SHARE ERIC
SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE

SHARE WORKING PAPER SERIES

Recall error in the year of retirement

Julie M. Korbmacher

Working Paper Series 21-2014
Version 2 / Nov 29 2016

SHARE-ERIC | Amalienstr. 33 | 80799 Munich | Germany | share-eric.eu

mea | MAX-PLANCK-INSTITUT FÜR SOZIALRECHT UND SOZIALPOLITIK
Munich Center for the Economics of Aging
Horizon 2020
Federal Ministry of Education and Research
DFG
NIH National Institute on Aging

Recall Error in the Year of Retirement

Julie M. Korbmacher

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

Abstract

This project analyzes recall error in the year of retirement by comparing the self-report of respondents of the Survey of Health, Ageing and Retirement in Europe (SHARE) with administrative records of the same person provided by the German Pension Fund. A comparison of the two data sources show that the majority of respondents (63.5%) reported the year they retired correctly. Deviations occurred in 36.5% only, mainly within the range of +/- two years. Based on research from cognitive psychology, different determinants to explain the absolute error are identified. A hurdle regression model is estimated consisting of two steps: first a logistic regression analyzing whether or not respondents make an error and second a zero-truncated negative binomial model to predict the size of the error. The results show that cognitive abilities as well as characteristics of the event are correlated with the absolute error in both steps. In addition, some effects differ for male and female respondents. The results change when considering the direction of the error: no other variables determine whether respondents report the event too early or too late. It is the error's variance which differs between subgroups of respondents, not the direction of the error.

Keywords: Recall error, record linkage, validation

***Acknowledgement**: I would like to thank Johanna Bristle, Helmut Farbmacher, Barbara Felderer, Thorsten Kneip, Frauke Kreuter, Luzia Weiss and Mathias Weiss for valuable comments and suggestions. All remaining errors are my own.*

This paper 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). In addition, 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, as well as the Forschungsnetzwerk Alterssicherung for the funding of the continuation of SHARE-RV.

1 Introduction

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 paper 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; Calder-

wood 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 whether variables from survey and administrative data are measuring the same and whether the variable of interest derived from the administrative data is measured without error.

As more and more surveys have started to link survey data and administrative data, 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 paper 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 (unlikely social desirability and 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) an event which already took place for a large fraction of the SHARE population (50+), and (4) 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 paper is structured as follows: Section 2 describes the validation of the year of retirement based on the comparison of survey and administrative data. Sections 2.1 and 2.2 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 2.3 provides the model and results whereby Section 2.4 closes with some final conclusions.

2 Validating the Year of Retirement Using SHARE-RV

The project SHARE-RV, which combines SHARE survey data with the administrative records of the German Pension Fund (Börsch-Supan et al., 2013; Korbmacher and Czaplicki, 2013), allows analyzing the error respondents make when reporting their year of retirement, as this information is included in both datasets. The data used is based on the German sub-sample of the fourth wave of data collection. This sample consists of respondents of the panel sample (Release 1-1-1) which participated for at least two waves of data collection and respondents of a refreshment sample (unpublished internal data)[1] , 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. The linkage rate, which combines respondents' consent, the availability and the 'linkability' of the administrative records, is 48.5% for the panel sample and 34.3% for the refreshment sample.[2] Within the project SHARE-RV, two different datasets can be combined with the survey data: the VSKT (Versichertenkontenstichprobe) which includes respondents working histories as well as the RTBN (Rentenbestand) which is cross-sectional and includes all information which are used to calculate the pension. The RTBN is only available for respondents who are retried. The linkage rate include both the data of the employment histories as well as pension data; 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 1).

Table 1: 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 precisely: after the SHARE interview but before the end of 2012 (the release version

[1] see Kneip (2013)

[2] 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.

Table 2: 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

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).*[3] Only if the respondents declare that they are retired, are they asked about the year in which they retired.[4] 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 2 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.[5] 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.

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

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

[5] 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.

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.

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

$$dif_{year_{abs}} = |year_{reported} - year_{admin}| \tag{1}$$

$dif_{year_{abs}}$ is the absolute deviation between the report of the respondent ($year_{reported}$) and the value provided by the administrative data ($year_{admin}$). Figure 1 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.
To provide a better impression of the errors' extent, Figure 2 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 1: more than 60% of the respondents report the year correctly.

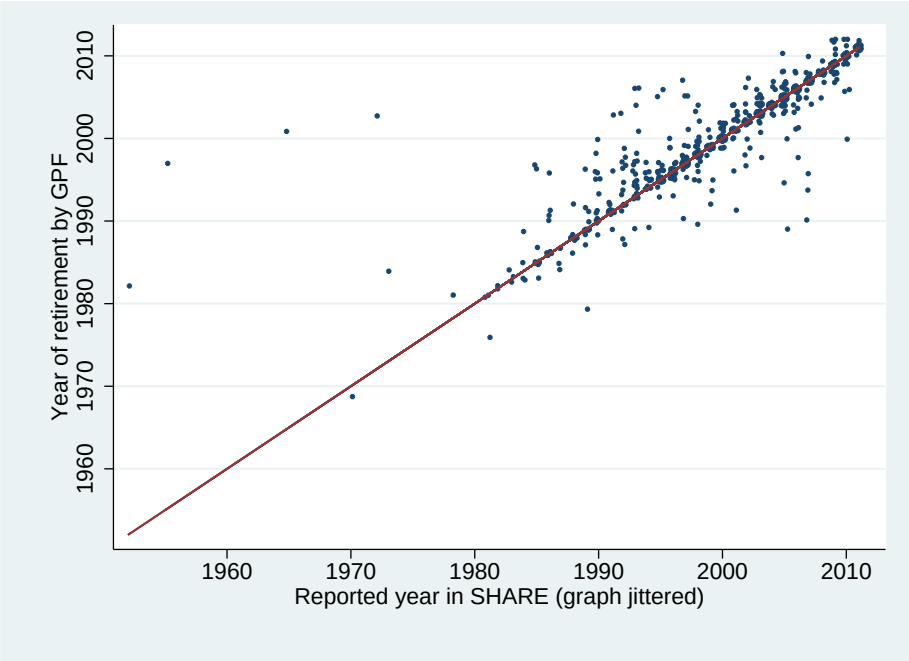

Figure 1: Difference in Reported and True Values of the Year of Retirement

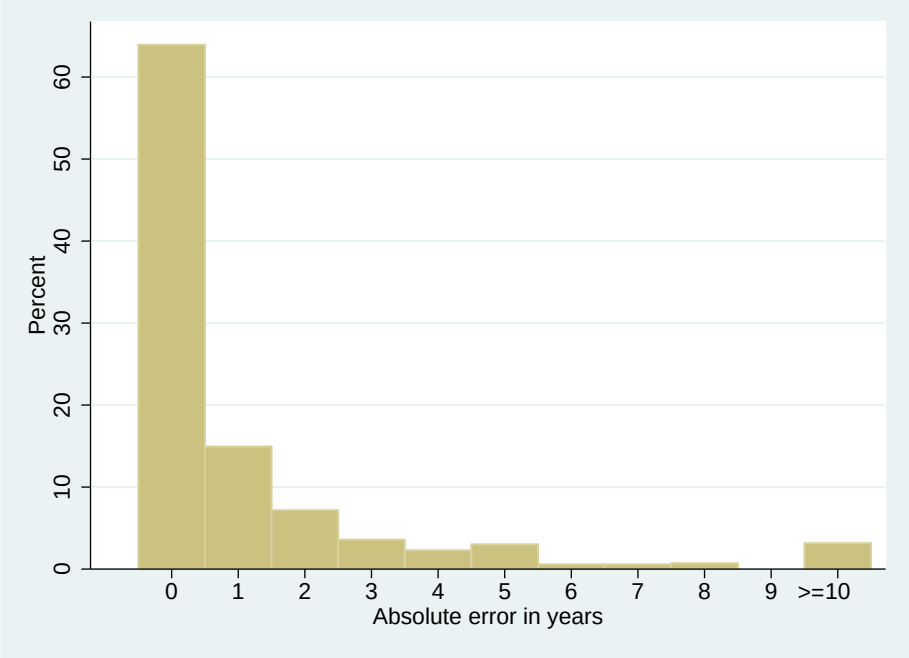

Figure 2: Distribution of Absolute Error

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.

2.1 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 3. 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 3: 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[6]". 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*[7]" 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 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 A.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[8], 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,

[6] `www.http://www.ldoceonline.com/dictionary/retire`

[7] `www.http://www.duden.de/rechtschreibung/Rente`

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

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.

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 Section 2.3.

2.2 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 3. 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.[9] 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 4. At first glance, the share of prominent years 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.[10] Therefore, the difference between reports in the administrative data and the SHARE data is not statistically significant.

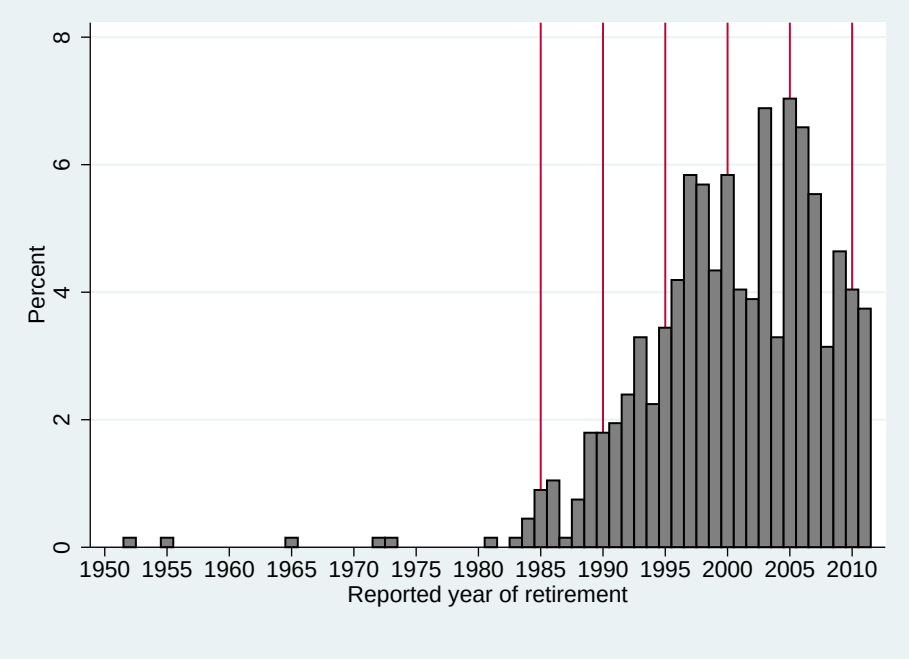

Figure 3: Distribution of Reported Years

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

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

Table 4: 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%)

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 5). 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 5: 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 6 shows, there is little variation in respondents' calculation ability when referring to the German Wave 4 sample.[11] The majority of respondents (67%) made no mistakes and 19% made only one mistake.

Table 6: 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

[11] 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.

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 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 6 gives an example of how the outcomes are categorized.

Table 7: 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.[12] 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 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

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

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 4 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.[13] 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[14].

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.

[13] 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

[14] 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.

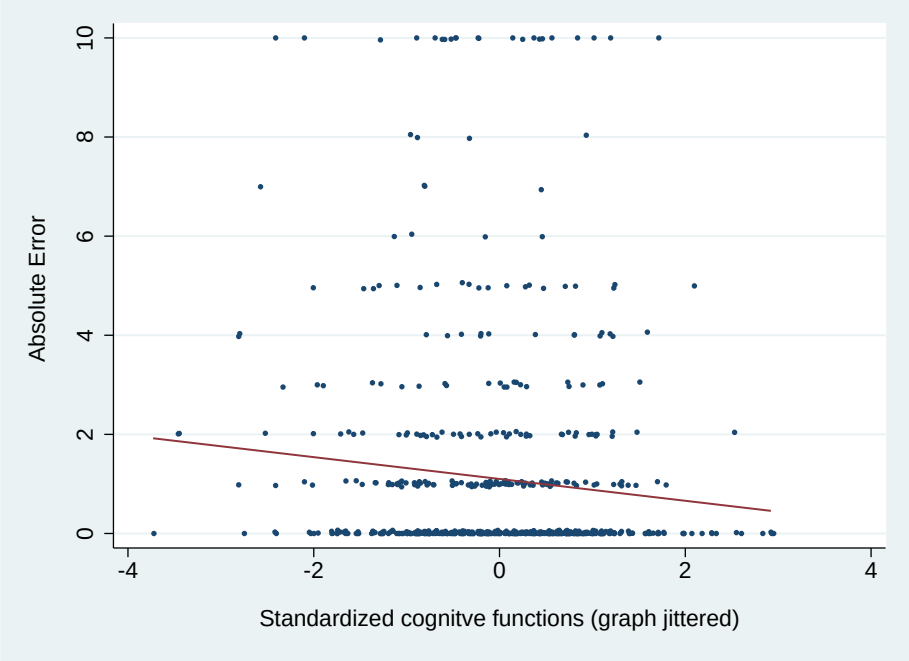

Figure 4: Cognitive Functions and Measurement Error

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 5. 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 linearity of the effect (results not shown) for cognitive abilities. The results for elapsed time are not that clear. Figure 6 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

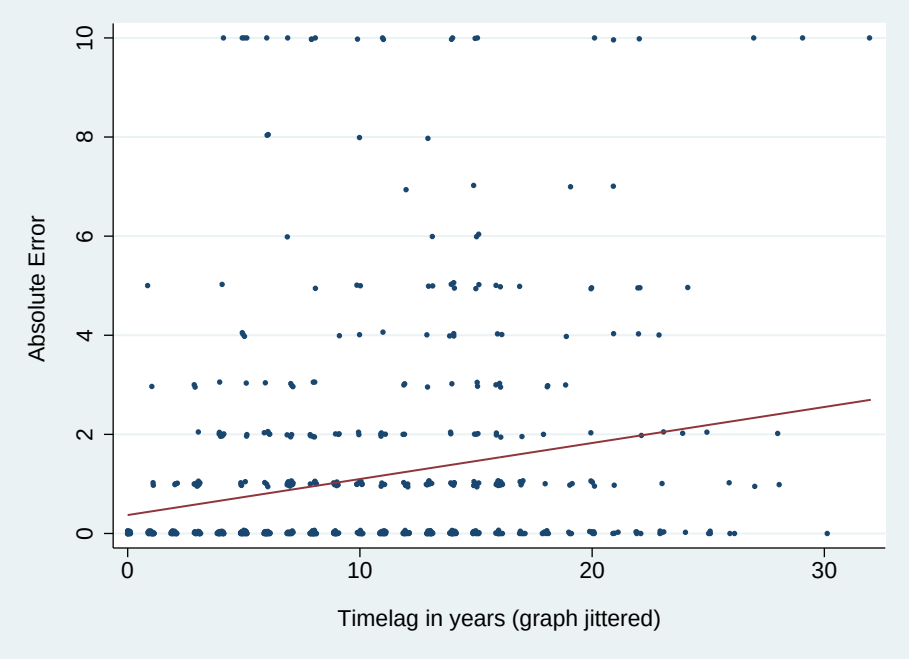

Figure 5: Time-lag and Recall Error

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.

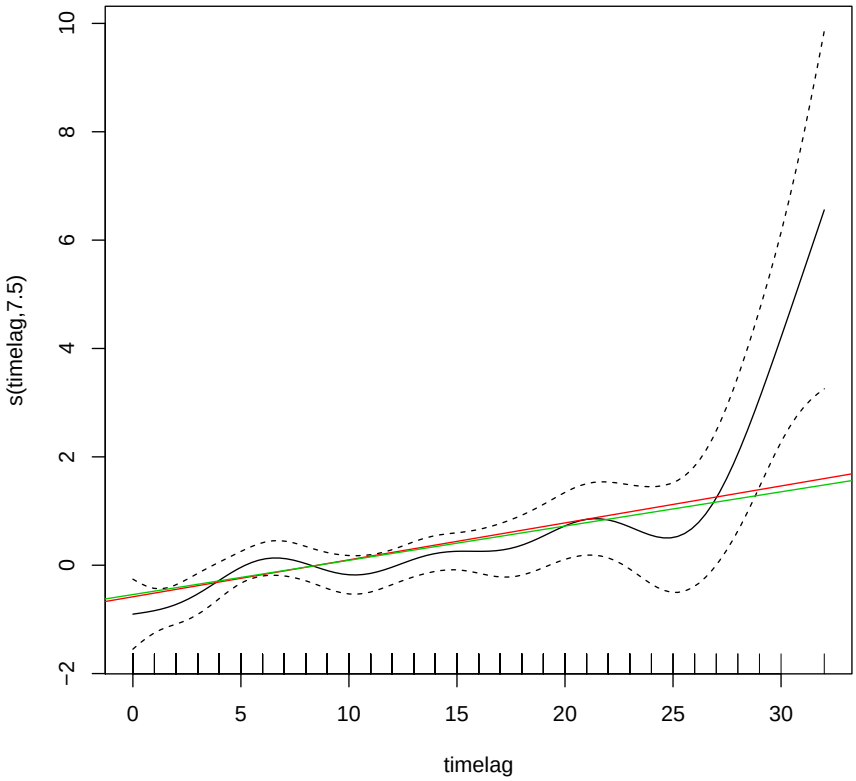

Figure 6: gam Regression: Time-lag and Measurement Error

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 8). 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[15]. 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' [16]), as well as those of category 3 (unpaid care), category 5 (disabled), and categories 6-8 (unemployment). The variable "*Soziale Erwerbssituation*" of the administrative 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.

[15]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

[16]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.

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 9). Surprisingly, there are only small differences between men and women when considering their last employment status.

Table 8: 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

Table 9: 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			

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 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.

2.3 Model and Results

The Sample

The sample of the following analysis consists of 668 cases. Table 10 gives an overview of the stepwise reduction of the sample of linked cases reported in Table 1. 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.

The following section is divided into two 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).

Table 10: 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 Absolute Error

Referring to Equation 1 on page 7, 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 2 on page 8 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:

$$Pr(y_i = 0|x_i) = \frac{exp(x_i\gamma)}{1 + exp(x_i\gamma)} = \pi_i \tag{2}$$

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:

$$Pr(y_i|x_i) = (1 - \pi_i)Pr(y_i|y_i > 0, x_i) \text{ for } y > 0 \tag{3}$$

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

$$\mu_i = E(y_i|x_i) = [\pi_i \times 0] + (1 - \pi_i) \times E(y_i|y_i > 0, x_i) \tag{4}$$

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

$$E(y_i|y_i > 0, x_i) = \frac{\mu_i}{1 - (1 + \alpha\mu_i)^{-1/\alpha}} \tag{5}$$

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

$$Var(y_i|x_i) = E(y_i|x_i) = \mu_i = exp(x_i\beta) \tag{6}$$

The Negbin I model implies a constant variance–mean ratio and can be written as[17]

$$Var(y_i|x_i) = \mu_i + \alpha\mu_i \tag{7}$$

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

$$Var(y_i|x_i) = \mu_i + \alpha\mu_i^2 \tag{8}$$

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.

$$Var(y_i|x_i) = \mu_i + \alpha\mu_i^P \tag{9}$$

Following Farbmacher (2013), I calculated three different Negbin versions (I, II, and P) to find the adequate model with the best model fit.[18] Table 11 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.

[17] 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 $Var(y_i|y_i > 0, x_i)$.

[18] 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)).

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. 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).

Table 11: 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

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:

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

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'[19] (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 (10). 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

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

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 11) when analyzing the absolute error. Whether the time-lag also influences the direction of the error, will be analyzed in the following. Figure 7 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 7.

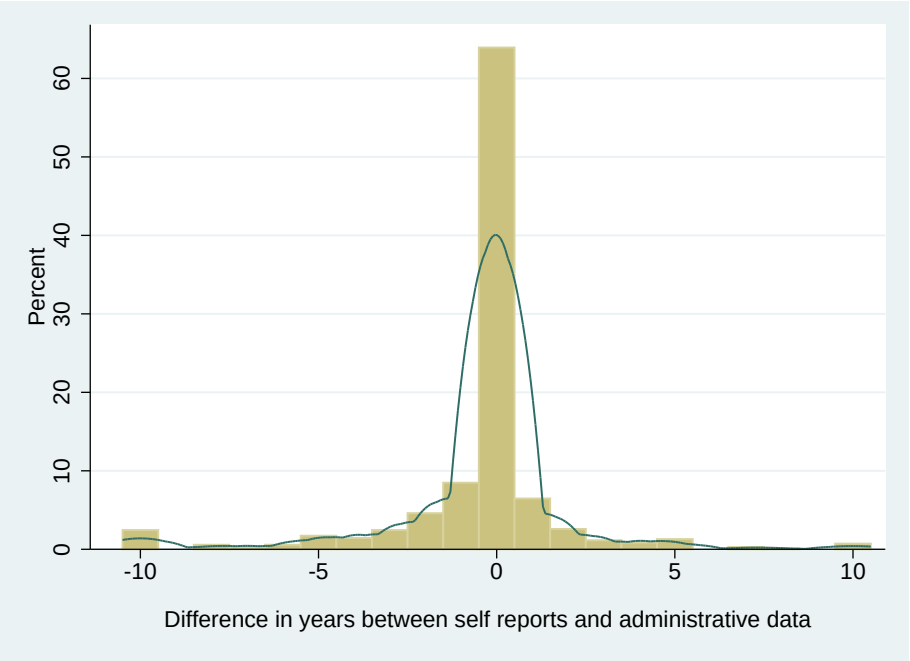

Figure 7: Distribution of Error

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 12 shows the results for the multinomial logistic regression, which includes 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[20]. Table 13 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.

[20]It is also possible to compute an LR test but since the results of the Wald and the LR test provide similar results, I decided to use the Wald test as the LR cannot be calculated while using robust standard errors.

Table 12: 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 13: 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

2.4 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, for those respondents who misreported the year, different determinants could be identified which are correlated with the error. The first model presented 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 is 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, that is no 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.

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 is 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 was 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 2.3, 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).

The reduction in the number of cases was based on different reasons, such as the availability 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 Korbmacher and Schröder (2013) 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.

3 Bibliography

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.

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.

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.

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.

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.

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.

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.

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.

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.

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., 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.

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.

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.

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

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.

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. 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., 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.

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.

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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

A Appendix

A.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[21]. Question 'ep213' asked about the year they first received a pension, distinguishing between the different income sources [22]. 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 14 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.

[21] **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?"

[22] **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 14: 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

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.

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

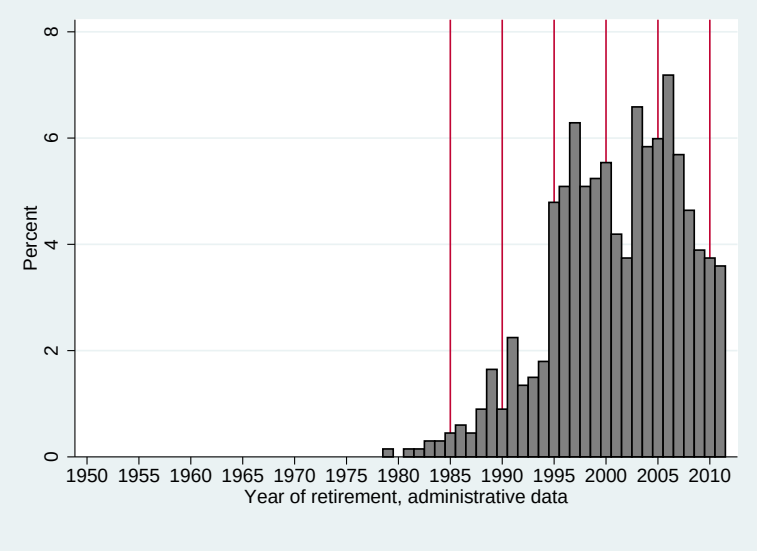

Figure 8: Distribution of Reported Years