also takes negative values into account, count models as used for the absolute error are not longer sufficient. The huge number of zeros also argues against a linear regression. I decided to use a multinomial logit model to simultaneously estimate binary logits among the three alternatives: (1) making a negative error, (2) making no error, or (3) making a positive error. In a multinomial logit regression model with an outcome of J categories, $J-1$ binary logit regressions will be estimated. There is always one base category (in Stata, by default, this is the category with the most frequent outcome) to which the other categories are compared to. Here the base category is (2) making no error. Table 19 shows the results for the multinomial logistic regression, which includes

[59] the term 'telescoping' is inspired by looking at something through a telescope which shrinks the real distance to the object (Rubin and Baddeley, 1989)

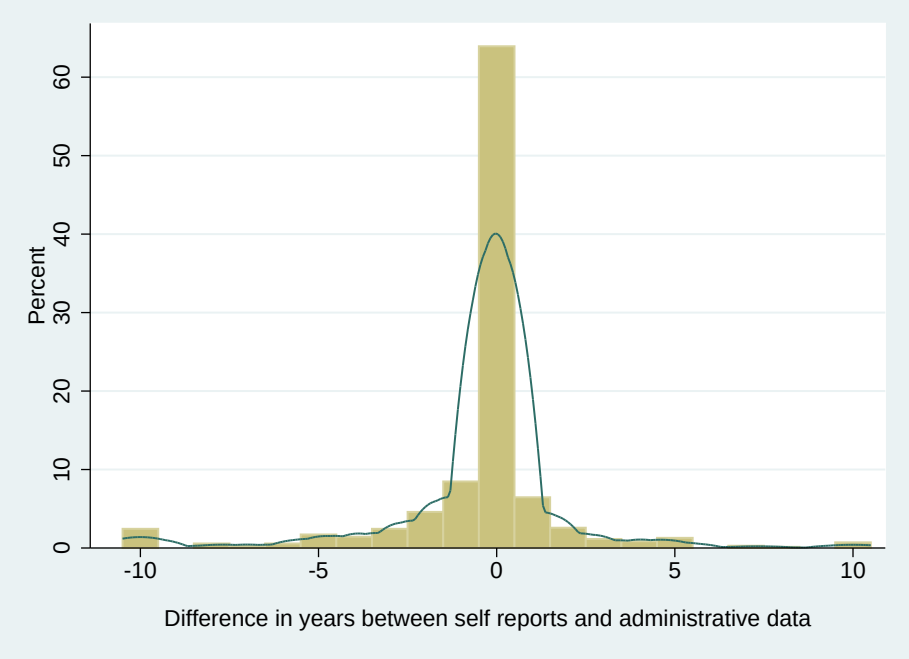


Figure 19: Distribution of Error

the same variables as discussed above for the two comparisons: (1) negative error vs. no error and (2) positive error vs. no error. If the effects of the independent variables are symmetric (whereby I mean that the effect of all independent variables are comparable in effect, size, and significance, for both comparisons) there is no systematic error and the model with the absolute error seems to be sufficient. In contrast, a systematic error would result in coefficients which significantly differ between the two categories. For example, if men have a significantly lower chance of making a negative error and simultaneously a significantly higher chance of making a positive error, this would mean that men (in comparison to women) rather report the event earlier than it took place. At first glance, those variables which show a significant effect in both comparisons are symmetric, suggesting that there are no significant differences between the coefficients of model (1) and model (2). Instead of comparing each pair of effects separately, I used Stata's postestimation command '*mlogtest*' which provides different tests (see Freese and Long, 2000). The adequate test for my question (are there differences between two sets of coefficients) is the 'test for combining alternatives'. If there are no differences, the two categories (negative and positive error) are indistinguishable. The hypothesis which is being tested can be written as: $H_0 : (\beta_{1,-1|0} - \beta_{1,1|0}) = ... = (\beta_{K,-1|0} - \beta_{K,1|0}) = 0$. With the command '*mlogtest, combine*' a Wald test for combining alternatives is calculated[60].

[60]It is also possible to compute an LR test but since the results of the Wald and the LR test provide

Table 20 shows the results test: the hypothesis that categories -1 and 1 (making a negative and making a positive error) are distinguishable cannot be rejected. In contrast, I can reject the hypothesis that categories -1 and 0 as well as categories 1 and 0 are distinguishable. As the results are very similar for both categories, there seems to be no systematic error in a specific direction.

similar results, I decided to use the Wald test as the LR cannot be calculated while using robust standard errors.

Table 19: Multinomial Logistic Regression with Three Categories

	(1)	(2)
	-1 (negative error)	1 (positive error)
Male	-0.24	-0.70
	(0.44)	(0.50)
Cognitive functions	-0.16	-0.26**
	(0.11)	(0.12)
Time-lag (years)	0.06***	0.07***
	(0.02)	(0.02)
Not working	1.27***	0.03
	(0.37)	(0.43)
Several spells	2.02**	1.40*
	(0.90)	(0.79)
Contribution months	0.23	-0.23
	(0.19)	(0.23)
Typical ret. age	-1.29***	-1.15***
	(0.33)	(0.43)
Turn of year	0.54**	0.01
	(0.21)	(0.28)
Interactions		
Male*not work.	-0.26	0.86
	(0.45)	(0.54)
Male*several	-1.33	-1.49
	(1.01)	(1.05)
Male*months	-0.60**	0.12
	(0.24)	(0.30)
Male*typical	1.32***	1.11**
	(0.43)	(0.54)
Missing data	1.14**	-1.01
	(0.54)	(1.17)
Constant	-1.99***	-1.91***
	(0.35)	(0.42)

*, **, *** mark significance on the 10, 5, 1 percent level, respectively

Base category= no error; Robust standard errors in parentheses

Table 20: Wald Test for Combining Alternatives

Alternatives tested	chi2	df	P>chi2
-1 vs. 1	18.803	13	0.129
-1 vs. 0	77.278	13	0.000
1 vs. 0	40.882	13	0.000

6.6 Summary and Discussion

The goal of this chapter was to learn more about recall error when asking respondents about the year they retired. The availability of external validation data allows me to identify the error by comparing the self-reports with the 'true values'. The results can be summarized as follows:

First, the majority of respondents (63.5%) report the year correctly.

Second, motivated by the 'Model of the Response Process' of Tourangeau et al. (2000), different determinants could be defined which influence the error. My first model deals with absolute error, meaning that the direction of the error is not considered. The model consists of two separate steps: first, a binary regression comparing the two outcomes of making no error with making an error. The second step deals with the size of the error conditional on making an error. Most of the variables show a significant effect on the first step. Even if the coefficient of gender is not significant in any of the two steps, respondents' gender matters with regard to the effects of the employment history, as the significant interaction terms show. Better cognitive abilities decrease the likelihood of making an error at all, but show no significant effect with regard to the size of the error. That's different for the the variable time-lag, which is the number of years between the event and the survey. More years in between the two events increase both, the likelihood of a misreport and the size of the error. The coefficients related to the respondent's work history differ by gender and in which of the two steps they show a significant effect. Respondents who didn't work before they retired have a higher chance of misreporting the year, but the effect is not significant when considering the size of the error. This effect is not significantly different for male and female respondents. Male and female respondents who have several retirement spells have a higher chance of making an error and also the size of the error is larger. The number of full contribution months only has an effect for male respondents, and is also significant in both steps. The positive effect of the variable typical retirement for female respondents as well as the effect of retiring close to the turn of the year are significant on the first step only, not on the size of the error.

Third, the error respondents make seems not to be systematic, meaning that other variables determine whether respondents report the event too early or too late. The results of the multinomial logistic regression and the subsequent test show that the coefficients

do not significantly differ between the two outcomes making a positive or making a negative error. In other words, it seems to be the error's variance which differs between subgroups of respondents, not the direction of the error. I used that vague wording as one has to take into account the low number of cases and the fact that the $p-$value is close to the 10% significance level.

One question which has not yet been considered here is about the consequences in terms of biased estimators when using a variable which is measured with error. It is not possible to formulate universal consequences, as they depend on various aspects. For example, one has to differentiate whether the variable measured with error is used as a dependent or an independent variable. In addition, the characteristics of the error are important (such as, distribution, variance, dependencies) as well as the analytical model which is used (for an overview of the consequences of measurement error see: Bound et al., 2001). Different hypothetical scenarios will be discussed in the following. The examples given refer an error structure in which the error in uncorrelated with other variables.

The first example refers to a linear model in which the age of retirement (which is calculated as the difference between the year of birth and the year of retirement) is used as an explanatory variable. According to the variance of the error, the estimated parameters are downward biased (attenuated) and inconsistent. This would mean that the coefficient of the age of retirement could be much smaller or even completely hidden compared to a model in which the age of retirement is measured without error. If other variables correlate with the miss-measured variable (as for example gender), the attenuation bias can even be accentuated when adding these variables to the model (Bound et al., 2001).

The second example also refers to a linear regression but assuming that the age at retirement is used as a dependent variable. In this case the estimates are consistent and unbiased, but they are less efficient. The effect of x could then be interpreted as not statistically significant even if it could be highly significant in the model without measurement error.

The third example is a more specific one, referring to an alternative regression model which is often used to analyze durations in time: the event history analysis. This type of modelling is used to analyze the time between two events (an initial event, e.g., the beginning of one's first job, and a terminal event, such as retirement) and how that time depends on different covariates (Holt et al., 2004). In a huge simulation study, Holt et al. (2004) considered the effect of measurement error on the duration in a state by varying the variance of the error. They compared the estimates of different scenarios with the one of an error free duration. The results of that simulation study show that unlike ordinary regression models, measurement error in the dependent variable can lead to biased estimators when using an event history analysis. As one would expect, the bias

is more severe when the variance is higher and the highest difference was shown if the variance is related to an independent variable of the model.

These examples show that it is hardly possible to formulate the consequences of measurement error in general. What the consequences are can differ from study to study even if the same variable is used. Therefore it is important to better understand the structure of the error.

The knowledge of the structure of the error allows correcting for it when using that variable in regressions. One simulation based method to correct for the bias which is introduced by measurement error with a known error variance is the SIMEX method (Simulation and extrapolation method) by Cook and Stefanski (1994). It uses the relation of the variance of the measurement error to the bias of the estimator when ignoring the measurement error. This is done by adding a simulated additional error with different variances to estimate the effect of the error on the estimated coefficient. The next step extrapolates the function back to the case without measurement error (Küchenhoff et al., 2006). This method is of special interest for complex models and error prone explanatory variables. But the results of the simulation study by Holt et al. (2004) show that in the case of an event history analysis, an error in independent variables can also lead to biased estimators. Therefore, the SIMEX method is also very helpful for error prone dependent variables. In addition, the most recent version of SIMEX also allows modelling heteroskedastic measurement error.

There are also some limitations to the present research. The number of cases available for the analysis is low, which has different consequences. First, it requires summarizing the different statuses of the last employment spell. Especially the comparison of the significance of effects between male and female respondents could be problematic, as some combinations of variables do not occur very often. When studying the direction of the error, I used a multinomial logistic regression to differentiate between negative and positive deviations. Given that all the errors of one direction are summarized into one category, there is a loss of information about the number of years the respondent's report differs from the true value. An adequate way to model the error structure would be a count model, as used in Chapter 6.5, which also considers negative outcomes. As a first step, one could split the count model into negative and positive errors to then compare the estimators. But given that only 36% of the respondents make an error (which correspond to 243 respondents) the results would hardly be valid when splitting them into 20 categories (-10 to + 10). But with the beginning of next year, a new data release of the linked dataset SHARE-RV will be available which includes many more cases. I'll then repeat the analyses presented here to have a deeper look into the direction of the error.

The reduction in the number of cases was based on different reasons, such as the avail-

ability of the data and the respondent's willingness to give their consent to link their survey answers with their administrative records. Both aspects can influence the external validity of the results, as one cannot rule out that the sample used here is selective. The availability of the data limits the results to people who have the obligation to contribute to social insurance (*sozialversicherungspflichtig*), while respondents who are civil servants or self-employed for nearly their whole employment history are not included in the dataset of the German Pension Fund. In addition, some records are not available for different reasons. Unfortunately we do not receive the information why some records are not available at the point in time the data is requested. The respondents' willingness to consent to the data linkage is the main factor which decreased the number of cases. Most of the respondents had been asked for consent in the third wave of data collection, where the consent rate was rather low. But as the results of Chapter 3 show, the characteristics of the interviewer are more influential than the characteristics of the respondent with regard to the likelihood of consenting. Therefore, I assume that this sub-population of SHARE (the consenting respondents) does not significantly differ in whether they remember the year of retirement correctly or not.

Even if it is not clear whether these results can be generalized to all respondents, this analysis is a first step in learning more about recall error in surveys using the example of a variable which is asked in a lot of different surveys.

7 Conclusions

The linkage of survey data with additional data sources such as administrative records or biomarkers has great potential to increase the quality of surveys. In both examples discussed here (record linkage and biomarker project), the linked data is more objective and detailed than the survey data alone. Besides the versatile substantive research questions which can be analyzed with help of these data, some important methodological aspects can also be considered.

The combination of the records of the German Pension Fund with the SHARE survey data allows us to learn a lot about the response process and behavior of the respondents. The results of Chapter 6 are just one example in which external validation data can help to evaluate the data quality of one specific variable. But there is of course room for additional validation studies. The administrative records include other information that allow validations of respondents' incomes or other aspects of their employment history. A comparison of the results of different validations would allow us to answer the question of whether general patterns of the relationship of respondent characteristics and response error can be detected or whether the error structure is different for different questions. The combination of survey data and the biomarkers can help to answer interesting methodological questions like how the subjectively and the objectively measured health statuses are related or whether the subjective or the objective health status is a better predictor for panel attrition.

But before we can fully tap into the potential of the linked data, we have to go one step back: to the 'linkability' of the survey data with the external data sources. The fewer cases that can be linked, the smaller the sample for analysis. This reduces the statistical power and limits the research questions that can be answered. An even more important aspect is the issue of the representativeness of the linked sample. Systematic differences between respondents whose data can be linked and those whose data cannot be linked would lead to bias. Different processes and actors influence the size of the linked sample. For example: (1) whether the data is available for all respondents (the administrative data are only available for people who have been subject to social insurance contributions), (2) whether the drops of blood allow for separating of the blood parameters, (3) whether all materials reached the correct address, and (4) whether the identification number to link the different datasets is unique and available for all cases. These are all examples of aspects and challenges which can reduce the final sample size. But the most important one for both projects is the willingness of the respondents to participate, which can be quantified as the consent rate.

At this point it is important to differentiate between the two players: the respondent and the interviewer. On the respondent level, different characteristics are correlated with the

willingness to consent to both projects. The explanatory variables on the respondent level differ between the two models as the topic and the focus of the two chapters are different. However, one consistent and intuitive finding over both models is that respondents differ in how willing they are in general to provide personal information in a survey. This conclusion is drawn from the result that in both models, respondents who refuse to provide information about their financial details are also more likely to refuse the record linkage and the biomarker projects. Knowing about the characteristics of the respondents which influence their willingness to participate is very helpful to learn more about potential bias. This information can also be used to calculate weights which can correct for differences between consenting and refusing respondents.

As the results of Chapters 3 and 5 show, the interviewers who ask for consent also have an important influence on respondents' willingness to consent. The interviewers are of special interest since, different from the respondents, they are under the researchers' control. This means that through interviewer training and interviewer selection their characteristics can be influenced. Therefore, learning more about the effect of the interviewers is an important step towards increasing the consent rate. When I first tried to look into interviewer effects on consent (Chapter 3) the information about interviewers was limited to some standard demographics and some information generated out of SHARE's paradata. But even here the share of unexplained variance could be reduced from 55% in the empty model to 35% in the final model. An even more impressive reduction of unexplained variance can be found in Chapter 5 which refers to the consent to the collection of dried blood spots. In this example, detailed information about the interviewers collected in the interviewer survey are available. I was especially interested in the influence of interviewers' expectations and experiences as these are two characteristics which can be manipulated by training and selection. These two examples demonstrate how important interviewers are for the success of a survey.

A lot of work and energy was (and will be) invested in innovating social surveys. These new ideas are important to improving research and increasing the quality of survey data. But from my point of view, much more attention needs to be paid to the interviewers and how these innovations change their job. Technological changes have also reached the survey world and surveys are getting more and more complex, requiring more and more specialized skills of the interviewers. This change has to be considered when training interviewers.

There is a growing strand of research dealing with interviewer effects on nearly all aspects of a survey. But as in the example of this dissertation, most of the research is very specific, focusing on one aspect of a survey only. From my point of view a next important step would be to combine these results and define skills and characteristics which are relevant to be a good and successful interviewer with regard to the entire

survey process. Collecting information on the interviewers is an important step within this process. Therefore the expansion of the interviewer survey, which was developed and implemented as part of this dissertation, to other SHARE countries and more waves is a promising step to learn more about interviewer effects in SHARE.

To summarize the results of this dissertation, two points are important: first linking survey data with additional data is a promising innovation in the world of surveys, and second, when implementing innovations in interviewer mediated surveys, the consideration of interviewers and how the new task influences their work tasks is very important for the success of the study.

8 Bibliography

Anderson, D. A. and M. Aitkin (1985). Variance Component Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society. Series B (Methodological) 47*(2), 203–210.

Antoni, M. (2011). Linking Survey Data with Administrative Employment Data: The Case of the ALWA Survey. Working paper. Accessed July 10th, 2012 at: `http://www.norc.org/PDFs/October%202011%20Utilizing%20Administrative%20Data%20Conference/4.%20Antoni%20LinkageOctober2011.pdf`.

Antoni, M. and S. Seth (2012). ALWA-ADIAB - Linked Individual Survey and Administrative Data for Substantive and Methodological Research. *Schmollers Jahrbuch 132 (1)*, 141–146. Nuremberg: Institut für Arbeitsmarkt- und Berufsforschung.

Auriat, N. (1993). "My Wife Knows Best" A Comparison of Event Dating Accuracy Between the Wife, the Husband, the Couple, and the Belgium Population Register. *Public Opinion Quarterly 57*(2), 165–190.

Bailar, B. A. (1983). Interpenetrating Subsamples. In N. L. Johnson and S. Kotz (Eds.), *Encyclopedia of Statistical Sciences*, pp. 197–201. New York: John Wiley & Sons, Inc.

Bar, H. Y. and D. R. Lillard (2012). Accounting for Heaping in Retrospectively Reported Event Data a Mixture-Model Approach. *Statistics in Medicine 31*(27), 3347–3365.

Beste, J. (2011). Selektivitätsprozesse bei der Verknüpfung von Befragungs- mit Prozessdaten. Record Linkage mit Daten des Panels „Arbeitsmarkt und soziale Sicherung" und administrativen Daten der Bundesagentur für Arbeit. *FDZ-Methodenreport 09/2011*, 1–28. Nuremberg: Institut für Arbeitsmarkt- und Berufsforschung.

Biemer, P. P., R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (Eds.) (2004). *Measurement Error in Surveys*. Hoboken, New Jersy: Wiley Series in Probability and Statistics.

Bingley, P. and A. Martinello (2014). Measurement Error in the Survey of Health, Ageing and Retirement in Europe: A Validation Study with Administrative Data for Education Level, Income and Employment. *SHARE Workingpaper Series 16-2014*, 1–33.

Blom, A. G., E. D. de Leeuw, and J. J. Hox (2011). Interviewer Effects on Nonresponse in the European Social Survey. *Journal of the Royal Statistical Society: Series B 47*(2), 203–210.

Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology 43*(1), 1–10.

Börsch-Supan, A. and M. Schuth (2013). Early Retirement, Mental Health and Social Networks. In A. Börsch-Supan, M. Brandt, H. Litwin, and G. Weber (Eds.), *Active Ageing and Solidarity Between Generations in Europe, First Results from SHARE After the Economic Crisis*, pp. 337–348. Berlin: De Gruyter.

Bound, J., C. Brown, G. J. Duncan, and W. L. Rodgers (1994). Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data. *Journal of Labor Economics 12*(3), 345–368.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement Error in Survey Data. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5 of *Handbook of Econometrics*, pp. 3705 – 3843. Elsevier.

Broome, J. (2014). How Telephone InterviewersŠ Responsiveness Impacts Their Success. *Field Methods*.

Calderwood, L. and C. Lessof (2009). Enhancing Longitudinal Surveys by Linking to Administrative Data. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys*, pp. 55–72. Chichester, UK: John Wiley & Sons.

Cameron, A. C. and P. K. Trivedi (1986). Econometric Models Based on Count Data. Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics 1*(1), 29–53.

Campanelli, P., P. Sturgis, and S. Purdon (1997). *Can You Hear Me Knocking? An Investigation into the Impact of Interviewers on Survey Response Rates*. London: Social and Community Planning Research.

Cannell, C. F., P. V. Miller, and L. Oksenberg (1981). Research on Interviewing Techniques. *Sociological Methodology 12*, 389–437.

Conrad, F. G., J. S. Broome, J. Benkŧiŋ, F. Kreuter, R. M. Groves, D. Vannette, and C. McClain (2013). Interviewer Speech and the Success of Survey Invitations. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 191–210.

Cook, J. R. and L. A. Stefanski (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association 89 (428)*, 1314–1328.

Couper, M. P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods 7 (3)*, 145–156.

Couper, M. P. and R. M. Groves (1992). The Role of the Interviewer in Survey Participation. *Survey Methodology 18*(2), 263–277.

Couper, M. P. and L. Lyberg (2005). The Use of Paradata in Survey Research [CD-ROM]. Proceedings of the 55th Session of the International Statistical Institute.

Czaplicki, C. and J. Korbmacher (2010). SHARE-RV: Verknüpfung von Befragungsdaten des Survey of Health, Ageing and Retirement in Europe mit administrativen Daten der Rentenversicherung. *DRV-Schriften Band 55/2010*, 28–37.

Dal Bianco, C., C. Garrouste, and O. Paccagnella (2013). Early-life Circumstances and Cognitive Functioning Dynamics in Later Life. In A. Börsch-Supan, M. Brandt, H. Litwin, and G. Weber (Eds.), *Active Ageing and Solidarity Between Generations in Europe, First Results from SHARE After the Economic Crisis*, pp. 209–223. Berlin: De Gruyter.

De Heer, W. (1999). International Response Trends: Results of an International Survey. *Journal of Official Statistics 15*(2), 129–142.

De Leeuw, E. D. and J. Hox (2009). International Interviewer Questionnaire (IQUEST): Development and Scale Properties. Working paper, Utrecht, The Netherlands: Department of Methodology and Statistics, Utrecht University.

Dentingen, E. and M. Clarkberg (2002). Informal Caregiving and Retirement Timing Among Men and Women: Gender and Caregiving Relationships in Late Midlife. *Journal of Family Issues 23*(7), 857–879.

Dex, S. (1995). The Reliability of Recall Data: a Literature Review. *Bulletin de Méthodologie Sociologique 49*(1), 58–89.

D'Orazio, M., M. Di Zio, and M. Scanu (2006). *Statistical Matching: Theory and Practice*. UK: John Wiley & Sons.

Dunn, K. M., K. Jordan, R. J. Lacey, M. Shapley, and C. Jinks (2004). Patterns of Consent in Epidemiologic Research: Evidence from Over 25,000 Responders. *American Journal of Epidemiology 159*(11), 1087–1094.

Durbin, J. and A. Stuart (1951). Differences in Response Rates of Experienced and Inexperienced Interviewers. *Journal of the Royal Statistical Society. Series A (General) 114*(2), 163–206.

Durrant, G. B., R. M. Groves, L. Staetsky, and F. Steele (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly 74*(1), 1–36.

Dwyer, D. S. and O. S. Mitchell (1999). Health Problems as Determinants of Retirement: Are Self-rated Measures Endogenous? *Journal of Health Economics 18*(2), 173 – 193.

Eisenhower, D., N. A. Mathiowetz, and M. David (2004). Recall Error: Sources and Bias Reduction Techniques. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*, pp. 127–144. Hoboken, New Jersey: Wiley Series in Probability and Statistics.

Esser, H. (1991). Die Erklärung systematischer Fehler in Interviews: Befragtenverhalten als "rational choice". In R. Wittenberg (Ed.), *Person - Situation - Institution - Kultur. Günter Büschges zum 65. Geburtstag.*, pp. 59–78. Duncker & Humblot.

Farbmacher, H. (2013). Extensions of Hurdle Models for Overdispersed Count Data. *Health Economics 22*(11), 1398–1404.

Freese, J. and J. S. Long (2000). sg155: Tests for the multinomial logit model. In *Stata Technical Bulletin Reprints*, Volume 10, pp. 247–255.

Gramlich, T., T. Bachteler, B. Schimpl-Neimanns, and R. Schnell (2010). Panelerhebungen der amtlichen Statistik als Datenquellen für die Wirtschafts- und Sozialwissenschaften. *AStA Wirtschafts- und Sozialstatistisches Archiv 4*(3), 153–183.

Greene, W. (2008). Functional Forms for the Negative Binomial Model for Count Data. *Economics Letters 99*(3), 585 – 590.

Groen, J. A. (2012). Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics 28*(2), 173–198.

Groves, R. M. (1989). *Survey Error and Survey Cost.* Hoboken, New Jersy: Wiley Series in Survey Methodology.

Groves, R. M. and M. P. Couper (1998). *Nonresponse in Household Interview Surveys.* New York: Wiley.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology.* Hoboken, New Jersey: Wiley Series in Survey Methodology.

Groves, R. M. and L. J. Magilavy (1986). Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. *The Public Opinion Quarterly 50*(2), 251–266.

Gustman, A. L. and T. L. Steinmeier (2004). What People Don't Know About Their Pensions and Social Security: An Analysis Using Linked Data from the Health and Retirement Study. In J. B. S. William G. Gale and M. J. Warshawsky (Eds.), *Private Pensions and Public Policies.*, pp. 57–125. Washington: D.C.: Brookings Institution.

Haider, S. and G. Solon (2000). Nonrandom Selection in the HRS Social Security Earnings Sample. RAND Labor and Population Program Working Paper Series 00-01, DRU-2254-NIA, Santa Monica, CA: RAND Corporation.

Hank, K., H. Jürges, and B. Schaan (2009). Die Erhebung biometrischer Daten im Survey of Health, Ageing and Retirement in Europe. *Methoden Daten Analysen: Zeitschrift für Empirische Sozialforschung 3(1)*, 97–108.

Hank, K. and J. M. Korbmacher (2011). Reproductive History and Retirement: Gender Differences and Variations Across Welfare States. In A. Börsch-Supan, M. Brandt, K. Hank, and M. Schröder (Eds.), *The Individual and the Welfare State*, pp. 161–167. Berlin Heidelberg: Springer.

Hank, K. and J. M. Korbmacher (2013). Parenthood and Retirement. *European Societies 15*(3), 446–461.

Hartmann, J. and G. Krug (2009). Verknüpfung von personenbezogenen Prozess- und Befragungsdaten - Selektivität durch fehlende Zustimmung der Befragten? *Zeitschrift für Arbeitsmarkt Forschung 42*, 121–139.

Holt, D., J. McDonald, and C. Skinner (2004). The Effect of Measurement Error on Event History Analysis. In P. P. Biemer, R. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*, pp. 665–685. Hoboken, New Jersey: Wiley Series in Probability and Statistics.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (2 ed.). New York: Routledge Academic.

Hox, J. J. and E. D. de Leeuw (2002). The Influence of Interviewers' Attitude and Behaviour on Household Survey Nonresponse: An International Comparison. In R. Groves, D. A. Dillman, J. L. Eltinge, and R. J. Little (Eds.), *Survey Nonresponse*, pp. 103–118. New York: Wiley.

Huang, N., S.-F. Shih, H.-Y. Chang, and Y.-J. Chou (2007). Record Linkage Research and Informed Consent: Who Consents? *BMC Health Services Research 7*(1), 18.

Huttenlocher, J., L. Hedges, and V. Prohaska (1988). Hierarchical Organization in Ordered Domains: Estimating the Dates of Events. *Psychological Review 95(4)*, 471–484.

Jäckle, A., P. Lynn, J. Sinibaldi, and S. Tipping (2013). The Effect of Interviewer Experience, Attitudes, Personality and Skills on Respondent Co-operation with Face-to-Face Surveys. *Survey Research Methods 7*(1), 1–15.

Jäckle, A., E. Sala, S. P. Jenkins, and P. Lynn (2004). Validation of Survey Data on Income and Employment: The ISMIE Experience. Working Papers of the Institute for Social and Economic Research, 2004-14, Colchester: University of Essex.

Japec, L. (2007). Interviewer Error and Interviewer Burden. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. Link, and R. L. Sangster (Eds.), *Advances in Telephone Survey Methodology*, pp. 187–211. Hoboken, New Jersey: Wiley.

Jaszczak, A., K. Lundeen, and S. Smith (2009). Using Nonmedically Trained Interviewers to Collect Biomeasures in a National In-home Survey. *Field Methods 21*(1), 26–48.

Jenkins, S. P., L. Cappellari, P. Lynn, A. Jäckle, and E. Sala (2006). Patterns of Consent: Evidence from a General Household Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(4), 701–722.

Kho, M. E., M. Duffett, D. J. Willison, D. J. Cook, and M. C. Brouwers (2009). Written Informed Consent and Selection Bias in Observational Studies Using Medical Records: Systematic Review. *British Medical Journal 338:b866*, 1–8.

Klevmarken, A., P. Hesselius, and B. Swensson (2005). The SHARE Sampling Procedures and Calibrated Designs Weights. In A. Börsch-Supan (Ed.), *The Survey of Health, Aging, and Retirement in Europe - Methodology*, pp. 28–69. Mannheim: MEA.

Kneip, T. (2013). Survey Participation in the Fourth Wave of SHARE. In F. Malter and A. Börsch-Supan (Eds.), *SHARE Wave 4: Innovations & Methodology*, pp. 140–155. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Koch, A., A. Blom, I. Stoop, and J. Kappelhof (2009). Data Collection Quality Assurance in Cross-National Surveys : The Example of the ESS. *Methoden Daten Analysen: Zeitschrift für Empirische Sozialforschung 3*(2), 219–247.

Korbmacher, J. and C. Czaplicki (2013). Linking SHARE Survey Data with Administrative Records: First Experiences from SHARE-Germany. In F. Malter and A. Börsch-Supan (Eds.), *SHARE Wave 4: Innovations & Methodology*, pp. 47–53. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Korbmacher, J. and C. Czaplicki (2014). SHARE-RV User Guide. Technical report, Max-Planck-Institute for Sozial Law and Social Policy. `http://share-dev.mpisoc.mpg.de/fileadmin/pdf_documentation/SHARE-RV/SHARE-RV_User_Guide_2_0_0.pdf`.

Korbmacher, J. M. and M. Schröder (2013). Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer. *Survey Research Methods 7(2)*, 115–131.

Kreuter, F. (2013). Improving Surveys with Paradata: Introduction. In F. Kreuter (Ed.), *Improving Surveys with Paradata*. Hoboken, New Jersey: Wiley Series in Survey Methodology.

Kreuter, F., G. Müller, and M. Trappmann (2010). Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opinion Quarterly 74*(5), 880–906.

Kreuter, F., J. Sakshaug, and M. Trappmann (2014). The 2010 PASS Interviewer Survey. Collecting Data for Research into Interviewer Effects. FDZ-Methodenreport 2, Research Data Centre of the German Federal Emplayment Agency at the Institute for Employment Research, Nuremberg.

Küchenhoff, H., S. M. Mwalili, and E. Lesaffre (2006). A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX. *Biometrics 62*(1), 85–96.

Lamla, B. and M. Coppola (2013). Please Sign Here: Asking for Consent in Self-Administered Surveys. Talk at the 5th Conference of the European Survey Research Association, July 2013 Slovenia.

Lang, I. A., N. E. Rice, R. B. Wallace, J. M. Guralnik, and D. Melzer (2007). Smoking Cessation and Transition Into Retirement: Analyses from the English Longitudinal Study of Ageing. *Age and Ageing 36*(6), 638–643.

Lehtonen, R. (1996). Interviewer Attitudes and Unit Nonresponse in Two Different Interview Schemes. International Perspectives on Nonresponse. In S. Laaksonen (Ed.), *Proceedings of the Sixth International Workshop on Household Survey Nonresponse*. Helsinki: Statistics Finland.

Lessof, C. (2009). Ethical Issues in Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys*, pp. 35–54. UK: John Wiley & Sons.

Lessof, C., J. Banks, R. Taylor, K. Cox, and D. Philo (2004). Linking Survey and Administrative Data in the English Longitudinal Study of Ageing. Talk presented at ESRC Research Methods Programme Seminar: Linking Survey Response and Administrative Records, London. Retrieved December, 12, 2011, from `http://www.ccsr.ac.uk/methods/events/linkage/Lessof.pdf`.

Lipps, O. and A. Pollien (2011). Effects of Interviewer Experience on Components of Nonresponse in the European Social Survey. *Field Methods 23*(2), 156–172.

Loeys, T., B. Moerkerke, O. De Smet, and A. Buysse (2012). The Analysis of Zero-inflated Count Data: Beyond Zero-inflated Poisson Regression. *British Journal of Mathematical and Statistical Psychology 65*(1), 163–180.

Long, J. S. and J. Freese (2006). *Regression Models for Categorial Dependent Variabels Using Stata* (2 ed.). Texas: Stata Press.

Mathiowetz, N. A. and G. J. Duncan (1988). Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment. *Journal of Business & Economic Statistics 6*(2), 221–229.

Matschinger, H., S. Bernert, and M. C. Angermeyer (2005). An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview. *Journal of Official Statistics 21*(1), 657–674.

Mazzonna, F. and F. Peracchi (2012). Ageing, Cognitive Abilities and Retirement. *European Economic Review 56*(4), 691 – 710.

McFall, S., C. Booker, J. Burton, and A. Conolly (2012). Implementing the Biosocial Component of Understanding Society- Nurse Collection of Biomeasures. *Understanding Society Working Paper Series 04*.

McFall, S., A. Conolly, and J. Burton (2014). Collecting Biomarkers Using Trained Interviewers. Lessons Learned from a Pilot Study. *Survey Research Methods 08*(1), 57–66.

McKelvey, R. D. and W. Zavoinab (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology 4*(1), 103–120. doi: 10.1080/0022250X.1975.9989847.

Meschi, E., G. Pasini, and M. Padual (2013). Economic Crisis and Pathways to Retirement. In A. Börsch-Supan, M. Brandt, H. Litwin, and G. Weber (Eds.), *Active Ageing and Solidarity Between Generations in Europe, First Results from SHARE After the Economic Crisis*, pp. 101–109. De Gruyter.

Mika, T. and C. Czaplicki (2010). SHARE-RV: Eine Datengrundlage für Analysen zu Alterssicherung, Gesundheit und Familie auf der Basis des Survey of Health, Ageing and Retirement in Europe und der Daten der Deutschen Rentenversicherung. *RVaktuell 12*, 396–401.

Morton-Williams, J. (1993). *Interviewer Approaches*. Aldershot and Brookfield, VT: Dartmouth Publishing.

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics 33*(3), 341 – 365.

Olson, J. (1999). Linkages with Data from Social Security Administrative Records in the Health and Retirement Study. *Social Security Bulletin 62*(2), 73–85.

Olson, K. and A. Peytchev (2007). Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes. *Public Opinion Quarterly 71*(2), 273–286.

O'Muircheartaigh, C., S. Eckman, and S. Smith (2009). Statistical Design and Estimation for the National Social Life, Health, and Aging Project. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 64B*, i12–i19.

Pyy-Martikainen, M. and U. Rendtel (2009). Measurement Errors in Retrospective Reports of Event Histories A Validation Study with Finnish Register Data. *Survey Research Methods 3(3)*, 139–155.

Rabe-Hesketh, S. and A. Skrondal (2008). *Multilevel and Longitudinal Modeling Using Stata* (2 ed.). Texas: Stata Press.

Rasner, A. (2012). *The Distribution of Pension Wealth and the Process of Pension Building - Augmenting Survey Sata with Administrative Pension Records by Statistical Matching*. Dissertation, Technische Universität, Berlin, Germany. Available from `http://opus.kobv.de/tuberlin/volltexte/2012/3384/`.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer.

Rehfeld, U. G. and M. Mika (2006). The Research Data Centre of the German Statutory Pension Insurance (FDZ-RV). *Schmollers Jahrbuch 126*, 121–127.

Rubin, D. and A. Baddeley (1989). Telescoping is not Time Compression: A model. *Memory & Cognition 17*(6), 653–661.

Rust, J. P. (1990). Behavior of Male Workers at the End of the Life Cycle: An Empirical Analysis of States and Controls. In D. A. Wise (Ed.), *Issues in the Economics of Aging*, pp. 317–382. University of Chicago Press.

Sakshaug, J. W. (2013). Using Paradata to Study Response to Within-Survey Requests. In F. Kreuter (Ed.), *Improving Surveys with Paradata. Analytic Uses of Process Information*, pp. 169–186. Hoboken, New Jersey: John Wiley & Sons.

Sakshaug, J. W., M. P. Couper, and M. B. Ofstedal (2010). Characteristics of Physical Measurement Consent in a Population-Based Survey of Older Adults. *Medical Care 48*(1), 64–71.

Sakshaug, J. W., M. P. Couper, M. B. Ofstedal, and D. R. Weir (2012). Linking Survey and Administrative Records: Mechanisms of Consent. *Sociological Methods & Research 41*(4), 535–569.

Sakshaug, J. W. and F. Kreuter (2012). Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data. *Survey Research Methods 6*(2), 113–122.

Sakshaug, J. W., M. B. Ofstedal, H. Guyer, and T. J. Beebe (2014). The Collection of Biospecimens in Health Surveys. In T. P. Johnson (Ed.), *Handbook of Health Survey Methods*. Wiley. forthcoming.

Sakshaug, J. W., V. Tutz, and F. Kreuter (2013). Placement, Wording, and Interviewers: Identifying Correlates of Consent to Link Survey and Administrative Data. *Survey Research Methods 7(2)*, 133–144.

Sala, E., J. Burton, and G. Knies (2012). Correlates of Obtaining Informed Consent to Data Linkage: Respondent, Interview, and Interviewer Characteristics. *Sociological Methods & Research 41*(3), 414–439.

Schaan, B. (2013). Collection of Biomarkers in the Survey of Health, Ageing and Retirement in Europe (SHARE). In F. Malter and A. Börsch-Supan (Eds.), *SHARE Wave 4: Innovations & Methodology*, pp. 38–46. MEA, Max Planck Institute for Social Law and Social Policy.

Schaeffer, N. C., J. Dykema, and D. W. Maynard (2010). Interviewers and Interviewing. In P. V. Mardsen and J. D. Wright (Eds.), *Handbook of Survey Research*, pp. 437–470. Binley, UK: Emerald Group Publishing.

Schnell, R. (2009). Biologische Variablen in Sozialwissenschaftlichen Surveys. *RatSWD Working Paper Series 107*, 1–4.

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

Schröder, M. (2011). *Retrospective Data Collection in the Survey of Health, Ageing and Retirement in Europe. SHARELIFE Methodology.* Mannheim: MEA.

Singer, E., M. R. Frankel, and M. B. Glassman (1983). The Effect of Interviewer Characteristics and Expectations on Response. *Public Opinion Quarterly 47*(1), 84–95.

Singer, E. and L. Kohnke-Aguirre (1979). Interviewer Expectation Effects: A Replication and Extension. *Public Opinion Quarterly 43*(2), 245–260.

Singer, E., J. van Hoewyk, and R. J. Neugebaue (2003). Attitudes and Behavior: The Impact of Privacy and Confidentiality Concerns on Participation in the 2000 Census. *Public Opinion Quarterly 67*(3), 368–384.

Skowronski, J. J. and C. P. Thompson (1990). Reconstructing the Dates of Personal Events: Gender Differences in Accuracy. *Applied Cognitive Psychology 4*(5), 371–381.

Snijders, T. A. and R. J. Bosker (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2 ed.). London: Sage Publishers.

Snijkers, G., J. J. Hox, and E. D. de Leeuw (1999). Interviewers' Tactics for Fighting Survey Nonresponse. *Journal of Official Statistics 15*, 185–198. Reprinted in: D. de Vaus (2002). Social Surveys, Part Eleven, Nonresponse Error. London: Sage, Benchmarks in Social Research Methods Series.

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Prognosen der Rational-Choice Theorie und des Modells der Frame-Selektion. *Zeitschrift für Soziologie 33(4)*, 303–320.

Stocké, V. and C. Hunkler (2007). Measures of Desirability Beliefs and Their Validity as Indicators for Socially Desirable Responding. *Field Methods 19*(3), 313–336.

Sudman, S. (1980). Reducing Response Error in Surveys. *Journal of the Royal Statistical Society. Series D (The Statistician) 29*(4), 237–273.

Sudman, S., N. M. Bradburn, and N. Schwarz (1996). *Thinking About Answers. The Application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass Publishers.

Tisch, L. and S. Tophoven (2011). The lidA Project. Paper presented at the 4$^{\text{th}}$ Workshop of Panel Surveys in Germany, Nuremberg.

Torelli, N. and U. Trivellato (1993). Modelling Inaccuracies in Job-search Duration Data. *Journal of Econometrics 59*(12), 187 – 211.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Trappmann, M., B. Christoph, J. Achatz, C. Wenzig, G. Müller, and D. Gebhardt (2009). Design and Stratification of PASS: a New Panel Study for Research on Long-term Unemployment. IAB Discussion paper 5/2009, Nuremberg: Institut für Arbeitsmarkt- und Berufsforschung.

Tutz, G. and M.-R. Oelker (2014). Modeling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. Technical Report 156, Department of Statistic University of Munich.

van Solinge, H. (2007). Health Change in Retirement: A Longitudinal Study among Older Workers in the Netherlands. *Research on Aging 29*(3), 225–256.

Weir, D. (2008). Elastic powers: The integration of biomarkers into the health and retirement study. In Committee on Advances in Collecting and Utilizing Biological Indicators and Genetic Information in Social Science Surveys. Maxine Weinstein (Ed.), *Biosocial Surveys*, pp. 78–95. Washington, DC: The National Academies Press.

Weiss, L. M. (2013, July). True Blood? Validation Approaches for Dried Blood Spots Collection and Analyses. Talk at the weekly Seminar of the Max Planck Institute for Social Law and Social Policy, Munich.

Woolf, S. H., S. F. Rothemich, R. E. Johnson, and D. W. Marsland (2000). Selections Bias from Requiring Patients to Give Consent to Examine Data for Health Services Research. *Archives of Family Medicine 9*, 1111–1118.

9 Appendix

A Consent to Record Linkage

Table 21: Tabular Literature Overview (replicated from Antoni (2011), with permission of the author)

	Beste (2011)	Gustman and Steinmeier (2004)	Haider and Solon (2000)	Hartmann and Krug (2009))	Jenkins et al. (2006)	Olson (1999)	Sakshaug and Kreuter (2011)	Sala et al. (2012)	Singer et al. (2003)
Respondent									
Male	ns	ns	ns	+	ns	ns		+	ns
Foreign, ethnic minority	-	-	-	-		-	-	-	-
Native language									
Region of residence	ns			sig	sig	ns		sig	ns
Age	ns			ns	sig	ns	+	-	+
Qualification	ns	-	ns	ns	ns	-		+	ns
Cognitive skills									
Labor market status	ns		sig	ns	sig	ns	ns		
Income	+		+	+	ns	+		ns	+
Refused income information	-	-		-	ns			-	
Wealth, assets		-	-			-			
Existing relationship/marriage		+		ns	+	+		ns	
Children		ns			+			ns	
Interviewer									
Male	+			ns				ns	
Age	+			+				ns	
Qualification	-			+				ns	
Experience before study								ns	
Prior interviews within actual study	ns							-	
Similarity of respondent and interviewer									
Sex				ns					
Age				ns					
Qualification	ns								
Interview situation									
Weekday/time of interview									
Share of refused answers									
Share of answers like ”don’t know”									
Duration of interview				ns	+				
Disturbances/problems during interview					-				
Cooperation in other consent questions	+								

Notes: +/-/ns/sig denote significantly positive/significantly negative/no significant/overall significant influence on consent, respectively. “Sakshaug and Kreuter (2011)” refers to an earlier version of Sakshaug and Kreuter (2012).

Table 22: Multilevel Estimation of the Consent Decision: Using an East/West Indicator instead of State Indicators

	with GDR indicator	without GDR indicator
Age	1.340**	1.353**
Age^2	0.998**	0.998**
Female	1.091	1.115
Years of Education	1.012	1.012
Currently employed	0.799	0.817
Number of jobs	1.118*	1.134**
Lives with Partner	1.871**	1.919***
Ever married	1.234	1.195
Ever divorced	0.535**	0.585**
Ever lived in GDR	4.327***	
Household in urban area	0.686	0.661
Household in 1- or 2-family house	1.068	1.058
Foreigner in household	0.725	0.721
Income is missing	0.236***	0.242***
1^{st} income quartile	0.701	0.703
2^{nd} income quartile	0.512*	0.521*
3^{rd} income quartile	0.712	0.731
Living in East	0.153**	0.526
Intra-Class Correlation	0.534	0.538

Notes: *, **, *** mark significance on the 10, 5, 1 percent level, respectively.
Dependent variable in both models is the dichotomous variable "consent to record linkage". Both models are estimated with 1,055 observations in a multilevel logistic regression with Stata's xtlogit command with a random intercept on the interviewer level. Coefficients are odds ratios. χ^2-values are the respective test statistics.
Reference category: Income: 4^{th} income quartile.

Table 23: Multilevel Estimation of the Consent Decision: Testing the Influence of Missing Interviewer Information

	Respondent Characteristics	Interview Situation	Interviewer Characteristics
Age	1261*	1.216	1.211
Age^2	0.998**	0.998	0.999
Female	1.059	1.077	1.093
Years of Education	1.023	1.010	1.007
Currently employed	0.794	0.827	0.820
Number of jobs	1.128**	1.103*	1.111*
Lives with Partner	1.797**	1.642**	1.649***
Ever married	1.147	1.283	1.290
Ever divorced	0.534**	0.569**	0.556**
Ever lived in GDR	4.800***	4.002**	3.917**
Household in urban area	0.637	0.554*	0.518**
Household in 1- or 2-family house	1.008	1.038	1.022
Foreigner in household	0.579	0.659	0.650
Income is missing	0.250***	0.440**	0.438**
1st income quartile	0.635	0.665	0.681
2nd income quartile	0.567*	0.535*	0.553*
3rd income quartile	0.766	0.761	0.767
Interviewer is known		0.877	0.882
Respondent comprehension		1.625**	1.642**
Seconds per question (net Interviewer)		1.006	1.010
Missing rate: financial questions		0.989**	0.989**
Missing rate: non-financial questions		0.751**	0.760**
Interviewer's experience: interview 6-10		0.821	0.829
Interviewer's experience: interview 11-20		0.581*	0.580
Interviewer's experience: interview 21-50		0.354***	0.362***
Interviewer's experience: interview 51+		0.226***	0.225***
Interviewer is male			2.535*
Average seconds per question (Interviewer)			1.047
Quality: too few multiples of 5 and 10			0.186*
Quality: too many multiples of 5 and 10			0.487
Interviewer information missing	1.038	0.693	0.380
ICC	0.467	0.439	0.404
χ^2 (1) of LR-Test for interviewer information	0.003	0.230	1.520
Observations	1,172	1,172	1,172

Notes: *, **, *** mark significance on the 10, 5, 1 percent level, respectively.
Dependent variable in all models is the dichotomous variable "consent to record linkage". All models are estimated with a multilevel logistic regression with Stata's xtlogit command with a random intercept on the interviewer level. All estimations include state fixed effects. The χ^2-values refer to the test statistics from a test of two nested models including the indicator for missing interviewer information.
Reference categories: Income: 4th income quartile; Experience: interview 1-5; Quality: rounding not questionable.

Table 24: Multilevel Estimation of the Consent Decision: Including Previous Wave Information on the Interview Situation

	Basic Model (similar Table 2, Column 4)	Check 1: previous waves	Check 2: previous waves
Interviewer is known	0.866	0.808	0.692
Respondent comprehension	1.848**		
Respondent comprehension (w1/w2)		1.265	0.973
Seconds per question (net Interviewer)	1.007	1.029	1.027
Missing rate: financial questions	0.986**		
Missing rate: non-financial questions	0.768*		
Missing rate: financial questions (w1/w2)		0.977***	0.977***
Missing rate: non-financial questions (w1/w2)		0.893	0.920
Respondent willingness to answer (w1/w2)			2.838***
Interviewer's experience: interview 6-10	0.678	0.765	0.729
Interviewer's experience: interview 11-20	0.712	0.845	0.881
Interviewer's experience: interview 21-50	0.368***	0.447**	0.442**
Interviewer's experience: interview 51+	0.233***	0.278***	0.260***
ICC	0.458	0.512	0.499
Observations	1,046	1,046	1,046

Notes: *, **, *** mark significance on the 10, 5, 1 percent level, respectively.

Dependent variable in all models is the dichotomous variable "consent to record linkage". All models are estimated with 1,046 observations in a multilevel logistic regression with Stata's xtlogit command with a random intercept on the interviewer level. All estimations include state fixed effects and all variables also included in Appendix Table 3. The coefficients represent odds ratios. "w1/w2" refers to data coming from previous waves of SHARE: from wave 2, if they were available there or otherwise from wave 1.

Reference categories: Experience: interview 1-5.

B Interviewer Survey

SHARE “50+ in Europe”

Interviewer-Questionnaire

You as interviewer play a key role in the success of our SHARE study. Therefore, we from [COUNTRY’S INSTITUTION] want to get to know you; your attitudes, your experiences as a successful interviewer and your opinion concerning the interview situation. Your participation is of course voluntary. However, with your participation you help us immensely in better understanding the interview situation. Your answers do **not** serve to an assessment of your performance and will **not** be passed down to [SURVEY ORGANISATION]. [FURTHER INFORMATION ABOUT WHAT HAPPENS TO YOUR DATA YOU WILL FIND IN THE ENCLOSED DATA PROTECTION LEAFLET.]

Please fill in your interviewer-number!

Interviewer-number: _____

Job as an interviewer

1) How long in total have you been working as an interviewer?

☐ years and ☐ months ☐ 99 (don't know)

2) How many hours a week do you currently approximately work as an interviewer?

☐ hours ☐ 99 (don't know)

3) There are different reasons for working as an interviewer. How important are the following aspects to you?

Please provide an answer in each row using the following scale. Value 1 means: not important at all, value 7 means: very important. With the values between 1 and 7 you can grade your opinion.

	1= not important at all ⟷ 7= very important							don't know
Payment	☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐9
Interesting work	☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐9
Opportunity to interact with people	☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐9
Gaining insight into other people's social circumstances	☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐9
Involvement in scientific research	☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐9
Involvement in research that serves society	☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐9
Possibility to determine own working hours	☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐9

4) Below follows a series of statements about difficult respondents and contact attempts. We would like to know from you, how you react in the following situations.

Please provide an answer in each row using the following scale!

The statement applies to me ……	perfectly	some-what	not really	not at all	don't know
If the respondent doesn't understand a question, I explain what is actually meant with the question.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If the respondent has difficulties with a question, I don't help, but read out the exact wording again.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If I notice that the respondent has difficulties listening to me, I shorten long question texts.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If I notice that the respondent has difficulties understanding the question, I speak more slowly.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If I notice that the respondent is in a hurry, I speak faster.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If I know from the course of the interview what an answer will be, I complete the answer myself.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If I remember answers from previous waves and notice that nothing has changed, I complete answers myself.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If I notice that the respondent doesn't speak [FORMAL ENGLISH – COUNTRY EQUIVALENT], I also speak regional dialect.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
I always exactly stick to the interviewer instructions, even if I don't consider them sensible.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9

5) Sample persons have different reactions to the request to participate in a study: Some agree spontaneously, others hesitate or refuse immediately. In the following statements, please tell us your opinion as an experienced interviewer.

Please provide an answer in each row using the following scale!

	strongly agree	some-what agree	some-what disagree	strongly disagree	don't know
Reluctant respondents should always be persuaded to participate.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
With enough effort, even the most reluctant respondent can be persuaded to participate.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
An interviewer should respect the privacy of the respondent.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If a respondent is reluctant, a refusal should be accepted.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
One should always emphasise the voluntary nature of participation.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
It does not make sense to contact reluctant target persons repeatedly.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
If you catch them at the right time, most people will agree to participate.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Respondents that were persuaded after great effort do not provide reliable answers.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9

General attitudes and behaviour

6) Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?

Please use the scale from 0 to 10, where 0 means that you can't be too careful in dealing with people and 10 means that most people can be trusted. With the values in between you can grade your opinion.

You can't be too careful. ⟷ Most people can be trusted.	don't know
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10	☐ 99

7) What would you say? To what extend do the following statements apply to you?

Please provide an answer in each row using the following scale!

The statement applies to me ……	perfectly	some-what	not really	not at all	don't know
My first impression of people generally turns out to be right.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
I am uncertain about my judgements.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
I know exactly why I like something.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
I don't say anything, if I receive too much change.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
I am honest with others.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9

You as a respondent

8) In the last 5 years, how often have you taken part in a survey as a respondent (not counting this survey)?

☐ 999 (don't know)

9) If you have previously taken part in surveys, what kind of surveys were they?

Please tick one answer only!

Predominantly scientific surveys (e.g. studies like SHARE, election studies)	…… ☐ 1
Predominantly commercial surveys or market research	…… ☐ 2
Both scientific and commercial surveys to the same extent	…… ☐ 3
I have never taken part in any survey.	…… ☐ 8

10) Have you received any incentive/compensation for your participation in these studies?

Please tick one answer only!

Predominantly yes	☐ 1	Approximately both to the same extent	☐ 3
Predominantly no	☐ 2	I have never taken part in any survey.	☐ 8

Data protection

11) How concerned are you about the safety of your personal data?

Please tick one answer only!

Very concerned	☐ 1
Quite concerned	☐ 2
A little concerned	☐ 3
Not concerned at all	☐ 4

☐ 9 (don't know)

12) How concerned are you about computers or other technologies being used to invade your privacy?

Please tick one answer only!

Very concerned	☐ 1
Quite concerned	☐ 2
A little concerned	☐ 3
Not concerned at all	☐ 4

☐ 9 (don't know)

What would you do?

In the following we want to ask you to imagine yourself in different hypothetical situations. What would you do if you were in one of the following situations?

13) You are a respondent to a survey of [NATIONAL STATISTICAL OFFICE]. As part of this survey you are asked to provide the following pieces of information. For each of these the interviewer gives you plausible reasons why he/she needs the information.

Please provide an answer in each row using the following scale!

How likely is it that you would provide the following information?	very likely	quite likely	quite unlikely	very unlikely	don't know
Your national social insurance number	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your date of birth	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your place of birth	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your private telephone number	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your complete name	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your mother's maiden name	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your private address	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your credit card number	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Name and address of your health insurance	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your health insurance number	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9

14) In the same study you are asked to consent to the linkage of your survey data with administrative data. How likely is it that you would consent to the [NATIONAL STATISTICAL OFFICE] linking your answers with the following data sources?

Please provide an answer in each row using the following scale!

	very likely	quite likely	quite unlikely	very unlikely	don't know
Your income tax assessment	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your debts and loans	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your employment history, i.e. information about previous periods of employment and unemployment	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Your medical data, held by your doctors	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Information about your health insurance	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Information about receipt of social security benefits such as unemployment benefits or social welfare	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9
Information from your school files (diplomas etc.)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 9

15) In some of their surveys [SURVEY ORGANISATION] asks respondents to consent to have their survey data linked to the administrative data from the [ADMINISTRATIVE DATA SOURCE, E.G. SOCIAL SECURITY REGISTER]. This concerns for example additional information about [PREVIOUS PERIODS OF EMPLOYMENT, UNEMPLOYMENT AND THE PARTICIPATION IN WORK PROGRAMMES DURING UNEMPLOYMENT]. What do you think, how many of your respondents (in percent) would consent to this?

☐ % ☐ 999 (don't know)

16) Would you as a respondent agree to such a linkage?

Yes ..... ☐ 1 No ..... ☐ 2 ☐ 9 (don't know)

Expectations about wave 4 of the study 50+ in Europe

18) Have you worked as an interviewer on previous waves of SHARE?

Yes ..... ☐ 1 No ..... ☐ 2

19) Studies vary as to whether they reward respondents for their survey participation and how much respondents receive. Please imagine that your respondents receive the following incentives.

Please indicate your expectations in each row!

What do you expect, which percentage of your sample persons will agree to the interview, if...	Expected response rate in percent
[NATIONAL SCENARIOS]	%
	%
	%
	%

20) Social surveys very often ask about respondents' income. How many of your respondents (in percent) in SHARE do you expect will provide information about their income?

%

☐ 999 (don't know)

21) In SHARE respondents are asked to consent to some physical measurements, such as blood pressure, height, waist circumference and the collection of small blood spots.

Please give your expectations in each row!

What do you think, which percentage of your respondents will consent to the following measurements?	Expected consent rate in percent
Measurement of blood pressure	%
Measurement of body height	%
Measurement of waist circumference	%
Collection of small blood spots	%

22) Please imagine that you are a respondent to SHARE or a similar scientific study. Which measurements would you as a respondent consent to?

Please tick all that apply!

Measurement of blood pressure	☐ 1
Measurement of body height	☐ 2
Measurement of waist circumference	☐ 3
Collection of small blood spots	☐ 4

☐ 9 (don't know)

23) Do you personally have experience with measuring blood sugar levels, either because you or someone you know has diabetes?

Yes ☐ 1 No ☐ 2

24) Do you donate blood?

Please tick one answer only!

Yes, regularly	☐ 1	No, not anymore	☐ 3
Yes, occasionally	☐ 2	No, I have never donated blood	☐ 4

Personal details

25) Are you male or female?

Male ☐ 1 Female ☐ 2

26) In which year were you born?

Year of birth: ☐☐☐☐

27) Do you use social networks in the internet like Facebook, Myspace or Twitter?

Yes ..... ☐ 1 No ..... ☐ 2

28) Do you use the internet for online-banking?

Yes ..... ☐ 1 No ..... ☐ 2

29) Do you hold the [COUNTRY'S] citizenship?

Yes ..... ☐ 1 No ..... ☐ 2

30) Please state whether you, your mother and your father were born in [COUNTRY].

	Yes	No	
You yourself	☐ 1	☐ 2	☐ 9 (don't know)
Your mother	☐ 1	☐ 2	☐ 9 (don't know)
Your father	☐ 1	☐ 2	☐ 9 (don't know)

31) Apart from your job as an interviewer do you have any other job? Are you...

Please tick all that apply!

full-time employed	..... ☐ 1	retired	..... ☐ 7
part-time employed	..... ☐ 2	on parental leave	..... ☐ 8
[COUNTRY SPECIFIC]	..... ☐ 3	a homemaker	..... ☐ 9
in vocational training or occupational re-training	..... ☐ 4	a student	..... ☐ 10
unemployed	..... ☐ 5	other	..... ☐ 11
[COUNTRY SPECIFIC]	..... ☐ 6	none of these	..... ☐ 12

32) Which is your highest level of education?

Please tick your highest level of education only!

Graduated from lower-level secondary school [NATIONAL EQUIVALENT]	..... ☐ 1
Graduated from medium-level secondary school [NATIONAL EQUIVALENT]	..... ☐ 2
Advanced technical college entrance qualification or graduated from upper-level secondary school [NATIONAL EQUIVALENT]	..... ☐ 3
University degree [NATIONAL EQUIVALENT]	..... ☐ 4

33) How many persons do currently live in your household?

34) All in all, approximately what was the average monthly income of your household after taxes in the last year?

€

Thank you very much for participating!

C Consent to the Collection of Biomarkers

C.1 Full Model also Including Respondent Level Variables

Table 25 displays the full model including all variables at both the respondent and the interviewer levels. The respondents' standard demographics do not show a significant effect on their willingness to consent to the collection of dried blood spots. The respondent's health status shows an ambiguous effect: on the one hand, respondents who are diabetic are more willing to consent than are those who do not have this disease. On the other hand, the more reported limitations in their daily living activities, the lower their likelihood of consenting. These results are in line with the results of Weiss (2013) and are not that surprising, considering that diabetics are used to the technique. Respondents who did not report their income are also less likely to consent to the biomeasure. In comparison to the results of Weiss (2013), who used to entire German Wave 4 sample, only one coefficient differs: the positive effect of having high cholesterol is significant in Weiss (2013) but not significant here. All other variables at the respondent level have the same sign and level of significance even if the sample used by Weiss (2013) is twice the size. The robustness of the results at the respondent level in this sub-sample supports the assumption that the sample used here is not selective.

Table 25: Multilevel Logistic Regression: Full model

		Full Model
Respondent chrarcteristics:		
Male	1.07	(0.19)
Age: <=59	0.89	(0.26)
Age: 60–64	1.12	(0.31)
Age: 65–69	0.84	(0.25)
Age: 70–75	1.20	(0.33)
Low educational level	0.96	(0.29)
Medium educational level	1.05	(0.21)
DDR	0.98	(0.33)
High cholesterol	1.44	(0.33)
Diabetic	1.69*	(0.46)
Difficulties with activities	0.80***	(0.04)
Income missing	0.45***	(0.12)
Living in urban area	0.98	(0.23)
Interviewer chrarcteristics:		
Age	1.03**	(0.02)
Male	0.62	(0.19)
Low educational level	0.16***	(0.10)
Medium educational level	1.30	(0.42)
Member of social networks	1.48	(0.50)
Hypothetical own consent to DBS	1.16	(0.45)
Motivation: "socialize"	0.51*	(0.19)
Motivation: "research"	0.96	(0.34)
Experience in measuring blood sugar	0.83	(0.27)
1–5. interview	0.56***	(0.11)
Years of experience	0.96*	(0.02)
Years of experience^2	0.99***	(0.00)
Expected consent rate	1.04***	(0.01)
Years*Expectations	1.01***	(0.00)
ICC	0.09	
Number of interviewers	55	
Number of cases	843	
χ^2 against logistic regression	5.93 ***	
(degrees of freedom; p-value of LR test)	(27; 0.000)	

Notes: *,**,*** mark significance on the 10, 5, 1 percent level, respectively

Exponentiated coefficients; Standard errors in parentheses

Dependent variable in all models is the dichotomous variable "consent to dbs collection"

All models are estimated in a multilevel logistic regression with Stata's xtlogit command with a random intercept on the interviewer level. Coefficients are odds ratios.

χ^2 are the respective test statistics.

C.2 Excluding Interviewers

The results demonstrate that interviewers' experience seems to be an important determinant in predicting interviewers' success in getting respondents' consent. One limitation of that study is the low number of interviewers. Of those 55 interviewers, two are conspicuous because they differ a lot from all other interviewers in two aspects: they both interviewed a high number of respondents (in sum about 10% of the sample), and have a 0% consent rate (see Fig. 10, page 64). In addition, they are very experienced (with 15 and 40 years of experience). As these two interviewers could be assumed to be very influential, the analyses were repeated by excluding these two interviewers from the sample. Due to the exclusion of these interviewers, the sample is reduced by two level-two units (interviewers) and 78 level-one units (respondents). A comparision of the two models can be found in Table 26. Two main changes can be reported: first, the effect of an interviewer's experience disappears, and second, the curves of the predicted probabilities are much closer and smoother. Fig. 20 displays the predicted probabilities

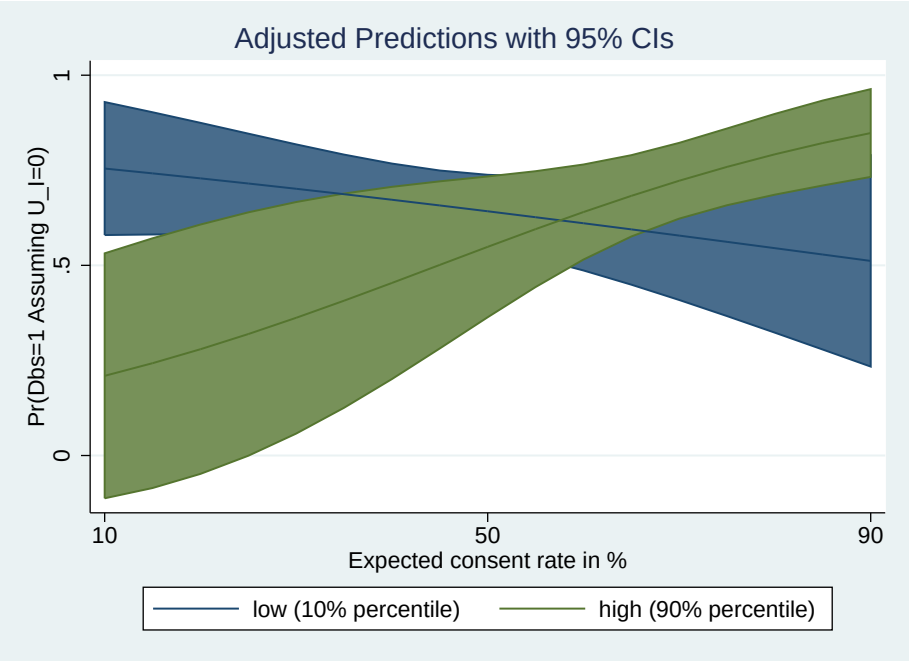


Figure 20: Predicted probabilities of consenting to the collection of dried blood spots: Reduced sample

for the reduced sample of interviewers (excluding two interviewers) for less experienced and for highly experienced interviewers. [61]

[61] The predicted probabilities for interviewers with average experience are not included in the graph as these are very close to the blue curve, which would make it very hard to read the graph.

Table 26: Multilevel Logistic Regression: Comparison of Reduced and Full Sample

	Model 1 Reduced Sample		Model 2 Full Sample	
Respondent characteristics:				
Male	1.08	(0.19)	1.07	(0.19)
Age: <=59	0.94	(0.27)	0.89	(0.26)
Age: 60–64	1.21	(0.33)	1.12	(0.31)
Age: 65–69	0.94	(0.28)	0.84	(0.25)
Age: 70–75	1.23	(0.33)	1.20	(0.33)
Low educational level	0.91	(0.27)	0.96	(0.29)
Medium educational level	1.02	(0.20)	1.05	(0.21)
DDR	1.13	(0.31)	0.98	(0.33)
High cholesterol	1.42	(0.32)	1.44	(0.33)
Diabetic	1.79**	(0.49)	1.69*	(0.46)
Difficulties with activities	0.80***	(0.04)	0.80***	(0.04)
Income missing	0.43***	(0.11)	0.45***	(0.12)
Living in urban area	0.98	(0.20)	0.98	(0.23)
Interviewer characteristics:				
Age	1.04***	(0.01)	1.03**	(0.02)
Male	0.79	(0.21)	0.62	(0.19)
Low educational level	0.29**	(0.15)	0.16***	(0.10)
Medium educational level	1.72**	(0.46)	1.30	(0.42)
Member of social networks	1.50	(0.39)	1.48	(0.50)
Hypothetical own consent to DBS	0.71	(0.24)	1.16	(0.45)
Motivation: “socialize”	0.54**	(0.17)	0.51*	(0.19)
Motivation: “research”	1.04	(0.30)	0.96	(0.34)
Experience in measuring blood sugar	1.12	(0.31)	0.83	(0.27)
1–5. interview	0.53***	(0.10)	0.56***	(0.11)
Years of experience	1.02	(0.02)	0.96*	(0.02)
Years of experience^2	1.00	(0.00)	0.99***	(0.00)
Expected consent rate	1.01	(0.01)	1.04***	(0.01)
Years*Expectations	1.00**	(0.00)	1.01***	(0.00)
ICC	0.02		0.09	
Number of interviewers	53		55	
Number of cases	765		843	

Notes: *,**,*** mark significance on the 10, 5, 1 percent level, respectively

Exponentiated coefficients; Standard errors in parentheses

Dependent variable in all models is the dichotomous variable “consent to dbs collection”

All models are estimated in a multilevel logistic regression with Stata’s xtlogit command with a random intercept on the interviewer level. Coefficients are odds ratios.

D Multilevel Modeling

As the models used in Chapter 3 and 5 are identical, some general aspects which are relevant for both chapters will be discussed here.

D.1 Assumptions of the Model

Even of the random intercept model seems to be a valid way to model the heterogeneity among interviewers which allows to estimate the parameter of interest: the share of variance on the interviewer level there are some drawbacks which are the assumptions of the model. As mentioned in Chapter 3 the random effect is assumed to be normally distributed (Rabe-Hesketh and Skrondal, 2008; Tutz and Oelker, 2014) an assumption which can hardly be tested. Using the 'xtmelogit' allows to predict the random intercepts but they should not be used for model diagnostic within the logistic regression (Rabe-Hesketh and Skrondal, 2008). The assumption of normally distributed random intercepts implicitly assumes that all interviewers differ in their intercept (Tutz and Oelker, 2014). The second assumption refers to the independence of the random effects and the covariates (Tutz and Oelker, 2014). An alternative model which is recommended by Tutz and Oelker (2014) is the fixed effects model as this overcomes the assumptions of the random effects model. But the disadvantage of this model is that one cannot include group-specific explanatory variables. As this is the main interest of this work, fixed effect models cannot be used.

D.2 Scale Correction

In logistic regressions, the scale of the unobserved latent variable is standardized to the same distribution in each model. By adding explanatory variables, one would expect to find smaller variance components in the full model. But in logistic regressions, the latent variable is rescaled so that the lowest level residual variance is again $\frac{\pi^2}{3}$. As a consequence, regression coefficients and variance components cannot be compared across models (see (Hox, 2010)). To correct for that fact, Hox (2010) follows the approach of McKelvey and Zavoinab (1975) by calculating a scale correction factor. The scale correction factor for the variance components is the ratio of the total variance of the null model $\sigma_0^2 = \sigma_{u0}^2 + \frac{\pi^2}{3}$ to the total variance of the model which includes only level-one characteristics, $\sigma_m^2 = \sigma_F^2 + \sigma_{u0}^2 + \frac{\pi^2}{3}$, where σ_F^2 is the variance of the linear predictor from the fixed part of the model. Before calculating the intra class correlation of the final model, all variance components have to be multiplied by this factor. The scale correction factor for the model of Chapter 3 is 0.89. This reduces the ICC of the final model only slightly. For Chapter 5 the scale correction factor is 0.93, this does not change the reported ICC.

E Recall Error in the Year of Retirement

E.1 What do Respondents Report When Asked About Retirement?

All respondents participating for the first time (refreshment sample) are asked very detailed questions about their employment status. In addition to the questions 'ep005'(current job situation) and 'ep329' (the year they retired), question 'ep050' asked when the last job before retirement ended[62]. Question 'ep213' asked about the year they first received a pension, distinguishing between the different income sources [63]. The combination of the three measures allows differentiating between the two concepts: leaving the workforce and entering into retirement. I compared the year they reported in question 'ep329' with the two other questions and summarized the difference in each case into three categories:

- negative difference (ep329 reported to be before leaving the workforce/receiving the first payment)
- no difference
- positive difference (ep329 reported to be after the the workforce/receiving the first payment)

Table 27: Difference of Reported Year of Retirement and the Year Leaving the Workforce/Receiving the First Payment

	Leaving job			
Payment	- difference	No difference	+ difference	Total
- difference	9	64	12	85
No difference	15	337	181	533
+ difference	0	13	21	34
Total	24	414	214	652

[62] **ep049**:"We are now going to talk about the last job you had before you retired."; **ep050**:"In which year did your last job end?"

[63] **ep213**: In which year did you first receive this [public old age pension/public old age supplementary pension or public old age second pension/public early retirement or pre-retirement pension/main public disability insurance pension, or sickness benefits/secondary public disability insurance pension, or sickness benefits/public unemployment benefit or insurance/main public survivor pension from your spouse or partner/secondary public survivor pension from your spouse or partner/public war pension/public long-term care insurance/occupational old age pension from your last job/occupational old age pension from your second job/occupational old age pension from a third job/occupational early retirement pension/occupational disability or invalidity insurance/occupational survivor pension from your spouse or partner's job]?

Table 27 summarizes the differences for the two variables for all respondents of the refreshment sample who were already retired. 337 respondents (52%) reported the same year in all three questions. For 414 respondents (63%) the reported year of retirement (ep329) and the reported year they left the workforce (ep050)are the same (including the 337 cases mentioned above) and 533 (82%)reported the same year for retirement (ep329) and the first receipt of a pension (ep213)(again including the 337 cases mentioned above). When limiting the sample to the refreshment cases which are in the final sample, the distribution looks pretty much the same. These results show that the majority of respondents seem to understand the question as expected: the year they retired is the year they received a pension for the first time. 77 respondents (12%) instead answered the year they left the workforce, all others (42, 6%) answered something completely different. How this affects the error (the difference between the reported year of retirement and the year provided by the German Pension Fund) can only be evaluated for the respondents who could be linked successfully. The share is with 11% (24 respondents) the same as for the whole refreshment sample. All but three of them made an error in reporting the year in retirement in terms of the dependent variable of this chapter.
Unfortunately, the two additional questions which are used here are not available for the panel sample in the same wave, so that it is not possible to add a variable controlling for "reporting the year of leaving workforce" to the model. But a deeper look into the characteristics of these 24 respondents show that 2/3 of them have the status 'not working,' which is controlled for in the model.

E.2 Distribution of Years Respondents Retired Based on the Administrative Data

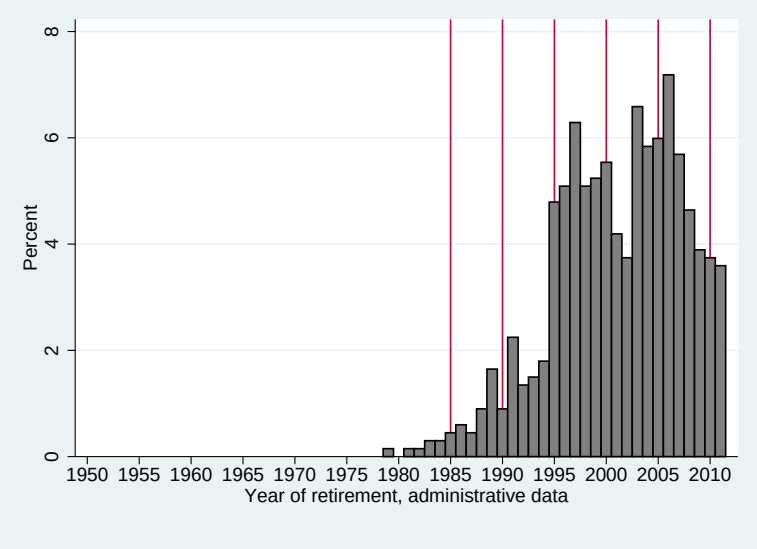


Figure 21: Distribution of Reported Years

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Name, Vorname

....

Ort, Datum Unterschrift Doktorand/in

Formular 3.2