Institut für Arbeitsmarkt- und Berufsforschung
Die Forschungseinrichtung der Bundesagentur für Arbeit

IAB

IAB-Bibliothek 362
Die Buchreihe des Instituts für Arbeitsmarkt- und Berufsforschung

Patterns and impact of longitudinal measurement error for welfare receipt

Johannes Eggs

Dissertationen

wbv

Institut für Arbeitsmarkt- und Berufsforschung
Die Forschungseinrichtung der Bundesagentur für Arbeit

IAB

IAB-Bibliothek 362

Die Buchreihe des Instituts für Arbeitsmarkt- und Berufsforschung

Patterns and impact of longitudinal measurement error for welfare receipt

Johannes Eggs

Dissertationen

wbv

Bibliografische Information der Deutschen Nationalbibliothek

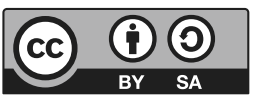
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors der Sozial- und Wirtschaftswissenschaften (Dr. rer. pol.) der Otto-Friedrich-Universität Bamberg

vorgelegt von

Dipl. Ver.Wiss. Johannes Eggs, M.Sc.,
geb. am 21.05.1981 in Freiburg im Breisgau

1. Gutachter	Prof. Dr. Mark Trappmann
2. Gutachter	Prof. Dr. Guido Heineck
Tag der Einreichung	04.08.2015
verteidigt am	09.12.2015



Herausgeber der Reihe IAB-Bibliothek: Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit (IAB), Regensburger Straße 104, 90478 Nürnberg, Telefon (09 11) 179-0 ■ **Redaktion:** Martina Dorsch, Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit, 90327 Nürnberg, Telefon (09 11) 179-32 06, E-Mail: martina.dorsch@iab.de ■ **Gesamtherstellung:** W. Bertelsmann Verlag, Bielefeld (wbv.de) ■ **Rechte:** Kein Teil dieses Werkes darf ohne vorherige Genehmigung des IAB in irgendeiner Form (unter Verwendung elektronischer Systeme oder als Ausdruck, Fotokopie oder Nutzung eines anderen Vervielfältigungsverfahrens) über den persönlichen Gebrauch hinaus verarbeitet oder verbreitet werden.

In der „IAB-Bibliothek" werden umfangreiche Einzelarbeiten aus dem IAB oder im Auftrag des IAB oder der BA durchgeführte Untersuchungen veröffentlicht. Beiträge, die mit dem Namen des Verfassers gekennzeichnet sind, geben nicht unbedingt die Meinung des IAB bzw. der Bundesagentur für Arbeit wieder.

ISBN 978-3-7639-4111-7 (Print)
ISBN 978-3-7639-4112-4 (E-Book)
Best.-Nr. 300923 www.iabshop.de www.iab.de

Vorwort

Die statistische Modellierung von Zusammenhängen benötigt in der Regel Annahmen, um unverzerrte Schätzer zu ermitteln. Meistens ist es jedoch nicht möglich, diese Annahmen zu testen. Im Rahmen meiner Arbeit am am Forschungsbereich E3 Panel „Arbeitsmarkt und soziale Sicherung" des Instituts für Arbeitsmarkt– und Berufsforschung (IAB) in Nürnberg war es mir jedoch möglich, mit Daten zu arbeiten, mit denen sich Annahmen zu Messfehlern im Längsschnitt testen lassen. Dazu wurden Angaben zum SGB-II-Bezug aus Umfragedaten mit Einträgen aus Registerdaten der Bundesagentur für Arbeit verglichen. Solche zusammengespielten Daten ermöglichen eine große Anzahl von Forschungsvorhaben, die meiner Meinung nach erst oberflächlich angekratzt worden sind.

Die Forschungsfragen der Arbeit ergaben sich im Nachhinein fast automatisch aus den Forschungsschwerpunkten des Bereiches, da er sich unter Anderem mit der Untersuchung der sozialen Folgen des Bezuges von „Hartz IV"–Leistungen und Methoden der Umfrageforschung beschäftigt.

Die Arbeit wäre ohne eine Reihe von Personen nie entstanden. Hier ist an erster Stelle Prof. Dr. Mark Trappmann zu nennen. In Personalunion als Erstbetreuer, Bereichsleiter und Co-Autor war er der leise, aber unermüdliche Antrieb hinter der Arbeit. Danken muss ich auch Annette Jäckle, die als Co-Autorin zweier Artikel dieser Arbeit fungierte. Während der gemeinsamen Arbeit an den Artikeln habe ich viel über das Schreiben gelernt. Ebenso muss ich mich bei Prof. Dr. Guido Heineck bedanken, der sich bereit erklärte, als Zweitbetreuer dieser Arbeit zu fungieren.

Hervorzuheben sind auch eine ganze Reihe von ehemaligen Kollegen am IAB. Hier sind vor allem Arne Bethmann, Jonas Beste, Stephanie Gundert, Anita Tisch und Silke Tophoven zu nennen. Bedanken muss ich mich auch bei meiner Familie, die in der Zeit immer ein Rückhalt war und die sich im Laufe des Entstehungsprozesses beträchtlich vergrößert hat.

Neben den genannten Personen gibt es noch eine Vielzahl weiterer Personen, die bewußt oder unbewußt zu dieser Arbeit beigetragen haben. Auch ihnen gilt mein herzlicher Dank. Falls jedoch in den Texten noch Fehler zu finden sind, sind diese selbstverständlich dem Autor allein vor die Füße zu legen.

Bonn, im November 2016

Contents

List of Tables

List of Figures

1 Introduction

Empirical research in economics and the social sciences depends in many cases on the use of survey data. However, survey data is always influenced by a range of errors that can distort the findings. In order to derive methods that might decrease survey errors, the analysis of the extent and impact of survey errors is important. The studies in this thesis focus on the analysis of one particular type of survey error, the measurement error. Measurement error occurs, if the response of an individual differs from the true value. E.g. if an actual smoker states that he or she is a non-smoker when asked about the smoking behaviour. Measurement error can cause response bias, which can endanger the validity of survey results. The thesis focuses on measurement error in longitudinal panel surveys. In longitudinal panel surveys individuals are participating repeatedly so that change in individuals over time can be observed. If characteristics of a respondent can change over time, so can measurement error. However, most research on the influence and extent of measurement error in surveys is conducted using cross-sectional and not longitudinal data. Some research was conducted for two subsequent panel waves (e.g. Bollinger and David 2005; Bound and Krueger 1991; Freeman 1984; Lynn et al. 2012). Lack of longitudinal research is caused by lack of longitudinal validation data, which is rarely available. In this work, longitudinal validation data is available for five consecutive years. The validation data can be linked on individual level to survey responses of an annual panel study. Focusing on measurement error for welfare receipt, a range of research questions will be tackled in this thesis: (1) How does measurement error behave over time? (2) How does the measurement error affect panel regression models? (3) How does the measurement error affect time-to-event models? (4) To which extent can flaws in the survey administration increase measurement error?

1.1 The Total Survey Error framework

The analysis of measurement error can be embedded in the wider field of research that focuses on the assessing of errors in surveys. A survey is "a systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the entities are members" (Groves et al. 2004, p. 2). To put it slightly different, the aim of a given survey is to measure or to derive a statistic X or a set of statistics $\mathbf{X}$ from a target population. The statistics can be the mean, the median, the prevalence, the incidence or the estimation results of multivariate methods. However in reality, not x but $x*$ will be the outcome of the analysis as the empirical survey estimate $x*$ of the parameter of interest will always be somewhat different from the true underlying

parameter. Also, random factors will cause variance between surveys, even if they are identically administered. Analyzing survey errors and trying to decrease the difference between x and $x*$ as well as the variance is the aim of the scientific field of survey methodology. In survey methodology, one of the most commonly used framework for the assessment of survey errors is the Total Survey Error (TSE) framework (Groves et al. 2004) that is shown in graph 1.1.

The TSE framework uses the core elements that are necessary to derive a survey statistic as its foundation. The survey elements are depicted as rectangles in the graph. The outcome of each survey step can be influenced by its respective error type, which are depicted as ellipses. Depending on whether the error-generating process is random or non-random, errors can increase the variance of and bias the estimator. The graph consists of a measurement branch and a representation branch. A measurement branch is needed to construct quantitative descriptors for each unit. A representation branch is needed to construct a set of units that represent the target population. The left branch of the graph focuses on the measurement dimension for a given item, while the right branch focuses on the representation of the target population.

Giving a brief overview of the graph 1.1 and beginning with a description of the left measurement branch, the aim of the researcher is to measure a construct or item μ for each single unit or respondent i. However, the quality of the measurement is influenced by the validity of the instrument, respectively how well is the measure related with the underlying construct. This relation might not be perfect. Thus, in reality Y_i is measured instead of μ_i. Even if the instrument is valid, measurement error can cause further bias, if respondents can not retrieve the true answer or might edit their response in the answering process. Thus, y_i will be recorded for each respondent. According to Alwin (2007), measurement errors can be most critical in empirical research. As the analysis of measurement error is the topic of this thesis, measurement error will be described in detail in the next section. Errors can also occur during the data entry respectively data processing of the response. E.g. outlying values might be censored or deleted, even if they are actually true. As the final value of these measurement steps, one would subsequently derive an edited response y_{ip} for each respondent in the data. Thus, on the measurement level, for each unit a value y_{ip} is used for the analysis that deviates from the value of the true construct μ_i.

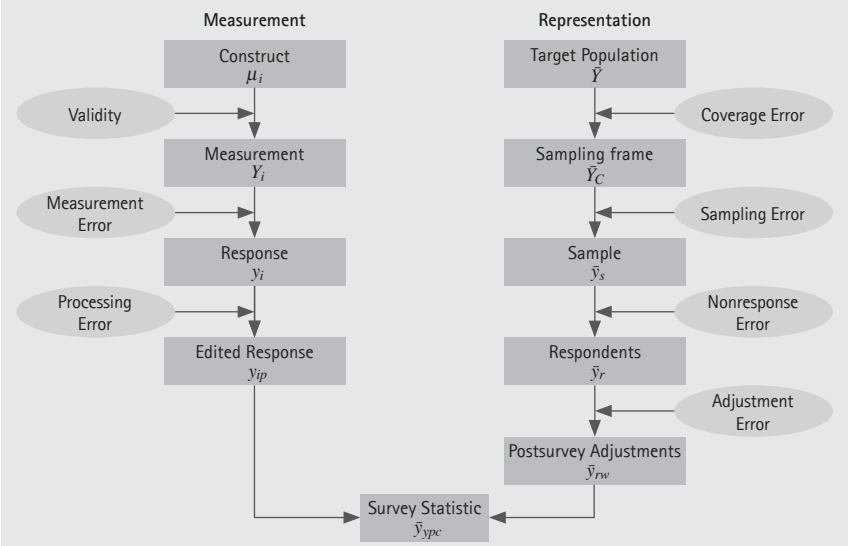
The right branch of the graph focuses on the representation of the underlying target population in the resulting statistic. Using the mean for a given variable as the example of choice, the aim of a given survey is to measure $\bar{Y}$ of a target population. In order to sample from a target population, a sampling frame is needed. If there are systematic differences between the statistic of the sampling frame $\bar{Y}_C$ and that of the target population, one speaks of coverage error. Sampling error is an inherent

error source in sample surveys and can be splitted into sampling bias and sampling variance. Sampling variance as one type of sampling error arises as the parameters of each possible sample realization vary from each other. The second type of sampling error is sampling bias which is caused by faulty selection probabilities in the sampling frame. The statistic for the sample composition is then $\bar{y}_s$. The composition of the resulting sample can be influenced by non-response error as individuals, that were drawn in the study sample, may choose to not participate in the survey, which can in turn cause non-response bias, if the decision to respond is caused by a non-random process. Thus, the statistic for the participating respondents will be $\bar{y}_r$.

Post-survey adjustment can be used to increase the generalizability of the survey if the survey used a stratified sampling design or is affected by non-random non-response. The adjustment can be complicated and normally results in the calculation of survey weights. However, it is possible that due to adjustment error the application of weights would cause an increase and not a decrease of the bias for the statistic of the weighted sample $\bar{y}_{rw}$.

The sample composition $\bar{Y}_{rw}$ and the edited responses y_{ip} of each individual are then used to derive the final survey statistic $\bar{y}_{ypc}$. The total survey error for each estimator can then be summarized using the mean squared error. The mean squared error is the sum of the squared bias and the variance. Each error type can contribute to an increase of both error and variance. A more detailed description of each of the steps is beyond the scope of this introduction, but discussions of the TSE can be found in e.g. Groves et al. (2009).

Figure 1.1: The Total Survey Error framework according to Groves et al. (2004)



The main advantage of the TSE is that it splits the survey process into definable and manageable entities. Thus, the TSE framework enables the researcher to analyse specific error sources and deliberate on the minimization of the mean squared error before and after the actual data collection by choosing the appropriate survey design option for each step of the survey process. It has to be mentioned that the concept has some weaknesses and peculiarities. The application of the TSE framework does not result in a quality score for a survey. On the one hand, there is not one single score for one survey, but each specific survey statistic would have its own specific score as the combinations of different representations of each construct and of the different response patterns cause unique ramifications on the specific survey statistics. As a consequence, design choices that could decrease the amount of error for one statistic could increase the error for a different statistic. Hence, researchers have to focus on key statistics, when trying to minimize survey error. On the other hand, the mean squared error for a given statistic can normally not be calculated, as the true value for the complete statistic is mostly not measurable or only under great costs. However, it is possible to evaluate the impact of survey errors for selective survey steps. Most of the evaluation studies focus on descriptive statistics. However, how the errors might affect the results of multivariate models is mostly disregarded, even if applied researchers are commonly relying on multivariate modeling to analyse the data. The TSE framework also omits important aspects of survey quality as it focuses only on statistical properties of estimators. It does not discuss the requirements of users regarding the "fitness for use" of the survey. In order to provide the "fitness for use" , the survey has to be credible, relevant and on time. The basic TSE framework of graph 1.1 does not cover all potential error sources. Types of errors that are specific in longitudinal surveys are not explicitly described in the framework. As this thesis focuses on longitudinal data, specific errors in longitudinal data will be described in the next section.

1.2 The longitudinal perspective

This thesis focuses on longitudinal data where data from the same subjects is collected repeatedly over time. In the social sciences, longitudinal data is collected in so-called panel surveys.[1] The main advantage of longitudinal data is that it allows the observation of change in subjects over time. In contrast to cross-sectional studies, it is thus possible to analyse the gross composition of inand outflows, changes in units over time and time-related events. In principle, panel surveys are

1 Repeated cross-sectional studies that are sometimes called "trend studies" could also be defined as longitudinal studies. In this thesis, I will use the term "longitudinal" exclusively for studies that collect data from the same subjects at repeated time points.

affected by the same types of error as any other survey which were described in the previous section. However, the quality of longitudinal data can additionally be influenced by specific problems and varieties of survey errors that can only manifest themselves in longitudinal data.

One specific problem in longitudinal surveys is panel conditioning. Panel conditioning occurs, if attitudes, behaviors, and states of respondents are influenced by measuring them (Warren and Halpern-Manners 2012). E.g., it can occur if a survey question contains information which was previously unknown to the respondents and the respondent acts upon between the current interview and the next interview (Warren and Halpern-Manners 2012). Thus, a survey question can act as a kind of stimulus that is given to survey participants. This can endanger the external validity of the results of a panel study, as the stimulus was not given to the overall population.

Panel surveys can be influenced by a certain form of coverage error. In a panel survey, the initial sample tries to cover a specific target population at a specific point of time. However, over time a target population can change as outflows and inflows influence the composition of the target population. If no measures are taken to adapt for this change, the study sample will diverge more and more from the composition of the current target population as it only includes members of the original target population.

Non-response error has also a longitudinal version. In addition to initial non-response, panel surveys also face the problem of longitudinal non-response which is commonly known as panel or sample attrition. Panel attrition occurs if respondents do not participate in subsequent panel waves. This causes at least a decrease of the sample size and thus a loss of analytical power, if the attrition is based on a random process. However, if the attrition process is non-random and related to unknown factors, panel attrition can cause substantial bias (Trivellato 1999).

Measurement error can cause specific problems in a longitudinal setting. As this is the central topic of this thesis, measurement error and its longitudinal properties will be discussed in more detail in the next section.

1.3 Measurement error

The total survey error approach distinguishes between different error types that can be tackled separately. This thesis focuses on one particular type of survey error, the measurement error. Measurement error occurs, if the measured value for a unit differs from the true value that might have been collected with the specific method of measurement. E.g., a telephone survey collects data on the individual body weight and a respondent states that he is four pounds lighter than he is.

In surveys, information can be collected directly for some items, if the interview is conducted in person. E.g., a scale is brought along by the interviewer to measure the weight of the individual. However, information is mostly collected by asking for the information. The information is given as a response to the question. The response to each question is the result of a response process. Thus, measurement error arises during the response process of the individual to a specific question. The most widely used model for the response process has been described by Tourangeau, Rips, and Rasinski (2000) and is based on the older model by Cannell, Marquis, and Laurent (1977). They separate the response process in four components. The first component is comprehension. The individual has to process the question and has to identify, which information is sought. The second component is retrieval. The cues in the question activate retrievable memories or impressions. The third component is judgement. The retrieved information has to be processed and evaluated. The fourth component is response. Before the response can be expressed, it has to be formulated and edited.

Errors can be caused in each of the four steps of the response process, which in turn will distort the outcome. Respondents might be unable to interpret or misunderstand the question. This can happen, if the question is too complex or not precise. Retrieval might not be possible or distorted, if respondents forgot the information or misplace the information on the internal timeline. This might happen in retrospective questions, where information about past behaviour and events is collected. This particular error is known as recall error. The information sought might also be stored in a different context and the question fails to activate the necessary cue. The respondents could employ a flawed estimation strategy or have sub-optimal judgements. Respondents can be also inclined to take cognitive shortcuts. Such behavior is coined as satisficing (Krosnick 1999). The retrieval and judgemental process can be tiring and exhaustive. Hence, respondents may stop the response process after having derived the first possible and acceptable answer in order to shorten the response process. Even if respondents derive the true answer, respondent can deliberately misreport due to social desirability (Krumpal 2011). When editing their response, individuals might judge their answer against social norms. If their response is adverse to social acceptable beliefs or norms, they edit the response accordingly. Respondents usually underreport socially undesirable outcomes like drug use and overreport socially desirable outcomes like voting. Acquiescence may also influence response behaviour. Acquiescence can occur in the judgement and editing phase. Respondents are more likely to agree than to disagree with survey items. Acquiescence, satisficing and social desirability are sometimes subsumed under the term of suboptimal responding (Thomas 2014). Having reached the response, individuals have to adapt it to the given response options of the questions. Errors might also occur even prior to the response process, if the respondents failed to store

the relevant information altogether. This is especially common in survey reports on diets. Individuals are unlikely to store relevant information on food intake and thus rely on random guesses when reporting their dietary intake (Freedman et al. 2015).

Thus, survey data that is based on responses is likely to be distorted by some forms of measurement error. Measurement error causes bias in statistical parameters and leads to a loss of power when estimating associations between variables or trying to establish a causal process (Carroll et al. 2006; Hernan and Robins 2014). For the field of nutritional studies, it is even sometimes argued that the scientific value of results based on survey responses is small due to measurement error (Ioannidis 2013).

When describing measurement error, the notations of classical test theory are widely used (Novick 1966). It is assumed that for each unit i, there is a true value xi and an observed value $x*_i$ for a given construct. Then, the measurement model can be stated as

$$x*_i = x_i + \varepsilon_i \tag{1.1}$$

where $_i$ is the additive measurement error for individual i. The relationship between these three values defines the error model. The most common error model is the classical measurement error model (Carroll et al. 2006). In this case, the measurement error has a mean of zero ($E(\varepsilon_i \mid x_i) = 0$) and thus adds random variability to the true score. Classical measurement error can lead to a loss of power and bias parameters of regression models. Common extensions of the classical error model assume that the error is uncorrelated with the true value and, when conducting multivariate analyses, uncorrelated with the additional model variables. If this is the case, one speaks of non-differential measurement error.

There are two common methods to assess measurement error in surveys. The first method is the use of an additional, external data source that contains the true value for the item of interest. To validate the measure of the primary data requires that information on the same construct can be collected in both survey and validation data. The alternative method is replication. Information on the same construct is collected repeatedly over time. The first method allows the assessment, whether the measurement is valid. The second method allows the assessment, whether the measurement is reliable.[2] However, validation data is rarely available and may only contain information on a subset of survey items. Thus, for a broad range of research questions, repeated measures are the only feasible approach to assess measurement error.

2 There are applications of repeated measures that might allow the assessment of validity like "Multitraitmultimethodmatrixes". Yet, such methods are based on assumptions and can not guarantee the validity of the measurement (Schnell, Hill, and Esser 2008, p. 160).

1.3.1 Measurement error in longitudinal studies

There are distinct features of measurement error in longitudinal settings Lynn (2009, p. 16). In longitudinal panel surveys individuals are participating repeatedly so that change in individuals over time can be observed. If characteristics of a respondent can change over time, so can measurement error. Thus, the extent of measurement error can be larger in longitudinal data than in cross-sectional data (Freeman 1984). If a variable is likely to be observed with error at each point in time, the proportion of false measures increases with each subsequent panel wave. If only a small number of observation change their status, the cross-sectional measurement error will decrease the number of correct transitions in the observational data and the incidence of change will be overestimated.

The bias for in- and outflows can be additionally increased, if information on specific states is collected retrospectively and not only for the current date of the interview. In this case, a so-called seam effect can be observed. The seam effect describes the phenomenon that one can find a heaping of reported changes on the seam between subsequent panel waves. The earliest description can be found in Czajka (1983). The seam effect is not only caused by reporting behaviour, but also by data processing (see Callegaro (2008) for an extensive review on the seam effect). The quality of the collected retrospective data in panel surveys can be also influenced by telescoping. Telescoping was described first by Neter and Waksberg (1964) and Sudman and Bradburn (1973) and describes the misreporting or mislocation of events on the time line. Individuals tend to place events closer to the present than they occurred in reality.

One way to decrease measurement error in longitudinal surveys is the use of dependent interviewing. With dependent interviewing, the answer of the previous interview is used in the present interview. The reminding can be implemented in different ways. Respondents can be proactively or retroactively reminded of their previous answer. With proactive dependent interviewing, respondents are reminded of their previous answers, either to ask whether their status has changed, or as an anchor for asking about events since the previous interview (Mathiowetz and McGonagle 2000). With retroactive dependent interviewing, the information is used as a corrective follow-up, and the previous response is only used, if the status has changed from one wave to the next. Dependent interviewing is now used in most panel surveys to improve the measurement of key economic information. Proactive dependent interviewing is most commonly used to collect information about labour market activities (e.g. in the UKHLS, CPS, NLSY97, HRS, ELSA, SLID), income sources and pension plans (e.g. UKHLS, SIPP, HRS), assets (SIPP, NLSY97), but also for partnership histories (e.g. UKHLS, SIPP, NLSY97, Pairfam), citizenship (SIPP),

or health behaviours and conditions (ELSA).[3] Dependent interviewing reduces the number of erroneous transitions and increases data quality as it provides respondent with additional mental cues and provides a temporal bound. A temporal bound can reduce the effects of telescoping, as it provides the respondent with a time frame, which helps individuals to place the events in temporal order.

1.4 Welfare receipt

The topic of the thesis is the analysis of longitudinal measurement error for welfare receipt. The most common welfare program in Germany is unemployment benefit II (UB II).

Unemployment benefit II was introduced in 2005 as one part of the "Hartz reforms". UB II is also commonly known as "Hartz IV" since it was as introduced as the fourth step of these reforms.[4]

Prior to 2005, unemployment benefit, unemployment assistance and welfare were the main pillars of the social security system in Germany. Unemployment benefit was insurance-based and was granted to people, who became unemployed and were employed for at least 12 months before the beginning of their unemployment. The amount of unemployment benefit was a proportion of the previous income. The entitlement expired after up to 32 month. The length of the entitlement depended on the age. Older individuals were entitled to longer entitlements.

Unemployment assistance was the second tier, means-tested and tax-based. The amount was related to the previous earnings, however the proportion was lower than for the unemployment benefit. It did not expire, but it had to be annually applied for. The amount was 53 % of the last income after taxes on individual level. Also, recipients had to take up only "suitable" jobs. Thus, unemployment assistance still secured the prior social status. Unemployment benefit and assistance were administrated by the German federal employment agency.

If no entitlement for unemployment benefit or assistance existed or if the amount of those was too low, the means-tested social assistance provided the third tier. Provision of social assistance was organized on municipal level. The amount of social assistance was fixed.

With the implementation of Hartz IV, the old unemployment benefit became the unemployment benefit I. The maximum duration of unemployment benefit I

3 UKHLS = UK Household Longitudinal Study, CPS = Current Population Survey, NLSY97 = National Longitudinal Survey of Youth 1997, HRS = Health and Retirement Study, ELSA = English Longitudinal Study of Ageing, SLID = Survey of Labour and Income Dynamics, SIPP = Survey of Income and Program Participation, Pairfam = Panel Analysis of Intimate Relationships and Family Dynamics.

4 For a detailed description of the German social security reforms, it is referred to Eichhorst, Grienberger-Zingerle, and Konle-Seidl (2010).

receipt was curtailed to 12 months for younger persons and to 18 months for older people. Unemployment assistance and social assistance were merged into the new unemployment benefit II. UB II is means-tested and granted to people who might be able to work but could not provide sufficient economic resources from other income sources or savings.

In contrast to the abolished unemployment assistance, UB II disregarded the level of previous income. Unemployment assistance was also seen as an insurance benefit, while UB II is seen as a welfare benefit with the associated lower social status. It provides basic provisions and the amount of monetary assistance is a lump sum, only depending on the size of the household. This meant, that the old principle of status protection was abolished and replaced by a welfare concept, where poverty should be prevented. Another major deviance from the prior system was that UB II is not distributed on the individual level but on the household level as the eligibility to claim UB II is based on the economic situation of the household.[5] The basic amount of UB II for a single person is 382 Euro as in 2013. In November 2012, 6.03 million people received UB II (Statistik der Bundesagentur für Arbeit 2013b).

One of the aims of the implementation of UB II was the activation of the large number of long-term unemployed. UB-II-recipients have to accept job offers, even if they are below the actual level of qualification. Also, a range of measures was introduced or redesigned that provide wage subsidies, start-up subsidies and enabled the creation of jobs with reduced social security contributions. It was thought that these kind of jobs could serve as stepping stones for UB-II-recipients out of benefit dependency. New public employment services were created especially for UB-II-recipients, which should promote the reintegration through a "carrot and stick" principle. Regular appointments have to be met by recipients, where the recipient has to prove his willingness to take up a job. If the recipient is not seen as cooperative, the amount of UB II can be temporarily cut. All household members, that are capable to work, have to comply to the demands of the public services. However, also retraining, skill enhancement measures or public employment opportunities are offered by the public services to enhance employability.

Not all recipients of UB II are necessarily unemployed respectively have to take up employment. A large proportion of UB-II-recipients are "Aufstocker" i.e. employed persons whose income from employment is below subsistence level and therefore elevated by UB II (Bruckmeier et al. 2013). In September 2012, 1.33 million people

5 To be more exact, UB II targets a benefit unit. A benefit unit consists of at least one adult plus their spouse (if applicable) plus any dependent children they are living with. As a benefit unit is in most cases congruent with the household. In this study household is used as a synonym for benefit unit, unless it is specied differently.

were working UB-II-recipients (Statistik der Bundesagentur für Arbeit 2013a). It is also possible, that recipients have to provide care to children or relatives and hence can not take up regular work (Beste, Bethmann, and Trappmann 2010).

1.5 Measurement error for welfare receipt

The analysis of welfare receipt is an important field of research in the empirical sciences as it provides policy makers with necessary information regarding the effectiveness and repercussions of the respective welfare programs. However, many of these studies rely on survey data and welfare receipt is known to be misreported by survey respondents (Bound, Brown, and Mathiowetz 2001). The resulting measurement error can endanger the validity of the results.

As has been discussed in a prior section, misreporting in survey responses can be caused by a range of factors. Respondents might underreport their welfare receipt. Underreporting can occur due to forgetting, if the respondent receives a multitude of benefits. For retrospective questions, respondents can misplace their receipt on their internal time axis on the outside of the reference period. Underreporting might also be due to social desirability, if respondents feel stigmatized by welfare receipt. A different reason for the underreporting might be that respondent might be unwilling to disclose information, which they feel is sensitive. Respondents might also overreport benefit receipt. Overreporting might be due misclassification of the benefit or if the receipt is misplaced on the inner side of the reference period.

For welfare receipt, a range of studies found that underreporting is more commonly observed than overreporting (see Bound, Brown, and Mathiowetz (2001) for a review). Lynn et al. (2012) evaluated the impact of questionnaire design options on measurement error for a variety of benefits in the United Kingdom. They found underreporting for all but one benefit (the statutory pension) and higher rates of underreporting for means-tested benefits than for non meanstested benefits. Using SIPP data, Bollinger and David (1997) analysed the measurement error for the receipt of food stamps in the United States and found an underreporting of 12 % and argue that underreporting is caused by deliberate misreporting and not by natural memory processes like forgetting. Other research findings were that individuals in higher income households were more likely to underreport and that the error considerably biased the results of probit modelings of food stamp receipt. Analysing the same data for two waves, (Bollinger and David 2005) found no substantial decrease of the error from one wave to the next and a high degree of autocorrelation for the error.

Using data from the same panel study and from the same source of validation data that will be used in this thesis, measurement error for UB II was analyzed by

Bruckmeier, Müller, and Riphahn (2014, 2015) and Kreuter, Müller, and Trappmann (2010). All three studies found substantial underreporting and differential error patterns for UB-II-receipt. Analysing data of the first panel wave, Kreuter, Müller, and Trappmann (2010) found a difference of -9.2 % between the proportions for recipients according to survey and validation data. Analysing data of the fifth panel wave, (Bruckmeier, Müller, and Riphahn 2014) found an underreporting of 10.5 % and an overreporting of 1.5 %. Analysing data of the fourth panel wave, Bruckmeier, Müller, and Riphahn (2015) found an underreporting of 12.2 % and an overreporting of 1.9 %. In all three studies, it was found that younger respondents were more likely to underreport. Bruckmeier, Müller, and Riphahn (2014, 2015) assessed correlations for the underreporting of UB-II-receipt. Assessing the correlations they found that recipients that were more similar to non-recipients in terms of income, socio-economic and employment status were more likely to underreport. In Bruckmeier, Müller, and Riphahn (2015) they analysed additionally the correlations between the propensity to underreport welfare receipt and the interview style, interview situation, interviewer characteristics and characteristics of the respondent. They found that if the interviewer showed similar characteristics as the respondent, the respondent was less likely to underreport. One reason for the underreporting might be that the receipt of UB II is stigmatized in parts of the German population (May and Schwanholz 2013). Summarizing the empirical evidence from a range of countries and for different benefits, misreporting for welfare receipt in population surveys seems to be considerable and not the result of a random process. However, the research questions how measurement error for welfare receipt evolves over longer periods of time and whether it can bias the results of longitudinal substantive research were not yet analysed.

1.6 Data

1.6.1 Survey data

The data of the German panel survey "Labour Market and Social Security" (PASS) will be used to analyse measurement error for welfare receipt. PASS is especially suited for this undertaking as the survey was established to study the impact of the "Hartz reforms" in Germany. It was designed to assess the dynamics of unemployment benefit II (UB II) and how the welfare reforms influence the social situation of affected households and the persons living in them. Thus, a sufficient number of respondents will receive UB II and may misreport the receipt in practice. The panel study is conducted by the Institute of Employment Research (IAB) and is funded by the German federal ministry for Employment and Social Affairs.

In order to compare recipients of UB II with non-recipients, PASS was set up as a dual-frame survey. It consists of a recipient sample and a sample drawn from the general population. The recipient sample was based on the register for recipients of UB II of the German federal employment agency ("Bundesagentur für Arbeit") (FEA). 300 primary sampling units (PSUs) were drawn from postcodes. The probability for each PSU depended proportionally on the size of the population. Within each PSU, benefit communities for the recipient sample or addresses for the population sample were drawn. The population sample was based on a commercial database of household addresses. The population sample was stratified disproportionately by social status in such a way that households with a low social status were oversampled. The design ensured a sufficient number of cases with and without UB-II-receipt in the data.

PASS was set up as a household survey. This was necessary, since UB II provides economic resources not on individual level but on benefit unit level, which is in most instances congruent with the household.

Prior to the first survey interview, each household receives an advance letter that informs the household about the study and it also includes a leaflet describing the data safety protocol. To collect the information regarding the household, the head of the household is asked to complete a household interview containing among others questions on household composition and the receipt of UB II. For the recipient sample the head of the household is defined as the person that applied for the receipt of UB II. In the population sample, the head of the household is defined as the person that is most familiar with the overall situation of the whole household. After the household interview, every member of the household aged fifteen or older is asked to complete a personal interview including questions concerning individual labor market status and subjective health. Proxy interviews for currently unavailable members of the household are not allowed. So it is possible that not all members of a household complete the personal interview. PASS is using a mixed mode design. This means that the data is collected using either computer-assisted telephone or computer-assisted personal interviews. In order to increase the response rates, incentives are distributed. The strategy for the incentives changed over panel waves. In earlier panel waves, respondents received lottery tickets, conditional on participation. A mix of cash and lottery incentives was employed in wave 2. Since panel wave 4, each member of a household that completes the personal interview receives 10 Euro as incentive at the time of the first interview. In subsequent panel waves, the incentive is posted together with the advance letter that informs the respondents of the upcoming panel wave. In order to assess socio-economic dynamics, the survey is administrated annually. PASS is annually approved to fulfill the requirements of the data protection laws in

Germany. More detailed information regarding study design, sampling and content can be found in Trappmann et al. (2013) and in the documentation of the panel study (Bethmann, Fuchs, and Wurdack 2013).

In wave 1, data from 6 804 households (9 386 persons) in the recipient sample and from 5 990 households (9 586 persons) in the population sample was collected. The wave 1 response proportion was 26.7 % (RR1 according to the American Association of Public Opinion Research (AAPOR 2009)). Subsequently, refreshment samples are drawn at every panel wave. The refreshment samples consist of households that are first time recipients of UB II. Sizes of the refreshment samples vary around 1 000 households and 1 400 individuals.

1.6.2 Administrative data

To asses the measurement error for UB II in PASS over successive panel waves, validation data for the same time span is needed. For PASS, this is possible by linking the survey data on individual level to the entries of administrative records. These administrative records are provided by the FEA and can be used by the IAB for scientific research. The data contains information on employment spells that are liable to social security, unemployment spells, participation on active labour market programs and spells of UB-II-receipt.

The administrative records are a by-product of the main statutory task of the FEA, the disbursement of unemployment benefits and welfare benefits. The administrative records are based on multiple steps. The foundation of the records are the entries of the caseworker in the respective software. The data is processed on the level of the local employment agencies and then sent to the central data warehouse of the FEA on a monthly basis. There, the single data sets are combined and processed in order to calculate the national official statistics for a given month. The data is then further processed by the IT department of the IAB to allow scientific analyses. In this step, data, provided by the National Pension Insurance, is appended that contains data on employment liable to social security. The administrative records for the scientific use are updated annually and contain the entire longitudinal history up to a reference date.[6]

In this thesis, only the information for welfare receipt is used. This information is drawn from the data of the "Leistungshistorik Grundsicherung" (LHG) (Geschäftsbereich ITM 2014). The LHG is based on data which is created during the administration of welfare benefits to claimants. The data entries of the administrative records are then used as the true score in order to validate the

6 See Köhler and Thomsen (2009) for an overview of the administrative record data of the IAB.

survey responses of the individuals. The information is based on the information, whether welfare benefits were distributed by the local agencies that are responsible for the administration of UB II. Administrative data can not be necessarily be used to validate survey responses (Groen 2012). Administrative records are collected for administrative reasons, irrespective of its potential analytical value. Hence, the record construct that serves as foundation for a specific variable might not be necessarily comparable to its counterpart in the survey data as they are based on different forms of data collection and data logics. Here, the information, whether welfare benefits were disbursed for a given time period, is compared with the information, whether welfare benefits were received for the same time period. Hence, in this case it seems feasible to compare the record entries with the survey entries to assess the measurement error in the survey data.

The quality of the administrative data is high for UB-II-receipt (Köhler and Thomsen 2009) and can be used to define the measurement error in the survey data. Still, it has to be mentioned that administrative records are neither free of error nor necessarily complete. Errors can occur during data entry, data processing and data transfer. For the LHG data, gaps can be found. Gaps in the data were caused by the use of different types of software to administer welfare claims. Data from a subset of local agencies using one specific software is missing for most of 2005. Data from this time period can not be used or cases administered in the affected local agencies have to be dropped. Both strategies will be employed in this thesis. Over the complete time period, gaps can also be caused, when a local agency did not deliver the monthly update. If the missingness is not corrected in a subsequent data transmission, this can cause an artificial break in the welfare spell. The problem will be tackled by filling smaller gaps between two succinct spells of welfare receipt in the administrative records.

An additional problem arises due to the definition of a household. As has been mentioned, UB II targets the household respectively a benefit unit. It is possible that more than one such unit is living in one household. An example would be a household where more than two generations are living together. For a respondent, it is hard to distinguish between household and benefit unit, as this is a term defined by the social code II. In the PASS survey, individuals are asked whether UB II was claimed on household-level. In order to circumvent this discrepancy, one solution is the exclusion of such households with multiple benefit units from the analyses. In this set of studies, such households were also excluded. Thus, a range of steps is necessary to allow a valid comparison of the data sources to analyse the extent of the measurement error.

Not all respondents of PASS can be linked to the administrative records. Respondents had to give their informed consent to the linkage. The proportion

of respondents that gave consent varies from 76% to 87% between the first five panel waves (Berg et al. 2012). Over all five waves combined, 79% of the respondents gave their consent (Antoni and Bethmann 2014). In order to link the administrative records with the survey data, the following variables were used: date of birth, adress, first name, last name and the identification number of the benefit unit.[7] As PASS is based on two different samples, two different strategies were employed for the record linkage. For the register sample, first the ID of the benefit unit was used to link sets of survey and register information. Then, within the benefit unit, an exact linkage was possible for most of the cases using the remaining personal characteristics. For respondents from the register sample that could not be linked with this procedure and for respondent from the population sample, a step-wise procedure was employed using the full data of the administrative records as the linkage frame for the survey respondents. As a first step, exact matches were necessary for all of the characteristics. This condition was then relaxed and only an exact match was required for a majority of characteristics. In a last step, error-tolerant probabilistic linkage procedures were employed using the mtb software described in Schnell, Bachteler, and Reiher (2005). With such procedures, an equality score is calculated for each pair of the characteristics. This score is summed up to a quality index score. If the score is above a predefined threshold, a match is assumed.

Not all respondents could be linked as not all individuals have a record entry. Self-employed and public servants are less likely to be found in the records as such forms of work are not recorded in the administrative data. It is also possible that no match was possible due to errors in the register or survey information. A rule of thumb is that the more frequent an individual is in contact with the social security system the higher is the probability for a match. As the contact rate of UB-II-recipients with the the employment agencies is high, more than 98% of all consenting UB-II-recipients could be linked to the register data.

Still, patterns of consent and patterns of linkage can cause selection bias. The patterns of consent were analysed for a range of studies. Korbmacher and Schröder (2013) give a good overview over the current research on patterns of consent in different studies and analyse patterns of consent for the SHARE study. The age of the respondent is positively and significantly correlated with the decision to consent. Also, the interviewer seems to have a significant impact on the decision to consent. For the PASS study, patterns of consent were analysed by Sakshaug and Kreuter (2012) and Beste (2011). They also find an impact of age and interviewer

7 This description follows closely the text by Antoni and Bethmann (2014) for the PASS-ADIAB data. While the PASS-ADIAB is an excerpt of the administrative records for wider use, the linkage procedures were the same for the data used in this thesis.

on the decision of the individual to give the consent to data linkage. It is further argued in both studies that respondents are more likely to consent if they are more cooperative as respondents are less likely to consent, if they do not want to participate in further panel waves or refused to answer questions on the employment status and the financial status. Both studies assess that using only consenters causes minor selection bias. Sakshaug and Kreuter (2012) also suggest that the non-response error and measurement errors are contributing more to the total error than the nonconsent biases.

1.7 The studies

Previous studies analysed the measurement error for cross-sections of the PASS data. However, in longitudinal data, the analysis of longitudinal measurement error should be of equal or even larger importance as most scientific analyses are based on the longitudinal data. This thesis consists of four studies that focus on the longitudinal measurement error for welfare receipt in a panel study. Using the data described in the previous section, the measurement error can be assessed and be analyzed for a longer period of time than in previous studies. This section gives a short overview over the aims of the studies in this thesis. The first study focuses on descriptive statistics for each wave and focuses on the development of the measurement error over time. Descriptive statistics should be the foundation of empirical research. However, as most applications of survey data use forms of multivariate analysis, it has been criticized that research under the TSE framework focuses too much on descriptive statistics and applies unrealistic assumptions when modeling the structure of the error (Groves and Lyberg 2010, p. 875). Thus, very little evidence is known for the likely effects of errors when benefits from individual programs are used as either dependent or explanatory variables (Bound, Brown, and Mathiowetz 2001, p. 3779). The second and the third study will evaluate the effect of the measurement error on models, as they focus on various assumptions on longitudinal measurement error and also analyze the impact of the measurement error on analytical models.

1.7.1 Will respondents eventually get it right? Changes in measurement error across five waves of a panel survey using dependent interviewing

The first study gives a detailed description of the development of the measurement error over time. As has been mentioned, the measurement of welfare receipt is affected by underreporting (Bound, Brown, and Mathiowetz 2001). In a panel survey there are however reasons to expect reporting quality to improve across

panel waves: respondents who are more likely to misreport may be more likely to attrit; over time respondents may gain trust in the survey and therefore underreport less; the stigma in the population associated with a newly introduced benefit type may fall over time, again reducing underreporting; and respondents may remember what they will be asked about in the interview and become better at remembering relevant information. In addition, panel surveys can make use of responses from previous interviews, to remind respondents of income sources they have received in the past.

In this study we use data from the first five waves of PASS, linked to individual administrative records on UB-II-receipt. This data is used to examine the following questions: (1) Does data accuracy change over waves of a panel survey? (2) Why does data accuracy change over waves of a panel survey? (3) Do changes in data accuracy alter substantive research conclusions?

1.7.2 Impact of measurement error for welfare receipt on panel models

Fixed-effects models are a popular tool for the analysis of panel data in economics (Angrist and Pischke 2008) and the social sciences (Allison 2009). However, fixed-effects models can be affected more strongly by measurement error than cross-sectional models as they solely rely on transitions from one state to another for the calculation of the estimates (Freeman 1984). Previous studies analyzing the influence of measurement error on fixed-effects models were only able to use two panel waves to analyze the impact of measurement error. In this study, the impact on fixed-effects models can be assessed for five panel waves.

Previous research has shown, that welfare receipt tends to be underreported in surveys (Bound, Brown, and Mathiowetz 2001). For Germany, it has been shown that unemployment benefit II is underreported up to 15% (Kreuter, Müller, and Trappmann 2010). However, it has also been shown that the extent of the measurement error for UB-II-receipt decreases over subsequent panel waves (Jäckle, Eggs, and Trappmann 2015). As a consequence, false transitions into and especially out of unemployment benefit-II-receipt will emerge in the data. As fixed-effects estimates are based on transitions, this raises the question to what extent the measurement error for UB II will impact fixed-effects estimates. Under the assumption of a classical measurement error model, measurement error would cause attenuation toward the zero in the model estimates.

Using information from register data as validation information for unemployment benefit-II-receipt, common assumptions about the form of the measurement error can be tested and joint distributions of the measurement error with model variables can be evaluated. Subsequently, panel models, analyzing the association between

unemployment benefit-II-receipt and subjective health, are recalculated with register information in order to assess the direction in which model estimates are biased by the measurement error for unemployment benefit-II-receipt. An attenuation of the estimates does not have to be necessarily the case, since the direction of the bias depends on model type and the covariance between the measurement error and all model variables (Bound, Brown, and Mathiowetz 2001, p. 3708). In a further step, approaches that can be easily implemented will be evaluated that might reduce the impact of the measurement error on the model estimates.

1.7.3 Errors in retrospective welfare reports and their effect on event history analysis

In the second study, the indicator for UB II is used as independent variable. In this third study, the indicator for UB-II-receipt is not used as independent variable but as dependent variable. However, for policy makers and labor market researchers individual transitions in and out the labor market states are of central importance. Thus, event history models are another popular class of statistical models to analyze panel data. Event history models require spells, a time span defined by two events, the time of beginning, the time of end.

The preferable way to collect information on events and their concrete date in surveys, are repeated measurements: "If change over time is of crucial interest, concurrent measures at different points in time are the only reliable way to assess it" (Schwarz 2007, p. 20–21). However, also panel surveys require retrospective recall of events, as the the event of interest can happen between the successive interviews. This is relevant, as many panel studies are conducted annually or even biannually.

The recall for such autobiographical events and spells is prone to a wide range of response errors (e.g. Cannell, Marquis, and Laurent (1977), Gray (1955), Paull (2002), and Thompson et al. (1996)). Respondents can omit, misdate, merge, misclassify or invent events or spells. These recall errors can cause a bias in the analyses (Pyy-Martikainen and Rendtel 2009).

Research in cognitive psychology (Roediger 2008) and in survey methodology (Belli, Bilgen, and Al-Baghal 2013; Tourangeau, Rips, and Rasinski 2000) has shown that the extent of such errors depend on factors like task characteristics, data collection modes, respondent abilities and individual response strategies. By validating the survey response for UB-II-receipt with entries from administrative records, we can identify the extent of measurement error for the reported UB-II-spells of the respondents. Thus, we can test, whether the factors related with the response process, influence the measurement error of the spells.

However, even if the recall is distorted by the aforementioned factors, the crucial point from a analytical point of view is: Do errors in autobiographical reporting bias the results of statistical analyses to such a degree that different conclusions would be drawn? Therefore, we compare time-to-event models based on administrative data with models based on respondent reports to assess the bias due to the measurement error.

1.7.4 Dependent interviewing and suboptimal responding

One way to reduce response errors in panel surveys is dependent interviewing. With Proactive Dependent Interviewing (PDI), respondents are reminded of the answer to a survey question they gave in a previous interview. The previous information is used to verify whether the respondent's status has changed, or as a starting point for asking about events since the previous interview. PDI is said to reduce spurious change between panel waves and increase overall data quality (Moore et al. 2009). However, concern is frequently voiced that measurement error from the previous wave will be carried forward into future waves of the survey caused by false confirmation of the preload by the respondent (Mathiowetz and McGonagle 2000). The false confirmation might be influenced by sub-optimal responding, as respondents are known to take cognitive shortcuts and are influenced by the social situation of the interview.

Prior to the interviews for wave 4 of PASS, the preload was faultily generated for a subgroup of 393 respondents regarding UB-II-receipt and respondents were given questions with false preload information. Only a part of the respondents contradicted the false preload. Thus, even if the rise of computers in survey research can facilitate and improve survey administration, new types of errors can arise.

However, by linking the survey response to individual administrative records on UB-II-receipt, the preload error allows us to exploit a rare research opportunity to address some questions regarding acquiesence to false preload. In this paper, following research questions are dealt with: (1) To what extent do respondents confirm previous information that is false? (2) To what extent is the false confirmation in fact caused by false reporting in the previous wave? (3) What factors explain the confirmation of the actual false preload? (4) To what extent is the false confirmation carried forward into the next wave of the survey?

2 Will respondents eventually get it right? Changes in measurement error in a panel survey using dependent interviewing: Results from a five-wave validation study

Johannes Eggs, Annette Jäckle and Mark Trappmann

Abstract

Measurement of state benefit receipt is typically affected by underreporting. In a panel survey there are however reasons to expect reporting quality to improve across waves: respondents who are more likely to misreport may also be more likely to attrit; over time respondents may gain trust in the survey and therefore underreport less; the stigma in the population associated with a newly introduced benefit type may fall over time, again reducing underreporting; and respondents may remember what they will be asked about in the interview and become better at remembering relevant information. In addition, panel surveys can make use of responses from previous interviews, to remind respondents of income sources they have received in the past. In this study we use data from five waves of the panel study "Labour Market and Social Security" (PASS), linked to individual administrative records on unemployment benefit receipt, to examine (1) whether data accuracy changes over waves of a panel survey, (2) which mechanisms lead to improved reporting, and (3) whether changes in data accuracy alter substantive research conclusions. The results show significant reductions in measurement error across waves. Part of this trend is due to dependent interviewing: in each wave a larger proportion of recipients receive the dependent question and are helped by the cues it provides. Part of the trend is related to selective attrition. However, whether the improved data quality decreases the bias for substantive research remains unclear.

Keywords: record linkage, welfare receipt, unemployment benefit, validation study

2.1 Introduction

The measurement of income from state benefit programmes is severely affected by underreporting. Evidence from validation studies comparing survey responses to individual administrative records suggests that, depending on the benefit type, up to 50% of recipients underreport in surveys, while overreporting is rare (Kreuter, Müller, and Trappmann 2010; Lynn et al. 2012). In panel surveys the reporting of state benefit receipt may however improve over time: First, respondents who are more likely to misreport might also be more likely to attrit from the panel (selection effects, see Bollinger and David 2001); second, over time respondents may gain

trust in the survey, and therefore feel less need to deliberately underreport; third, the stigma in the population associated with a newly introduced type of benefit may fall over time, again making it less likely that respondents deliberately underreport; fourth, respondents may remember what they will be asked about in the survey, and over time may be more likely to remember relevant details (panel conditioning effects on reporting behaviour, see Frick et al. 2006; Rendtel et al. 2004).

In addition, panel surveys offer unique opportunities for preventing measurement error, since the questionnaire script can make use of answers given in previous interviews. With proactive dependent interviewing, respondents are reminded of their previous answers, either to ask whether their status has changed, or as an anchor for asking about events since the previous interview (Mathiowetz and McGonagle 2000). Dependent interviewing is now used in most panel surveys to improve the measurement of key economic information. Proactive dependent interviewing is most commonly used to collect information about labour market activities (e.g. in the UKHLS, CPS, NLSY97, HRS, ELSA, SLID), income sources and pension plans (e.g. UKHLS, SIPP, HRS), assets (SIPP), but also for partnership histories (e.g. UKHLS, SIPP, NLSY97, Pairfam), citizenship (SIPP), or health behaviours and conditions (ELSA).[1]

A previous validation study contrasting dependent interviewing and traditional independent interviewing has shown that dependent questions reduce underreporting of benefit receipt (Lynn et al. 2012). When dependent interviewing is used in just one wave, as in Lynn et al. which is the only validation study so far, errors are however not eliminated. This is in part because only those respondents who have reported receipt in the past can be reminded of their previous receipt (see the discussion by Lynn et al. 2006). That is, dependent interviewing only helps some respondents. However, since dependent interviewing decreases underreporting, over time, an increasing proportion of recipients should receive the dependent question and benefit from the cue it provides. Dependent interviewing may therefore decrease underreporting to a larger extent when it is used over multiple waves of a panel survey. We use five waves of the panel study "Labour Market and Social Security" (PASS), linked to individual administrative records, to examine how measurement error in the reporting of unemployment benefit receipt changes over waves of a panel survey. In addition to examining measurement error in a population sample, as previous studies have done, we examine the special case of a sample of benefit recipients. We address the following questions:

1 UKHLS = UK Household Longitudinal Study, CPS = Current Population Survey, NLSY97 = National Longitudinal Survey of Youth 1997, HRS = Health and Retirement Study, ELSA = English Longitudinal Study of Ageing, SLID = Survey of Labour and Income Dynamics, SIPP = Survey of Income and Program Participation, Pairfam = Panel Analysis of Intimate Relationships and Family Dynamics.

1. Does data accuracy change over waves of a panel survey?
2. Why does data accuracy change over waves of a panel survey?
3. Do changes in data accuracy alter substantive research conclusions?

2.2 The panel survey and validation data

2.2.1 Survey design

PASS is a household panel survey that was started in 2007, to provide data for research on unemployment and poverty dynamics in Germany.[2] The main focus is on recipients of unemployment benefit II (UB II), a means tested benefit that had recently been introduced. The survey combines two samples: a sample drawn from administrative records of UB-II-recipients and an address-based sample of households. The recipient sample was drawn from administrative data held by the German Federal Employment Agency. 300 primary sampling units (PSUs) were drawn from postcodes with selection probabilities depending proportionally on the size of the population. Within each PSU, benefit units were drawn. The population sample was based on a commercial database of household addresses, where addresses were sampled within PSUs. The population sample was stratified disproportionately by socio-economic status such that households with low status were oversampled. Sample members are interviewed annually and we use the first five waves (2006–2011) of the data.

Households are first approached in CATI and non-respondents and households for whom no valid telephone numbers are known are followed up with CAPI. From wave 2 onwards households are first approached in the mode in which they were last interviewed. In each household, first an interview with the household target person is sought, followed by individual interviews with each member of the household aged 15+. In the recipient sample the household target person is the person registered with the agencies responsible for the provision of UB II. In the population sample the target person is the person who is most knowledgeable about household related matters.

In wave 1 PASS had a household response rate of 28.7 % for the recipient sample (23 736 households issued) and 24.7 % percent for the population sample (25 316 households issued) (RR1, according to AAPOR 2009). Row 1 in table 2.1 shows the development of both samples over the first five waves.

2 For a comprehensive overview of the PASS panel, see Trappmann et al. (2013). For detailed information, see the PASS User Guide (Bethmann, Fuchs, and Wurdack 2013). Field and methods reports as well as codebooks and data documentations for each wave can be downloaded from the website of the Research Data Center of the Federal Employment Agency at the Institute for Employment Research (http://fdz.iab.de/en/FDZ_Individual_Data/PASS/Working_Tools.aspx).

PASS uses dependent interviewing for several items, including benefit receipt, occupation, education, and household composition. We focus on UB-II-receipt as it can be validated against administrative records. Since UB II is paid to benefit units (defined as single persons or couples, with their dependent children) rather than individuals, the survey collects this information in the household questionnaire. The questions collect data in a spell format at monthly level. The independent version that is used for households who did not report receipt at the previous wave interview date reads:

> *Now we are only interested in unemployment benefit 2 ("Arbeitslosengeld 2") also known as Hartz 4. Thinking of the time since the last interview in ‹MONTH/YEAR›: What about you? Has your household obtained unemployment benefit 2 ("Arbeitslosengeld 2") at any time since ‹MONTH/YEAR›?*

A follow-up question collects the start and end date of the first spell of receipt in that period. The proactive dependent interviewing version is used for households who reported UB-II-receipt at the previous interview date:[3]

> *In the last interview in ‹MONTH/YEAR› you stated that the household you were living in then was obtaining unemployment benefit 2 ("Arbeitslosengeld 2") at the time. Until when was this benefit obtained without interruption? Please report the month and the year.*

Respondents can contradict the preloaded data (fewer than 0.5 % do). In this case the spell is considered to have ended in the month of the previous interview.

2.2.2 Administrative data and linkage

The administrative data used to validate survey reports are from the Integrated Employment Biographies (IEB). This dataset integrates and consolidates data from different sources, including data taken from social security notifications, from the administration of unemployment benefits, and from job applicant pools.

For this article we use the exact start and end dates of all spells of UB-II-receipt. This information is of high quality as it is directly produced by the software that administers benefit claims and payments (Jacobebbinghaus and Seth 2007; Köhler and Thomsen 2009).[4] The IEB is a person level dataset. Spells that refer to a benefit unit are therefore recorded for each person in that unit.

3 Households who were not interviewed in the previous wave are reminded of any receipt reported two years earlier. No interviews are sought with households who are not interviewed in two consecutive years.

4 The administrative data are incomplete prior to November 2005 (Geschäftsbereich ITM 2009). Hence, we use only data generated after November 2005 for validation purposes.

The linkage between PASS survey data and IEB administrative records requires informed consent of respondents. For the purposes of our analyses, linkage was only carried out if the person who completed the household interview gave consent. For the population sample respondents who consented were linked to the administrative records using their name and address, gender and date of birth. This was done using error tolerant procedures based on Jaro (1989).[5] For most cases in the recipient sample linkage was trivial, since the sample had been drawn from one of the IEB data sources. Only those members of households who had not been a member of the benefit unit at the sampling date had to be linked using the procedure used for the population sample.

2.2.3 Analysis sample

We exclude the following households from the analyses. The case numbers and resulting sample sizes for the population and recipient samples are documented in table 2.1:

- Refresher samples added after wave 1 – these are already excluded from row 1 in table 2.1.
- Households that were not interviewed in the previous wave – since we expect the measurement error implications to be different if the preload data are from two years rather than one year earlier. We do however include households that were non-respondents in any of the earlier waves.
- Households which do not contain anyone eligible for UB II – where no adult is below age 65 according to the survey data.
- Households which contain two or more benefit units (for example adult children living with their parents) – the household questionnaire asks whether anyone in the household received UB II, however the linked administrative records only cover receipt of the respondent's benefit unit. Depending on who answers the household questionnaire, it is therefore possible that the survey receipt status does not match the record receipt status.
- Households where the adult composition has changed since the previous interview – since a respondent who has just joined the household could potentially be led to report on receipt of their previous household (the dependent interviewing question reads "the household you were living in then was receiving ..."). We do however include households whose composition changed at any time before the last interview. We also include households where the person completing the

5 The linkage is documented in Antoni and Bethmann (2014) and for the linkage procedure the Merge Tool Box (MTB) software was used (Schnell, Bachteler, and Reiher 2005).

household questionnaire changes between waves. This is rare (for the combined population and recipient sample N = 2 at wave 2, N = 45 at wave 3, N = 112 at wave 4, N = 102 at wave 5). Since the respondent is reporting household level receipt, rather than acting as a proxy for individual receipt, we expect the quality of reports to be similar across respondents within a household.

- Households where the person completing the household interview did not consent to linkage in any of the first three waves of PASS.
- Households for whom the linkage failed – a person may not be found in the records if there are errors in the linkage variables, or if they are genuinely not included in the administrative datasets. The latter is the case for all persons who have never been unemployed, received UB II, officially searching for a job, or in employment subject to social insurance contributions (i.e. they have been students, housewives/men, self-employed or federal employees all their life).
- For wave 4 we further exclude 322 households for whom the preload data about UB-II-receipt contained an error, see Eggs and Jäckle (2015).

Table 2.1: Sample sizes

Population sample†	W1		W2		W3		W4		W5	
	N	%	N	%	N	%	N	%	N	%
HH responding	5990	100	3870	100	3813	100	2817‡	100	2524	100
• HH non-response t_1	0	0	0	0	585	15.3	102	3.6	79	3.1
• HH with 0 or 2+ BUs	1297	21.7	933	24.1	844	22.1	762	27.1	786	31.1
• HH comp. change	0	0	78	2	65	1.7	89	3.2	56	2.2
• HH no consent	848	14.2	296	7.6	208	5.5	159	5.6	133	5.3
• HH no linkage	666	11.1	447	11.6	350	9.2	290	10.3	279	11.1
Analysis sample	3179	53.1	2116	54.7	1761	46.2	1415	50.2	1266	50.2
HH current receipt in										
Survey	397	12.5	241	11.4	187	10.6	124	8.8	104	8.2
Records	451	14.2	262	12.4	186	10.6	130	9.2	96	7.6
Recipient sample†	**W1**		**W2**		**W3**		**W4**		**W5**	
	N	%	N	%	N	%	N	%	N	%
HH responding	6804	100	3472	100	3668	100	2418~	100	2044	100
• HH non-response t_1	0	0	0	0	982	26.8	141	5.8	131	6.4
• HH with 0 or 2+ BUs	568	8.3	307	8.8	265	7.2	242	10	320	15.7
• HH comp. change	0	0	125	3.6	86	2.3	117	4.8	86	4.2
• HH no consent	909	13.4	197	5.7	131	3.6	93	3.8	79	3.9
• HH no linkage	116	1.7	85	2.4	69	1.9	72	3	64	3.1
Analysis sample	5211	76.6	2758	79.4	2135	58.2	1753	72.5	1500	73.4
HH current receipt in										
Survey	3 951	75.8	1 941	70.4	1 391	65.2	1 144	65.3	879	58.6
Records	4 374	83.9	2 073	75.2	1 436	67.3	1 160	66.2	893	59.5

Notes: HH = household. BUs = Benefit Units, comp. = composition. † Excludes refresher samples. ‡ Excludes N = 42 households with preload error at W4. ~ Excludes N = 280 with preload error.
Percentages in rows 2 to 7 are based on the number of households responding in that wave; in the last two rows percentages are based on the number of households in the analysis sample.

2.3 Results

In examining how data accuracy changes across waves we treat the administrative records as the truth, and any deviations in the survey data as measurement error.

2.3.1 Does data accuracy change over waves of a panel survey

We firstly examine errors in survey responses, by testing for changes in the extent of under- and overreporting across waves. We then examine the accuracy of estimates derived from the survey responses, including the estimated stock of recipients, the duration of receipt, and inflows and outflows.

For each wave, we classify households according to their receipt status at the date of interview in the survey and the records. These classifications are used to calculate the rate of false negatives (the proportion of record recipients underreporting in the survey) and of false positives (the proportion of record non-recipients overreporting in the survey).

Figure 2.1 shows the error rates separately for the population and recipient samples. The false negative rate falls from wave 1 to wave 5: in the population sample from 17.5% to 7.2%, in the recipient sample from 12.6% to 7.3%. The false positive rate in the population sample is constant at around 1%. In the recipient sample it is higher, at 14.8% in wave 1, about 9% in waves 2 and 3, 11.6% in wave 4, and 8.4% in wave 5. In the recipient sample all households have, by definition been recipients at an earlier point in time and the higher rate of overreporting may therefore not be surprising.

To test whether the trends in false negative rates in figure 1 are significant, we estimate logit models of the probability of a household underreporting receipt, using only record recipients as the analysis sample and adjusting for clustering of observations in households (table 2.2). The results for Model 1 indicate that the probability of underreporting decreases by around 2% with every additional wave, and that this trend is significant for both the general population (average marginal effect (AME) = -0.027, S.E. = 0.009) and recipient samples (AME = -0.017, S.E. = 0.003). For false positives there is no apparent trend for the population sample in figure 1, and this is confirmed by the test in table 2.3 (population sample Model 1). For the recipient sample figure 1 suggests a slight downward trend in the false positive rate, which is confirmed by the test in table 2.3 (recipient sample Model 1, AME = -0.012, S.E. = 0.004).

Figure 2.1: Under- and overreporting rates in receipt at date of interview

Table 2.2: Significance tests for trend in false negatives

	Population sample				Recipient sample			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)
Wave	-0.027**	-0.020*	0.002	0.008	-0.017***	-0.013***	-0.016***	-0.011**
	(0.009)	(0.009)	(0.017)	(0.017)	(0.003)	(0.003)	(0.003)	(0.003)
Not in all waves		0.066*		0.063*		0.030***		0.031***
		(0.028)		(0.028)		(0.009)		(0.009)
CAPI			0.112*	0.106*			0.012	0.013
			(0.054)	(0.053)			(0.013)	(0.013)
CAPI*Panel wave			-0.041*	-0.038			-0.004	-0.005
			(0.020)	(0.020)			(0.006)	(0.006)
Psd. R^2	0.0117	0.0197	0.0179	0.0252	0.0000	0.0106	0.0079	0.0108
N	1 125	1 125	1 125	1 125	9 936	9 936	9 936	9 936

Notes: Average marginal effects (AME) and standard errors from logit models, adjusted for clustering of observations in households. Reference category: households in all waves (balanced panel) and interviewed in CATI. Analysis samples: households with record receipt. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.3: Significance tests for trend in false positives

	Population sample				Recipient sample			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)	AME (se)
Wave	0.001	0.001	0.000	0.001	-0.012***	-0.009*	-0.016**	-0.015*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)	(0.005)	(0.006)
Not in all waves		0.009**		0.007*		0.022		0.015
		(0.003)		(0.003)		(0.016)		(0.016)
CAPI			0.007	0.007			0.082**	0.081**
			(0.005)	(0.005)			(0.027)	(0.027)
CAPI*Panel wave			0.001	0.001			0.005	0.005
			(0.002)	(0.002)			(0.008)	(0.008)
Psd. R^2	0.0005	0.0145	0.0280	0.0368	0.0040	0.0055	0.0370	0.0378
N	8 610	8 610	8 610	8 610	3 421	3 421	3 421	3 421

Notes: Average marginal effects (AME) and standard errors from logit models, adjusted for clustering of observations in households. Reference category: households in all waves (balanced panel) and interviewed in CATI. Analysis samples: households with record receipt. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As the accuracy of survey responses improves from wave to wave, with respondents becoming less likely to underreport receipt, the accuracy of estimates derived from the responses also improves. Figure 2.2 shows the estimated stock of UB-II-recipients (or prevalence rate), based on the survey responses and the record data. The prevalence rates according to the records are shown as a dot for the register sample and as a triangle for the population sample. The prevalence rates estimated from the survey data also show the 95% confidence intervals. For the register sample the proportion of the sample receiving UB-II-benefits is underestimated compared to the records in waves 1 and 2. From wave 3 onwards the survey estimate overlaps with the estimate using record data. In the population sample the survey estimates overlap with the record estimates in all waves. The reduction of underreporting also improves estimates of the duration of receipt. As a measure of time spent in receipt we calculate the number of months of receipt for each household and wave, according to the records and the survey. For households where information about the dates of receipt was missing in the survey, but who had reported on-going receipt in the previous, current and subsequent wave, we impute the number of months of receipt as the months between interviews. All other households with missing date information are excluded from the analysis, to avoid confounding the evaluation of reporting errors with imputation methods (number of households excluded in W1:275, W2:63, W3:28, W4:18, W5:7). Figure 2.3 shows the proportion of the total months of receipt in the record data, which are reported in the survey. In both the population and the recipient samples the wave 1 survey captures only about 80% of the aggregate months of receipt in the record data. The coverage is slightly higher when the analysis

Figure 2.2: Prevalence rates based on the survey and record data, by sample

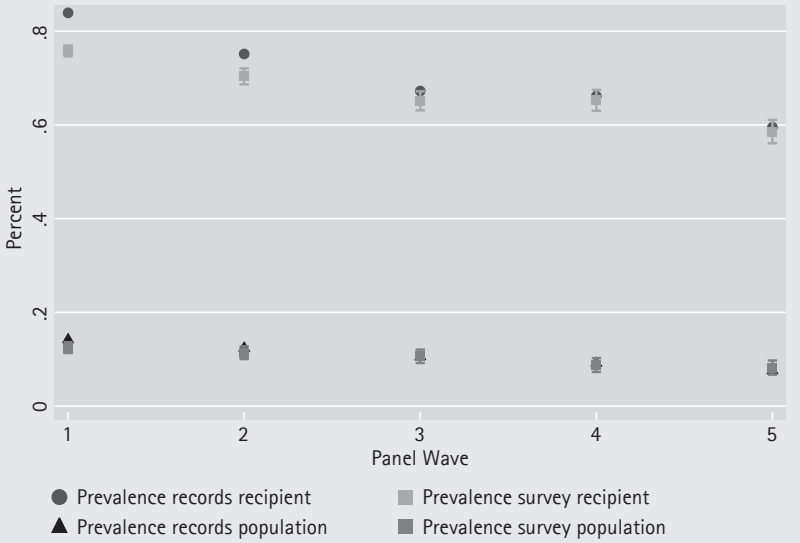
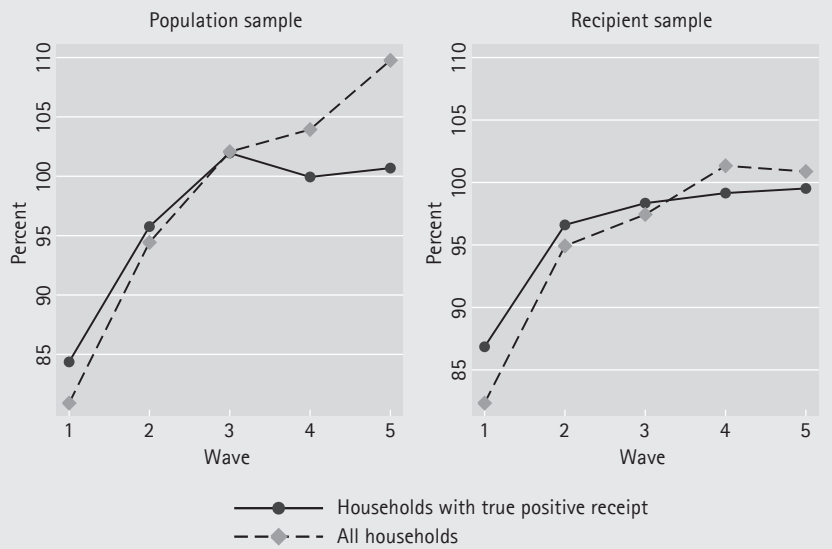


Figure 2.3: Proportion of months of receipt in the records reported in the survey



is restricted to households who according to the record data were recipients (84 % for the population sample, 86 % for the recipient sample). The underreporting of months of receipt can be a combination of underreporting any receipt (equal to reporting zero months) and misreporting the dates of receipt. Therefore, as the underreporting

of receipt falls across waves, so does the underreporting of months of receipt. From wave 3 onwards recipient households report around 100% of months in both the recipient and the population samples. In the full sample, including recipients and non-recipients, false positive reports lead to an overreporting of months of receipt in wave 4 and 5 in both samples.

To test whether the observed trend in the error in months of receipt is statistically significant, we focus on the sample of households who at any given wave are recipients according to the records. The corresponding line in figure 2.3 shows the ratio of the sum of months over all respondents in the survey, as a fraction of the sum of months in the records. This ratio can be expressed as

$$\frac{\sum_i s_i}{\sum_i r_i} = \frac{1}{\sum_i r_i} \cdot \sum_i r_i \cdot \frac{s_i}{r_i} \tag{2.1}$$

where $\frac{1}{\sum_i r_i}$ is a constant, $\frac{s_i}{r_i}$ is the ratio of survey over record months for each individual household i and $\sum_i r_i$ is a factor by which the ratio is weighted. We therefore test, whether the household level ratio $\frac{s_i}{r_i}$ significantly increases across waves by estimating an OLS regression of

$$\frac{s_i}{r_i} = \beta_0 \cdot \beta_{1wave} + \varepsilon_i, \; weighted\ by \sum_i r_i \tag{2.2}$$

The results in table 2.4 (Model 1) suggest that the trend of increasing coverage of months of receipt in the survey, that is observed in figure 2.3, is indeed significant in the population sample ($\beta_{wave} = 0.021$, S.E. = 0.008) and the recipient sample ($\beta_{wave} = 0.025$, S.E. = 0.002).

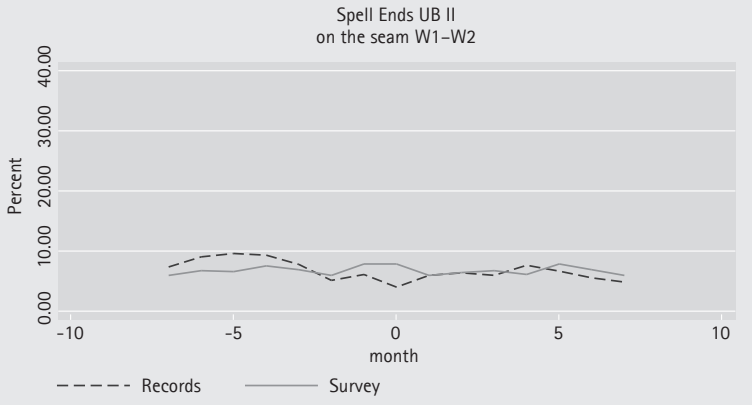
Table 2.4: Significance tests of trend in errors in months of receipt

	Population sample				Recipient sample			
OLS	(1) b (se)	(2) b (se)	(3) b (se)	(4) b (se)	(1) b (se)	(2) b (se)	(3) b (se)	(4) b (se)
Wave	0.021**	0.023*	0.002	0.003	0.025***	0.025***	0.028***	0.027***
	0.008	0.009	0.013	0.013	0.002	0.002	0.002	0.002
Not in all waves		0.013		0.015		-0.004		-0.005
		0.020		0.021		0.006		0.006
CAPI			-0.107**	-0.108**			0.045***	0.045***
			0.041	0.041			0.011	0.011
CAPI*Panel wave			0.026	0.026			0.010*	-0.009*
			0.016	0.016			0.004	0.004
Constant	0.914***	0.902***	0.994***	0.982***	0.893***	0.897***	0.881***	0.886***
	0.019	0.030	0.034	0.040	0.005	0.007	0.006	0.008
N	13 293	13 293	13 293	13 293	127 985	127 985	127 985	127 985
Adj. R^2	0.006	0.006	0.011	0.011	0.014	0.014	0.016	0.016

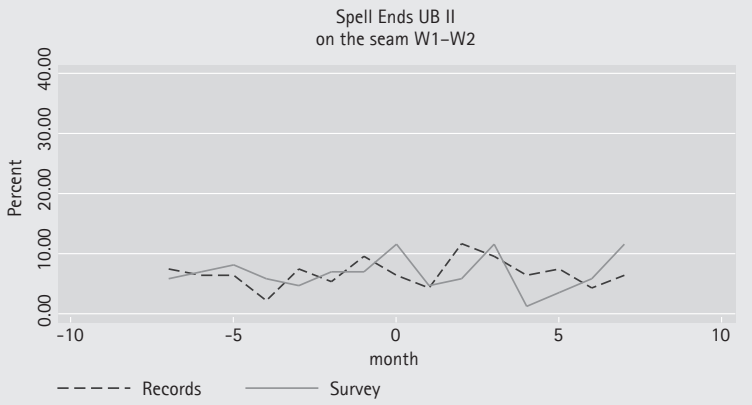
Notes: * p < 0.05, ** p < 0.01, *** p < 0.001

Finally, we examine the accuracy of estimated inflows and outflows. Figure 2.4 shows the monthly transition rates out of receipt, for the period reported on in waves 1 and 2 of the survey. For each month the graph shows the proportion of recipients who stopped receiving UB II by the next month. The x-axis is centred on the month of the wave 1 interview (month 0), which can correspond to different calendar months as fieldwork lasts for seven months. The month of the wave 1 interview constitutes the "seam" between reference periods: in the wave 1 interview respondents report on the months prior to month 0, in the wave 2 interview respondents report on the months since the wave 1 interview. The graph suggests that the transition rates estimated from the survey data are similar to the rates in the record data, for both the recipient and the population samples.

Figure 2.4: Monthly outflows, waves 1–2

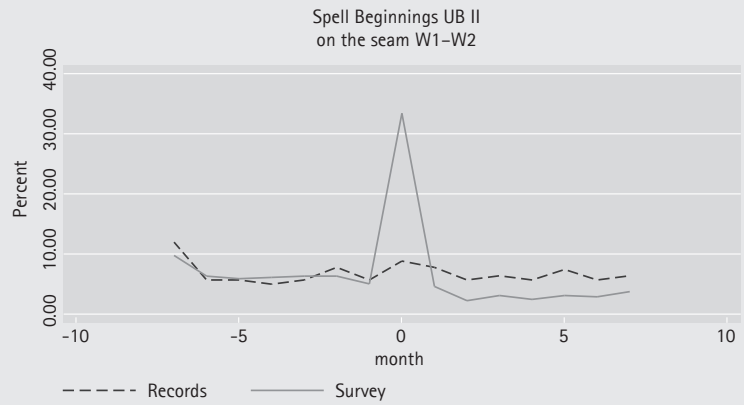


(a) Recipient sample

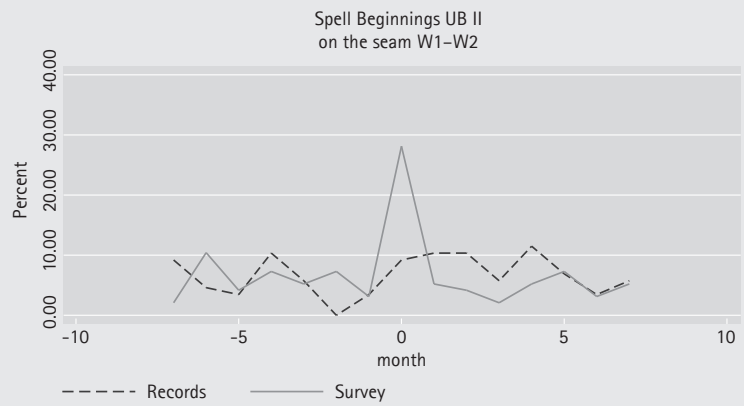


(b) Population sample

Figure 2.5: Monthly inflows, waves 1–2



(a) Recipient sample



(b) Population sample

Figure 2.5 similarly shows the rates of inflows into receipt. For each month the graph shows the proportion of recipients who started receiving UB II by the next month. In the recipient sample transition rates in the record data are relatively constant across months. However in the survey data there is a large spike in month 0. The transition rate at the seam between reference periods from wave 1 and wave 2 is 33%, compared to rates of around 3 to 8% in off-seam months. That is, there is a large proportion of respondents who in the wave 1 interview report no receipt of UB II, who in the wave 2 interview report a receipt of UB II in the month following the wave 1 interview. This concentration of transitions at the seam between reference periods (seam effect) is the result of respondent errors in the recall and the dating of receipt and survey processing (Callegaro 2008). This seam effect is similar in the population sample, although there are more fluctuations due to the smaller numbers of households entering receipt.

The patterns of estimated inflows and outflows are similar for the other waves (not shown). The results in table 2.5 confirm that there are no significant changes in the extent of seam effects across waves. The table shows results from logit models of the probability of a change occurring in a seam month, versus an offseam month, pooling all transitions covered by the five waves of the panel. The probability of a transition occurring in a seam month is no different for any of the wave pairs compared to the wave 1–2 seam, in either the population or the recipient sample. The only significant effect is that the probability of a transition occurring at the seam is lower in the wave 3–4 seam than the wave 1–2 seam.

2.3.2 Why does data accuracy change over waves of a panel survey?

As described in the introduction there are several alternative mechanisms that could lead to a reduction of measurement error in the reporting of benefit receipt over time. In this section we test for evidence of these mechanisms.

Table 2.5: Probability of a transition in receipt status being in a seam month

		Population sample		Recipient sample	
		Start date	End date	Start date	End date
		AME (se)	AME (se)	AME (se)	AME (se)
Survey			*Omitted: W1–2*		
	Waves 2–3	0.007	0.033	0.052	0.110
		(0.040)	(0.019)	(0.080)	(0.063)
	Waves 3–4	-0.165***	0.046	0.000	0.072
		(0.043)	(0.028)	(0.098)	(0.075)
	Waves 4–5	-0.041	0.019	0.033	0.060
		(0.047)	(0.023)	(0.117)	(0.076)
	Psd. R^2	0.0118	0.0051	0.0020	0.0159
	N	872	1424	195	220
Records			*Omitted: W1–2*		
	Waves 2–3	-0.009	0.020	-0.031	-0.018
		(0.026)	(0.013)	(0.049)	(0.036)
	Waves 3–4	-0.029	0.022	0.007	0.022
		(0.026)	(0.019)	(0.059)	(0.051)
	Waves 4–5	-0.056*	0.024	-0.073	0.113
		(0.025)	(0.019)	(0.046)	(0.067)
	Psd. R^2	0.0110	0.0054	0.0191	0.0383
	N	740	1517	209	257

Notes: Average marginal effects (AME) and standard errors from logit models of probability of spell starting or ending at an interview month (seam). Standard errors adjusted for clustering in households.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(1) Selection effects

The improvements in data accuracy may be due to selective attrition, whereby respondents who are more likely to misreport are also more likely to drop out of the panel. To test whether selective attrition explains the improvement in data accuracy across waves, we condition the analyses on whether or not the household was interviewed in all waves. Model 2 in table 2.2 suggests that households that are not interviewed in all waves are indeed more likely to underreport receipt (AME = 0.066, S.E. = 0.028 for the population sample and AME = 0.030, S.E. = 0.009 for the recipient sample). The time trend is slightly weaker (AME = -0.020 compared to -0.027 in Model 1 for the population sample, and -0.013 compared to -0.017 in Model 1 in the recipient sample), but remains significant. The results are similar when testing for the trend in overreporting of receipt (Model 2 in table 2.3). In contrast selective attrition does not appear to affect the reporting of months of receipt (Model 2 in table 2.4). We conclude that selective attrition explains some, but not all, of the improvement in data accuracy over time.

(2) Increased respondent trust in the survey

Improvements in data accuracy may also be caused by changes in respondents' trust in the survey. If respondents trust increases through their experience with the interviews, they may feel less need to deliberately underreport receipt of a benefit that could be stigmatizing. It is not a priori clear whether respondents would be more or less willing to disclose sensitive information in a face-to-face than telephone interview (see the review by Tourangeau and Yan 2007). In either case we would however expect the development of trust to be stronger with personal interviewing than telephone interviewing, and would therefore expect the reduction of underreporting across waves to be stronger with CAPI than CATI. To test whether increased trust affects the accuracy of reporting, we include an indicator for whether the household was interviewed in CAPI as opposed to CATI, and an interaction with wave. The results are somewhat mixed. The expected effect that reporting accuracy improves more across waves for CAPI than CATI respondents is found for the likelihood of underreporting in the population sample (AME of the interaction of CAPI and wave = -0.041, S.E. = 0.020, Model 3 in table 2.2). The time trend in the reporting accuracy becomes insignificant, suggesting that accuracy improvements are only in the CAPI sample. However in the other models testing for underreporting in the recipient sample, and for overreporting in both samples, the interaction of CAPI and wave is not significant and the estimated time trend in reporting accuracy is not affected. Testing for errors in the months of receipt shows similar results. In the recipient sample the interaction of CAPI and wave is again significant, with CAPI accuracy improving more than CATI across waves

(AME = -0.010, S.E. = 0.004, Model 3 in table 2.4). The trend in reporting accuracy however remains similar to Models 1 and 2. In the population sample the main effect of CAPI is significant, however the interaction of CAPI and wave is not, and the trend in reporting accuracy becomes insignificant. We conclude that there are some differences between CATI and CAPI respondents, which are consistent with the hypothesis that increased respondent trust in the survey could lead to improved reporting. The results remain unchanged, when controlling for attrition (Models 4 in tables 2.2, 2.3, 2.4). Thus, the increase in reporting accuracy seems to be also based on increased trust. The effects are however small.

(3) Reduced stigma in the perception of the population

Improvements in the accuracy of reporting receipt could also be caused by changes over time in the social stigma associated with UB II. At the start of the PASS survey UB II had only recently been introduced. It is possible that the population's perception of people receiving this benefit changed. If social stigma decreased over time, respondents may feel less need to deliberately underreport receipt. To test whether improvements in reporting accuracy could be due to changes in attitudes of the general population, we examined the error rates in the refreshment samples that were added at each wave of PASS. To rule out any effects due to prior experiences with the survey, we analyse households the first time they were interviewed, and check for differences in underreporting rates between the yearly refreshment samples. All refreshment samples were drawn from the administrative records of UB-II-recipients, using the same methodology as for the wave 1 recipient sample. If declining social stigma had any effect, we would expect households added in later waves to be less likely to underreport, than households interviewed in the earlier waves. Underreporting rates fluctuated slightly between 13 % and 15 % from waves 2 to 5, however there were no significant differences between waves ($\chi^2 = 1.524$, $p = 0.677$). We conclude that trends in the stigma associated with UB-II-receipt are unlikely to have caused the observed improvement in reporting accuracy.

(4) Panel conditioning effects on reporting behaviour

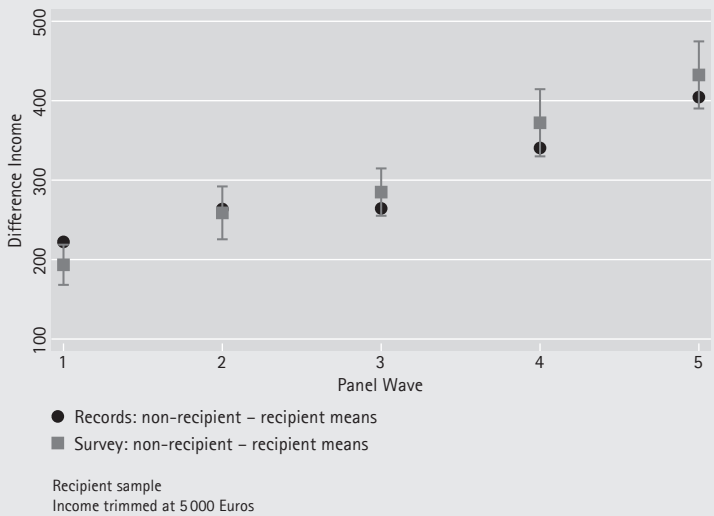
Improvements in reporting accuracy could also be caused by panel conditioning. Having experienced several interviews respondents are likely to remember that they will be asked to report on welfare receipt and over time may be increasingly likely to know the details they will be asked to report. The design of the survey unfortunately does not allow us to test for this mechanism. We return to this issue in the discussion.

2.3.3 Do changes in data accuracy alter substantive research conclusions?

One of the key research questions the PASS survey was designed for is to assess the material and social wellbeing of households in receipt of UB II (Trappmann et al. 2013). Previous research on underreporting of UB-II-receipt has shown that underreporters have characteristics that are similar to households that are not recipients (Bruckmeier, Müller, and Riphahn 2014). If there is a sizeable group of underreporters whose characteristics are similar to the characteristics of true non-recipients, for example in terms of their income or labour market attachment, then misclassifying them as non-recipients may exaggerate the apparent difference between recipient and non-recipient households.

We examine three indicators of wellbeing in which recipients and non-recipient households may differ: household income, material deprivation, and health status. For each indicator we calculate the mean or proportion for (1) recipients according to the records, (2) non-recipients according to the records, (3) recipients according to the survey, and (4) non-respondents according to the survey. We then examine whether reporting errors that cause households to be misclassified as non-recipients lead to an exaggeration of the differences in wellbeing between recipient and non-recipient households.[6]

Figure 2.6: Gaps for monthly household income between recipient and non-recipient households, according to survey and records



6 Similar effects are found in research on overreporting of voting: respondents who falsely report aving voted have characteristics that are similar to those who do vote. Misclassifying non-voters as voters exacerbates differences in characteristics of voters and non-voters (Ansolabehere and Hersh 2012).

Figure 2.6 shows the difference in means of monthly net equivalized household income, between recipient and non-recipient households. Non-recipient households are on average between € 200 and € 400 better off than recipient households. The estimated difference based on the survey data includes 95% confidence intervals. The results suggest that conclusions about the differences between recipients and non-recipients are not affected by classification errors, as the survey estimates overlap with estimates using the records to classify households.

Figure 2.7 shows the difference in means of a material deprivation index between recipient and non-recipient households. The index is a summed score of 23 items covering social activities and goods that the household cannot afford (see the appendix for the questions). A higher score on the index implies higher material deprivation. The graph shows the difference in mean scores of non-recipient and recipient households. For waves 1 and 2 the survey estimates overlap with the record estimates: non-recipient households can on average afford 2 to 2.5 fewer items than non-recipient households. In waves 3, 4 and 5 however the survey overestimates the difference in material deprivation between recipient and non-recipient households. This is unexpected. Given the improvements in data accuracy across waves, we would expect estimates of differences between recipient and non-recipient household to improve, not worsen, across waves.

Figure 2.7: Gaps for the material deprivation index between recipient and non-recipient households, according to survey and records

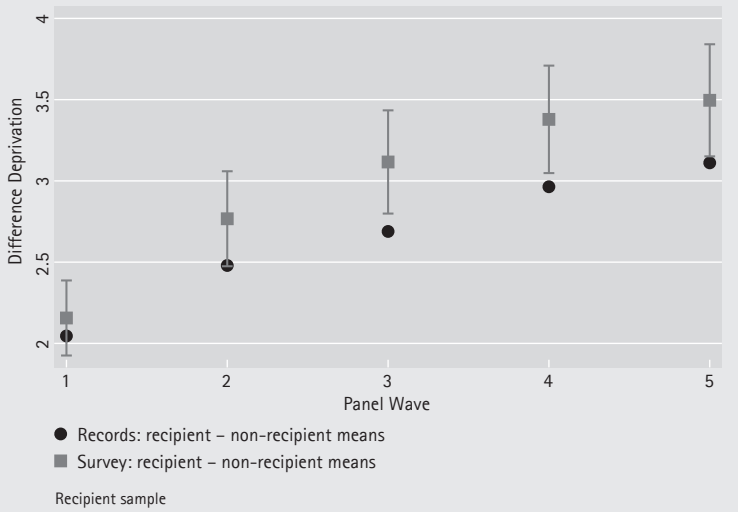


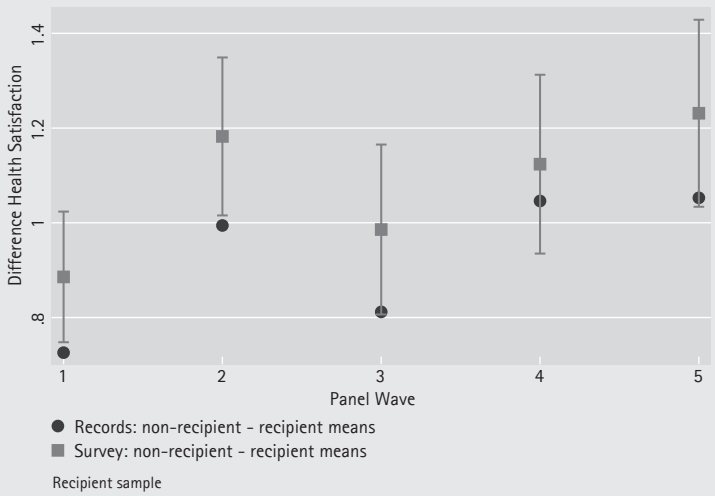
Figure 2.8 shows the difference for the self-reported health between household heads in recipient and non-recipient households, based on the question: *"On a scale from 0 to 10 how satisfied are you with your health?"* In waves 1 and 2 the survey data overestimate the differences between respondents in recipient and non-recipient households. In waves 3, 4 and 5 the survey estimates overlap with the record estimates. This improvement of estimates is consistent with the improvement in data accuracy across waves.

In sum, whether the observed improvements in data accuracy matter for substantive research conclusions, is unclear. For health we find the expected improvement of estimates across waves. However, for income we find no effects of misclassification errors on estimates and for material deprivation the survey estimates are worse in later waves.

2.4 Discussion

Using a unique source of validating data for five waves of a household panel survey, we examine how different aspects of measurement error in the reporting of state benefit receipt change across waves. We find that underreporting rates fall with each additional wave, overreporting does not increase, and therefore biases in estimated prevalence of receipt fall over time. Similarly, errors in the reported time spent in receipt are significantly reduced over time, and in aggregate close to zero by the fifth wave.

Figure 2.8: Gaps for the health indicator between recipient and non-recipient head of households, according to survey and records



The observed trends are partly related to selective attrition: respondents who are more likely to report with error are also more likely to drop out of the panel. This replicates previous findings by Bollinger and David (2001). We do not find support for the hypotheses that the reduction in measurement error could be due to increased trust of respondents in the survey, or a general decrease in the stigma associated with UB-II-receipt, or changes in reporting behaviour across time. Our results do not mirror findings by Bollinger and David (2005) who found a stronger persistence of underreporting of state benefit receipt in the first two waves of the 1984 SIPP.

The results are encouraging, especially since researchers frequently voice concern that proactive dependent interviewing can cause measurement error to be carried forward into future waves of the survey, and may lead some respondents to falsely confirm a previous status as still applying (Eggs and Jäckle 2015). The results may however be specific to the type of item studied, the average length of spells of receipt relative to the interval between interviews, and the specific wording of the dependent interviewing question used. Further validation studies of the effects of dependent interviewing over time would therefore be of value.

Seam effects are a common phenomenon, found in all panel surveys that collect continuous information about the timing and duration of spells, such as the duration of employment spells (Jäckle and Lynn 2007) or the duration of benefit receipt spells (Callegaro 2008). Although the extent of seam effects in our data does not change across waves, what is striking is that seam effects are only present in the transitions into receipt, not in the transitions out of receipt. That is, spell end dates seem to be reported well, spell start dates not so well. This difference in errors in the dates of events could be due to the asymmetry of the dependent interviewing question. Without dependent interviewing respondents are likely to report their current status as applying to all months of the reference period. Such "constant wave response" (Jäckle 2009) means that respondents who are currently not receiving a benefit are likely to underreport receipt that ended during the early months of the reference period. In this case the end date is erroneously shifted back to the date of the previous interview. Similarly, respondents who are currently receiving a benefit are likely to report receipt for all months of the reference period, even if it actually started after the previous month of interview. In this case the start date is erroneously shifted back to the previous month of interview. In PASS, respondents who reported receipt at the previous interview are reminded of their previous receipt and asked until when the receipt continued. This prevents respondents from forgetting about receipt at the start of the reference period, and thereby prevents the start date from erroneously being shifted back to the seam month. Respondents who at the previous date

of interview were not receiving UB II are however not reminded of this, and so constant wave responses are not prevented and start dates are erroneously shifted back to the interview month, creating the visible seam effect in transitions into receipt.

Although we used an unique data set to assess measurement error over time, there are some limitations. We used an interaction between panel wave and interview mode as a surrogate for trust in the survey. However, as the mode was not randomly allocated and respondents were put in the CAPI field that were harder to reach, the effect might be biased by selection. Also, individuals that agreed to the record linkage are a selective subsample (Beste 2011). In the PASS study, respondents that are older and report a higher income are more likely to consent. Selective attrition might have further influenced our results. Thus, a balanced panel was used as a sensitivity check (results available on request). For the population sample, a re-analysis was not always possible due to the low number of cases. We did not find any substantive deviances from the full analysis for both recipient and record sample, with one exception. For the recipient sample, we find smaller gaps between record and survey recipients in waves 4 and 5 for all three substantive variables. This supports the hypothesis regarding the association between selective attrition and lower data quality as was also found by Bollinger and David (2001). Due to the design of the study, we can not control for panel conditioning. However, panel conditioning would only be a problem, if the participation on the panel influenced the individual dynamics of welfare receipt, while in this case panel conditioning might increase the quality of reporting over time.

Still, we are confident that the decrease in measurement error over time is in part due to dependent interviewing: in each wave an increasing proportion of recipients are asked the dependent question and are therefore helped by the survey question to provide correct reports. However, whether the decrease in measurement error decreases the bias when comparing welfare recipients with non-recipients remains unclear. More detailed analysis of this phenomenon should be in the focus of further research.

3 Measurement error for welfare receipt and its impact on panel models[1]

Abstract

In this work, the extent and impact of measurement error for welfare receipt is evaluated for up to five panel waves. The extent of underreporting of welfare is known to be considerable in surveys. However, as respondent characteristics can change over time, so can measurement error. The change of measurement error over time can especially bias parameters of longitudinal fixed-effects models as they rely on transitions from one state to another. In this study, the measurement error is assessed by validating data of a German panel study with administrative records. It is found that the measurement error violates common assumptions and decreases over time. This causes an overestimation of the dynamic of welfare receipt. When estimating a fixed-effect model for subjective health, it is found that the measurement error is correlated with the model variables and causes a relevant overestimation of the effect of welfare receipt for sub-populations. In a last step, methods are evaluated to decrease the bias. No method decreases the bias for all subgroups, but such methods might be useful to assess the robustness of model results.

Keywords: panel data, measurement error, administrative data, fixed-effects models

3.1 Introduction

The use of longitudinal panel data is popular in economics and the social sciences. With panel data, subject level information is collected repeatedly on multiple time-points. Thus change in subjects can be observed. The observation of change has the advantage that it enables the estimation of quasi-causal effects with less assumptions than when just using cross-sectional data (Wooldridge 2002). A popular class of such estimators are fixed-effects models as they models rely solely on subject-level change.

Panel data like any other survey data is affected by measurement error. Measurement errors are deviations of the answers of respondents from the underlying true values (Groves 1991, p. 2). The research on measurement error is extensive as it can influence any survey outcome and estimate. However, most of

1 A shorter version of this chapter will be published in Eggs (2015).

this research is conducted cross-sectionally as it focuses on measurement error for one specific time-point. Yet, if characteristics of a subject can change over time, so can the respective measurement error of the characteristic. The change of measurement error over time can especially bias parameters of longitudinal fixed-effects models (see Angrist and Pischke (2008) for discussion). Previous studies (Chowdhury and Nickell 1985; Freeman 1984) found a a substantial attenuation of effect estimates as being biased toward zero. Yet, an attenuation of effect estimates is not necessarily the consequence of measurement error. The direction and size of the measurement bias, caused by measurement error, depend on the particular model specifications, which differ widely between applications.

Lack of longitudinal research on the impact of measurement error is related to the lack of longitudinal validation data, which is more difficult to acquire than cross-sectional validation data. Thus, most research on the influence and extent of measurement error in surveys is conducted cross-sectionally and not longitudinally. Some research was conducted for two subsequent panel waves (Bound and Krueger 1991; Lynn et al. 2012). In this work, validation data is available for a longer period. The extent and impact of measurement error can be evaluated for up to five panel waves. Survey data of the German household panel study "Labour Market and Social Security" (PASS) is used. The survey data is linked on individual level to register data, that is provided by the German employment agency and serves as the necessary source of validation. Comparing the data entries for welfare receipt between the two sources, the individual measurement error can be determined for respondents and its impact be evaluated for a longer period of time than in previous research.

This study focuses on measurement error for unemployment benefit II (UB II), a type of welfare. The extent of underreporting of welfare is known to be considerable in surveys (Bound, Brown, and Mathiowetz 2001; Czajka 2013). Cross-sectional studies by Bruckmeier, Müller, and Riphahn (2014) and Kreuter, Müller, and Trappmann (2010) also found systematic underreporting of UB II in PASS. Systematic underreporting can be caused by social desirable response behaviour (respondents edit and misreport their retrieved information in order to present themselves in a more favorable light (Krumpal 2011)), as the receipt of welfare is not a desirable trait in work-based societies.

This paper focuses on the following research questions: (1) Are classic assumptions about the distributions and correlations of measurement error met for the measurement error for UB-II-receipt? (2) In order to correct for measurement error bias, a range of measurement error models have been introduced over time. Assumptions for these models are also discussed. (3) Whether and in which direction does measurement error for UB-II-receipt distort estimates for fixed-effects models? For this purpose, analyses of the study by Eggs (2013) are recalculated,

using the register information for UB-II-receipt. (4) In a last step, methods to correct for a possible measurement error bias are evaluated.

3.2 Data

The empirical analysis is based on data from the household panel study "Labour Market and Social Security" (PASS), a survey designed for research on the labour market and poverty in Germany (Trappmann et al. 2013). Data from the first five panel waves (2007–2011) is used. In the first wave, about 18 000 individuals in 13 000 households were interviewed. PASS is based on a combination of two subsamples, one of which is drawn from the unemployment II registers of the Federal Agency of Employment, while the second is a general population sample. In each household, first an interview with the household target person is sought, followed by individual interviews with each member of the household aged 15+. In this study, UB-II-receipt is the variable of interest. UB II was introduced in 2005 as a part of the "Hartz" reforms, a major reform package of the social security system. UB II was supposed to be the new basic welfare scheme and as such supposed to provide the minimum resources necessary for an individual to meet his or her basic needs. UB-II-recipients are not necessarily unemployed or vice versa. E.g. if household earnings are not sufficient to provide the bare minimum, UB II can be claimed to bridge the resource gap. Information on UB II is collected in the household questionnaire. In the first interview, the household target person is asked:

> *What about your household? Have you or any other member of your household at any time since [January/Interview Year – 2] obtained unemployment benefit 2?*

If the answer is yes, the time span is collected. PASS is using an event-occurrence approach. This means that the beginning and the end for each spell are collected:

> *From when to when has your household without interruption obtained unemployment benefit 2? Please tell me the month and the year.*

In the subsequent wave dependent interviewing (DI)[2] is used for persons with ongoing spells of UB II at the time of interview of the previous wave.

> *In the last interview in ‹MONTH/YEAR› you stated that the household you were living in then was obtaining unemployment benefit 2 ("Arbeitslosengeld 2") at the time. Until when was this benefit obtained without interruption? Please report the month and the year.*

2 With DI, respondents are reminded of their previous answer. It has been shown, that DI enhances response quality in panel and follow-up studies (Jäckle 2009).

For the purpose of this study, PASS has the advantage of providing a sufficient number of UB-II-recipients and the possibility to link survey reports and register information. In wave 1 PASS had household response rates of 28.7 % for the recipient sample and 24.7 % for the population sample (RR1 according to AA-POR (2009)). In each subsequent wave, refreshment samples were drawn. The refreshment samples consist of households that are first time recipients of UB II. Sizes of the refreshment samples vary around 1 000 households and 1 400 individuals.

The administrative data used to validate survey data is drawn from the IEB (Integrated Employment Biographies) data. The IEB contains longitudinal information on employment, unemployment benefits, and UB-II-receipt. The linkage between PASS survey data and IEB data requires informed consent of respondents. Respondents in the population sample were linked by their name and address, gender and date of birth using error tolerant procedures based on Jaro (1989). Since most UB-II-recipients were sampled from the official registers, direct linkage with the register was possible for this subgroup.

Administrative data is not necessarily free of error or of better quality than survey data (Groen 2012; Kapteyn and Ypma 2007). However, the administrative data for UB-II-receipt is suited to analyse measurement error. For UB-II-receipt, register information is of high quality as it is directly produced by the software that administers benefit claims and payments (Jacobebbinghaus and Seth 2007; Köhler and Thomsen 2009). Information on UB-II-receipt can be extracted for the same point, the date of the interview, from both data sources. Also, the same construct is measured in both data sources (whether any UB-II-benefits were claimed at the date of the interview). Hence, differences between the two data sources can be defined as measurement error.

The analysis is based on 12 169 respondents with 36 909 observations. On average, respondents participate three times in the survey. Individuals that participated only once are excluded, since they do not provide longitudinal information. In order to study the association of UB-II-receipt with health, only respondents in the working age range ($17 < \text{Age} < 66$) who are employed or unemployed are kept in the analysis sample Individuals are discarded that are out of the labour force. For 10 458 respondents with 32 019 observations, the survey information can be linked with register information. According to Beste (2011), the use of data of the linked subgroup can lead to minor selection bias in comparison to the complete sample. Stata® 13.1 was used for statistical analyses.

3.3 Measurement error

Measurement error occurs, if the survey response of an individual deviates from the underlying true value. For any question the survey response of an individual is the product of a four-step response process (Tourangeau, Rips, and Rasinski 2000, p. 8). First, the individual has to understand the question. In a second step, he has to retrieve the information from memory. In a third step, the retrieved information is judged by the respondent. In a fourth step, the chosen response is edited and has to be communicated.

Measurement error can rise in any of the four steps. The understanding can be influenced by the amount of cognitive resources. The quality of the retrieval seems to depend on cognitive factors and the saliency of the event. The editing step can be influenced by social desirability, as respondents want to uphold their image in front of the interviewer. Respondents tend to systematically overreport social-desirable behaviours like voting (Ansolabehere and Hersh 2012) and underreport social-undesirable behaviours (Tourangeau, Rips, and Rasinski 2000, p. 269).

This study focuses on the measurement error for UB II at the time of the interview. Under normal circumstances, only the reported UB II status is known to researchers. The reported UB II status $UB\,II_{it}$ for respondent i and panel wave t can be seen as a function of the true status $UB\,II^{*}_{it}$ and an error term v_{it}:

$$UB\,II_{it} = UB\,II^{*}_{it} + v_{it} \tag{3.1}$$

Classical measurement error is assumed in most research (Carroll et al. 2006), where the measurement error has an expected mean of zero and is not correlated with the true score. These assumptions are tested by defining the measurement error as $v_{it} = UB\,II_{it} - UB\,II^{*}_{it}$. As the variable of interest is dichotomous, v_{it} can take three different values: 0, -1 (underreporting), +1 (overreporting).

In table 3.1 the percentage for over- and underreporting for UB-II-receipt over subsequent panel waves is shown for the analysis sample. Underreporting is more common than overreporting. A reason for this might be social desirable response behaviour as the receipt of welfare is mostly negatively connotated in western societies (Czajka 2013). The extent of measurement error due to underreporting is considerable in early panel waves. Thus, the expected mean for the measurement error can not be equal to zero, causing a systematic difference between observed and true sample prevalences for UB-II-receipt. There is, especially for underreporting, a substantial decline over subsequent panel waves. Bound and Krueger (1991) assume for their measurement error models that the probability for underreporting is equal to the probability for overreporting. Other approaches for error models

assume stable error probabilities over time (Biemer 2011). Both assumptions are not met in this case.

Table 3.1: Under- and overreporting of UB-II-receipt over five panel waves

	Overreporting	Underreporting
Wave 1	2.65 %	7.40 %
Wave 2	2.49 %	5.06 %
Wave 3	2.07 %	5.21 %
Wave 4	1.74 %	3.15 %
Wave 5	1.32 %	2.19 %
N = 10 458 Linked PASS-IEB data		

UB-II-receipt is a dichotomous variable. If the true value is receipt ($UB\,II^{*} = 1$), the value for the error can only take –1, 0. If the true value is no receipt ($UB\,II^{*} = 0$) the value for the error can only take 0, 1. Thus, the true score has to be negatively correlated with the measurement error. A negative correlation of the measurement error with the the true score is sometimes called meanreverting. The degree of the negative correlation $Corr(UB\,II_t^{*}, v_t)$ is depicted in table 3.2 for each panel wave. No trend can be seen for the size of the correlation over time.

Table 3.2: Correlations of the true value with the measurement error

	Corr.
Wave 1	-0.30
Wave 2	-0.28
Wave 3	-0.32
Wave 4	-0.29
Wave 5	-0.26
N = 10 458 Linked PASS-IEB data	

For longitudinal analyses, the degree of serial or auto correlation over time is also of importance ($Corr(v_t, v_{t+1})$). Some classes of panel models like dynamic Arellano-Bond or related auto-regressive estimators require the absence of serial correlation for the measurement error in order to reach unbiased results. Bound and Krueger (1991) and Chowdhury and Nickell (1985) assume for their measurement error models that the measurement error is uncorrelated over time. As can be seen in table 3.3, the errors are serially correlated over subsequent panel waves. The serial correlations vary between 0.21 and 0.39. While the degrees of serial correlation are smaller than found (0.51) in the study by Bollinger and David (2005) who analysed the response error for the receipt of food stamp, the correlations

are larger than found for unemployment (0.12) by Pyy-Martikainen and Rendtel (2009). This supports the hypothesis of Bollinger and David (1997) who assume that some respondents are more predisposed than others to provide the accurate response. To conclude this section, it is safe to say that the measurement error for UB II does not meet the assumptions of common measurement error models. It is unstable over time, not evenly distributed and auto-regressive. The degree of the serial correlation is neither constant nor does it evolve linearly over time.

Table 3.3: Serial correlations over two subsequent waves

	Corr.
$Corr(v_1, v_2)$	0.33
$Corr(v_2, v_3)$	0.29
$Corr(v_3, v_4)$	0.21
$Corr(v_4, v_5)$	0.39
N = 10 458 Linked PASS-IEB data	

3.4 Measurement error and fixed-effects models

In this section, it is assessed, how the measurement error affects effect estimates of linear fixed-effects models. Fixed-effects models are a popular class of models, since inherently all time-constant heterogeneity is controlled for via the process of time-demeaning the data (Allison 2009). For this analysis, the outcome of interest is subjective health.[3] The model is defined as

$$y_{it} = \alpha + UB\,II_{it}\beta_1 + U_{it}\beta_2 + log(H\,HI)_{it}\beta_3 + P_{it}\beta_4 + A_{it}\beta_5 + wave_i\beta + \mu_i + \varepsilon_{it} \quad (3.2)$$

y_{it} is the subjective health score for respondent i in wave t. The health score is a factor score derived from a set of questions on health (health assessment, hospitalizations, health restrictions). $UB\,II_{it}$ is the dichotomous UB II status (UB-II-receipt = 1). U_{it} is the dichotomous unemployment status (Unemployed = 1). Further control variables are the logarithm of household income (continuous), having a partner (yes = 1), age (continuous), panel wave dummy variables (wave 2 = 1 ,... , wave 5 = 1, wave 1 = reference category). α is the constant. μ_i are fixed unknown time-constant parameters, that will be controlled for via mean differencing. ε_{it} is the person and time specific model error. For such a model, it is assumed that $Cov\,(x, \mu_i) \neq 0$, $\varepsilon_{it} = 0$, $Cov\,(x_{it}, \varepsilon_{it}) = 0$.

3 For a detailed discussion regarding the association between subjective health and UB-II-receipt, see Eggs (2013).

Fixed-effects models require transitions in the independent variables for the estimation of the model coefficients. Table 3.4 depicts the number of transitions out of or into UB-II-receipt based on register and survey information. The results show that the dynamic of UB-II-receipt is overestimated in the survey as more entries and exits for UB-II-receipt are found in the survey data than in the administrative data. A large difference can be seen for the number of transitions into UB-II-receipt (entries). 41 % of all transitions into UB-II-receipt reported in the survey can not be validated with register information. This is directly related to the decrease of underreporting over time as seen in table 3.1. Respondents that are not reporting UB-II-receipt at time t_0 begin to report welfare receipt at time t_1. This causes false transitions into UB II. The improvement of cross-sectional data quality is thus directly related to errors in transitions and decreases longitudinal data quality. At one specific point in time, only a smaller number of observation is misclassified as has been presented in the prior section. However, these observations contribute a relative high number of erroneous transitions. As a result the proportion of observations in error will be much larger in a longitudinal analysis than in a cross-section analysis (Freeman 1984, p. 5). Thus, the results of table 3.4 illustrate, why fixed-effects models are more affected by measurement error than different model classes, as they rely solely on observations with transitions for their estimation (Angrist and Pischke 2008, p. 225). The size and direction of the impact of the measurement error on model estimates depends on its covariance with all other variables of the specific model. If studies discuss the problem of measurement error, non-differential measurement error is mostly assumed ($Cov(\nu, \gamma) = 0$, $Cov(\nu, \mathbf{Z}) = 0$). $\mathbf{Z}$ is the vector of control variables. Non-differential measurement error would lead to an attenuation of effect estimates. Non-differential measurement error is also assumed by measurement error models that try to correct for the measurement error bias (Carroll et al. 2006; Chowdhury and Nickell 1985; Hernan and Robins 2014).

Table 3.4: Transitions from and into UB II based on register and survey information

	Survey	Register
From UB II into UB II	9 461	10 716
From no UB II into no UB II	16 033	15 593
From no UB II into UB II (Entries)	1 363	793
From UB II into no UB II (Exits)	2 617	2 372
Total	29 474	29 474
Linked PASS-IEB data		

In order to test for non-differentation, the measurement error is modelled by using the model variables of the primary Model 3.2 as independent variables. Non-differentiation would imply that the model variables are not associated with

the error. As welfare receipt is a dichotomous variable, its measurement error can assume three different values (no error, underreporting, overreporting). As panel data with repeated measures is used, the assumption of non-differentiation is tested by estimating a multilevel multinomial logistic regression using the measurement error for UB-II-receipt at each time point as the dependent variable. Respondents are clustered over time. The model will be estimated separately for men and women, because the subsequent fixed-effect will also be estimated separately. The results for the models are shown in table 3.5. No measurement error serves as reference category.

Table 3.5: Multilevel multinomial logistic regression for over- and underreporting for UB-II-receipt at the time of interview

	Women				Men			
	Overreporting		Underreporting		Overreporting		Underreporting	
	AME	Std. err.	AME	Std. err.	AME	Std. err.	AME	Std. err.
Health score	-0.000	0.000	-0.001	0.001	-0.0005*	0.0002	0.001	0.001
Unemployed = 1	0.001***	0.000	-0.008***	0.002	0.001**	0.000	-0.007***	0.002
log(HH-Inc)	-0.001**	0.000	-0.015***	0.002	-0.001**	0.000	-0.009***	0.001
Age	-0.000	0.000	-0.00001***	0.000	-0.0001**	0.000	-0.0001*	0.000
Partner = 1	0.000	0.000	-0.004*	0.002	-0.000	0.000	-0.002	0.002
Wave 1	*Ref.cat.*				*Ref.cat.*			
Wave 2	0.000	0.000	0.008**	0.003	-0.000	0.000	-0.005*	0.002
Wave 3	0.001	0.000	0.001	0.003	0.000	0.000	0.000	0.002
Wave 4	0.001**	0.001	-0.008**	0.003	0.000	0.000	-0.010***	0.002
Wave 5	0.001*	0.001	-0.014***	0.003	0.000	0.000	-0.011***	0.002
N	16 100		16 100		15 354		15 354	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
AME: Average marginal effects
Basecategory: No measurement error
Linked PASS-IEB data

The dependent variable health score is negatively associated with overreporting for the male subsample. Unemployed recipients are more likely to overreport and less likely to underreport for both subsamples. Respondents, living in households with higher income, are less likely to over and underreport. Age is significantly associated with underreporting for women and with both types of error for men. Living in a relationship is associated with underreporting for the female subsample. Overreporting is more likely in later waves for women. As underreporting decreases over time, the indicators for later panel waves are negatively associated with the indicator for underreporting.

Using the model variables as predictors for the occurrence of measurement error, the results show that many of the model variables are significantly correlated

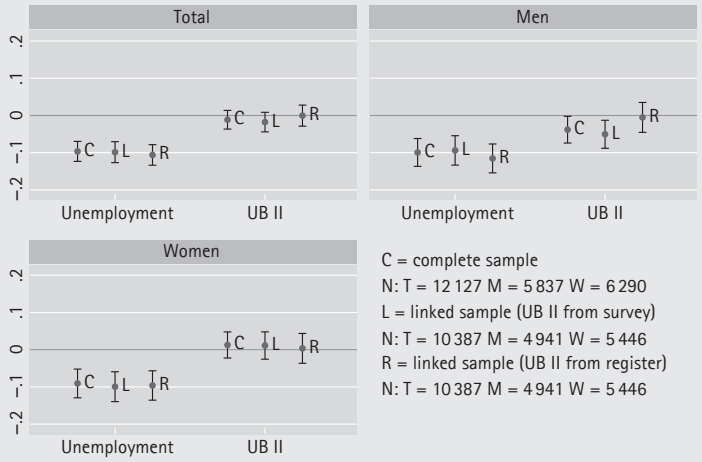
with the measurement error for welfare receipt, even if the size of the marginal effects is small. An explanation for the results for unemployment can be that employed recipients that might be on the brink to be eligible to claim welfare are more likely to misreport than individuals that are certain recipients. Similar results were found for underreporting by Bruckmeier, Müller, and Riphahn (2014, 2015) using cross-sections of linked PASS-IEB data. Respondents that are more integrated in the labour market might be less inclined to report their UB-II-receipt. This would divulge that their work does not provide sufficient resources to make ends meet. The associations for income can be explained as individuals in higher income are less likely to receive welfare and are more likely to classify their welfare status correctly. This study also investigated some correlates for overreporting. The results indicate that respondents are more likely to overreport, if they are similar to actual welfare recipients as unemployed individuals and respondents with lesser income are more likely to overreport.

It was shown in previous paragraphs that measurement error for UB-II-receipt does not fulfill the criteria for classical measurement error and that it is correlated with most of the model variables. Hence, an attenuation of effect estimates is not a necessity. In order to evaluate the effect of the error on the model estimates, a stepwise procedure is chosen. In a first step, models are calculated using only survey data and the complete analysis sample. In a second step, the models are recalculated for those respondents that could be linked, using otherwise the same information as in the first step. By comparing the results of the first two models, one can assess a possible selectivity of results due to the linkage. Large differences between the effect estimates between the linked and the complete sample would endanger the external validity of the results of the third step. In the third step, a register sample is created by replacing the survey entries for welfare receipt with the register entries. Thus, except the entry for welfare receipt, the model specifications for the second and third are otherwise identical and differences in the results can only be caused by the measurement error for UB-II-receipt in the survey data. The impact of measurement error can then be assessed by the calculation of the relative bias.

Results for the three model specifications are presented in graph 3.1 and the respective tables can be found in the appendix.[4] Effect estimates for unemployment and UB-II-receipt are shown for the complete sample, linked sample (linked data with survey information) and register sample (linked data with register entry). Results are shown separately for the male and female subsamples.

4 The full model results are shown in table A1 for men and in table A2 for women.

Figure 3.1: Adjusted FE linear coefficients and 95% confidence intervals for subjective health score; separated by gender



Adjusted for panel wave, age, log(HH–income), partner
Source: PASS Wave 1–5.

Comparing the effect estimates for UB-II-receipt for the complete analysis sample with the linked analysis sample, no differences can be seen for the estimates for women. For men, a restriction on the linked sample leads to larger negative effect for UB-II-receipt. For both gender-specific subsamples, a restriction of the analysis sample on linked individuals causes only minor differences on effect estimates for UB II. Estimates for unemployment are not affected by the restriction of the analyses to the linked sample. The linkage does not cause a sizable shift in the effect estimates.

When assessing the impact of measurement error in the next step, the measurement error does not bias the results for women considerably. The coefficient of the linked sample is .011, the coefficient of the register sample is .004. The relative bias due to measurement error is thus .373. A different picture emerges for men. Using survey information, UB-II-receipt has a negative, significant association with health of –.051. Using register information instead, UB-II-receipt has an association with health that is close to zero (–.005). Measurement error leads to a sizable relative bias of –2.102. For this example, measurement error does not cause attenuation, but causes a considerable overestimation of the effect estimate. Additionally, it causes a slight attenuation of the effect of unemployment for men. This also stresses the sometimes overlooked possibility that measurement error in one variable can also bias the estimates of correlated variables. For women on the other hand, the coefficient for unemployment remains unaffected by the measurement error for welfare receipt. The coefficients for the remaining control variables are not affected for both subgroups despite being correlated with the measurement error.

For this empirical example, minor bias is caused by the selective linkage but considerable bias is caused by measurement error for the male subpopulation. Differences between the effect estimates for men and women seemed to be mainly caused by measurement error. Hence, while the survey coefficients for UB II differ considerably between the male and the female sample, the coefficients based on register information equate to zero in both groups.

3.4.1 Measurement error models

In the previous section, it was shown that the measurement error causes a substantive overestimation for the UB II coefficient for men. In order to correct for measurement error, a range of correction methods were developed over time, e.g. Carroll et al. (2006), Hernan and Robins (2014), and Schneeweiß and Mittag (1986). However, the methods are mostly highly technical, hard to implement, target cross-sectional data and rely on assumptions that are most likely to be violated. Hence, the questions arises, how to to proceed from a practical point of view, when a variable is affected by measurement error and validation data is not available. In this section, four strategies are evaluated that are easy to implement and might reduce the bias due to measurement error or are used as sensitivity checks. The methods are evaluated using the subsample of respondents that agreed to the linkage.

The first method used to reduce the bias is a correction method proposed by Chowdhury and Nickell (1985). They propose to reduce the bias by averaging the model variables over two successive panel waves. The model is then estimated with the averaged information. The method is easy to implement. However, the method relies on the assumption that the measurement errors are uncorrelated over time, which was shown to be violated for this data in a previous section. The second strategy discards the information of the first interview for each respondent. This strategy is chosen as the degree of measurement error is highest in the first wave and data quality tends to to increase over panel waves (Rendtel 2012). Thus, the second strategy should decrease the number of erroneous transitions in the data. The third strategy also targets the information of the first interview. Instead of removing the first interview from the data, the survey entry for UB-II-receipt is replaced with the information from the sampling frame for the first interview. Thus, if the household of an individual was drawn out of the administrative record, his or her information will be set as having received at the time of the first interview, irrespective of the original survey response. For the fourth strategy, the sample is restricted to the balanced panel. In a balanced panel, only cases are kept that participated in every panel wave. Individuals that drop out from studies are more likely to misreport (Bollinger and David 2001). Thus, balanced panels are used as sensitivity checks or to control for selective panel attrition.

The results for the the different methods are presented in table 3.6.[5] The results for the linked sample in the first column serve as a reference in order to assess whether the use of a method causes a decrease of the bias. The last column shows the coefficients of the register sample that are assumed to be the "true" regression scores in this case. Using the averaged sample, the relative bias for men is nearly doubled in comparison to the linked sample. For women, the bias decreases. This is surprising as the absence of auto-correlation is a key assumption for the averaged method and the degree of auto-correlation does not differ gender-wise. The second strategy, the discarding of the first panel wave, increases the bias for men and slightly decreases the bias for women. Using the information from the sampling frame instead of the survey response for the UB II entry for the first panel wave considerably decreases the bias for the male subsample. Yet, this strategy increases the bias for the female subsample. Using the balanced panel increases the bias for men considerably, yet results in the smallest bias for women. Having applied four different methods to correct for the bias due to measurement error, no clear result emerges. While some methods decrease the bias for one subgroup, no method decreases the bias for men and women. For men, only the replacement of the survey entry with the proxy information derived from the sampling frame decreases the bias in comparison to the results of the linked sample. All other methods increase, sometimes even considerably, the bias. For women, using the averaged sample, the omission of wave 1 or the balanced panel decreases the bias. Irrespective of the method used, the results for women are fairly robust. Hence, if the bias is not large to begin with, different model specifications do not have singular impacts. Thus, even if the different model specifications do not necessarily decrease the bias, their use might still serve as an indicator for the general robustness of the results. The overall largest bias was seen for the balanced panel for the male subsample. Hence, its use might be a sensible check regarding the robustness of the results but not necessarily as a method to reduce the bias due to measurement error.

Table 3.6: Regression coefficients and bias of different model specifications for fixed-effects models for UB II on subjective health

	Linked sample		with averaged values		without t_1		using sampling frame entries		Balanced panel		with register information
	Coefficient	Relative bias	Coefficient	Relative bias	Coefficient	Relative bias	Coefficient	Relative bias	Coefficient	Relative bias	Coefficient
Men	-0.051	-2.210	-0.088	-4.055	-0.064	-2.846	-0.032	-1.286	-.0923	-4.244	-.005
Women	0.011	0.373	0.008	0.236	-0.001	-0.222	-0.012	-0.735	.001	-0.126	.003

Source: Linked PASS-IEB data.

5 The complete model results are shown in table A1 for men and in table A2 for women.

3.5 Discussion & conclusion

Linking register data with survey data on an individual level for five subsequent panel waves and using the register data as a way to validate the survey response, this paper provides new insights regarding measurement error for welfare receipt and its impact on longitudinal panel models. It is also possible to test the classical assumptions for measurement error as well as ascertain interdependencies between measurement error and model variables.

The paper focused on the measurement error for UB-II-receipt at the time of the interview. The measurement error is correlated across panel waves. Welfare receipt is underreported and to a lesser extent overreported. The extent of the measurement error decreases significantly over panel waves and thus the cross-sectional data quality increases over panel waves. However, this causes a high number of erroneous transitions into UB-II-receipt. Transitions are a necessary prerequisite for all kinds of longitudinal analyses and a necessary condition for the computation of fixed-effects models.

The study shows that the measurement error is correlated with most of the variables in the fixed-effect model. It is argued that measurement error for current UB-II-receipt is not caused by a random process like cognitive decay and thus can not be seen as "white noise". Instead it seems to be caused by social desirability as individuals tend to avoid the stigma attached to UB II (Booth and Scherschel 2010) and deliberately misclassify as age and employment status were correlated with the measurement error in this study. With increased age, social desirability decreases (Soubelet and Salthouse 2011). Employed individuals are more likely to underreport but less likely to overreport. It seems reasonable to assume that social desirability is more common for respondents that are closer to the labour market than for respondents that are more deeply entrenched in welfare receipt. This reporting behaviour also artificially increases the differences between recipients and non-recipients in the survey data as underreporting individuals are more similar to non-recipients than to recipients (Jäckle, Eggs, and Trappmann 2015). Similar results can be found in the reporting behaviour for voting. Non-voters that overreport voting have similar characteristics as voters (Ansolabehere and Hersh 2012).

Regarding fixed-effects models, previous research mostly assumed that such models tend to be severely attenuated in the presence of measurement error (Angrist and Pischke 2008). In this case, the measurement error is highly differential and the assumptions for classical measurement error are violated. Thus, an attenuation of effect estimates is not a necessity and also not found for this study. Measurement error causes a severe overestimation for the effect estimates for UB-II-receipt for men. For women, measurement error did not cause major bias in the effect

estimates. Four different strategies were evaluated in order to reduce the bias in model coefficients due to the measurement error. However, when applied, the error models performed neither well nor consistent for the male and female subsample.

There have been additional attempts to correct for measurement error in panel data. Most measurement error models depend on the assumption of a random error-causing process, implying a simple structure for measurement error (Battistin and Chesher 2014; Küchenhoff, Mwalili, and Lesaffre 2006). For UB-II-receipt, these assumptions do not hold and were shown to be violated. Using those methods could thus cause more harm than good, at least for outcomes like welfare receipt, since false assumptions about the error-generating process are used. Establishing a general model for correcting for impact of measurement error for welfare receipt seems not feasible, since the direction of the bias for model estimates depends on the covariance structure between the measurement error, all used variables, and on the chosen modelling strategy. All three factors vary widely between single analyses. The author is skeptical, whether a general error-correcting mechanism can be established in such circumstances. The non-random error-generating process, combined with the specific statistical model, requires a different strategy for each application.

Instead, when faced with a variable, that is known to be affected with non-standard measurement error, one could conduct a range of subgroup analyses and sensitivity checks in order to assess the robustness of the results. To assess the robustness, one could use the methods that were applied in this study and were previously thought to reduce measurement error. For men, where the measurement error has an impact on the coefficients, no robustness for the results can be determined due to large deviances between the method-specific results. For the female subsample, where the measurement error has only a minor impact, the results are more robust as the deviances between the results for the different model specifications are small. Another strategy would be the use of different types of estimators as the impact of measurement error depends on the model specifications (Angrist and Pischke 2008; Millimet 2011). Each modeling approach would be associated with some restrictions, but robustness could be assessed across a range of model results. Also, as data quality tends to improve over panel participation, one could use only later waves when using data of a long-running panel study.

This study has a range of limitations. The impact of the measurement error is only observed for a selective sample that could be linked to the register. However, since the linkage itself causes only minor shifts in effect estimates, it is assumed that the bias due to measurement error is not selective for the linked sample and a similar impact of the measurement error is anticipated for the complete sample. This analysis explored the impact of measurement error for one model

specification and one dependent variable. However, the analyses were recalculated using two additional health-related outcomes. The same patterns for the biases were found for the male and female subsamples. Also, the set of control variables age, panel wave, household income, and unemployment are highly correlated with the measurement error. These variables are a popular set of control variables and are widely used when analyzing welfare receipt. The impact of measurement error might also be caused by the mis-specification of the underlying model. Mis-specification is always a possibility. In this case, the measurement error only affects the model results for men. Thus, one would also assume that the mis-specification of the model only affects the male subsample and not the female subsample and subsequently that a different modeling strategy would be necessary for each sex. Using such an approach would also be hard to justify. This study investigates UB II, which is a nation-specific welfare scheme. While UB II is a German welfare program, measurement error for welfare can be observed in most western countries (Bound, Brown, and Mathiowetz 2001). It is more than likely that measurement error is also inconveniently behaved in these settings as well.

This study provides additional evidence that the properties of measurement errors are context-specific and classical measurement error is not the usual case. Using an empirical example it shows that measurement error can also lead to a severe overestimation of the effect estimates and that easily applicable error models do not necessarily reduce the measurement error bias, but might even increase the bias. The results of the study thus further bolsten the statement by Bound, Brown, and Mathiowetz (2001, p. 3 775) that "the possibility of non-classical measurement error should be taken much more seriously by those who analyze survey data, both in assessing the likely biases in analyses that take no account of measurement error and in devising procedures that "correct" for such error."

4 Errors in retrospective welfare reports and their effect on event history analysis

Johannes Eggs and Rainer Schnell

Abstract

Errors in autobiographical reports may bias labor force participation statistics based on surveys. Therefore, we compare time-to-event models based on administrative data with models based on respondent reports. Respondent data from a large German panel study (PASS) is matched to administrative records. Although differences in descriptive point estimates are obvious, the differences between the estimates of time-to-event models are small and not significant. Therefore, this study supports the use of survey reports for testing time-to-event models on labor force participation and welfare receipt.

Keywords: survey error, autobiographical memory, recall, survival analysis, welfare benefits, administrative data, validation, record linkage

4.1 Introduction

Individual transitions in and out the labor market are of central importance for policy makers and labor market researchers. As information from administrative micro data is limited, most studies on duration of unemployment or welfare spells use survey data. The only way to collect such data in cross-sectional surveys is the use of retrospective questions. Recall for autobiographical events and spells is prone to a wide range of response errors as respondents can omit, misdate, merge, misclassify or invent events or spells (e.g. Cannell, Marquis, and Laurent (1977), Gray (1955), Paull (2002), and Thompson et al. (1996)). Thus, the preferable way to collect such information, are repeated measurements: "If change over time is of crucial interest, concurrent measures at different points in time are the only reliable way to assess it." (Schwarz 2007, pp. 20–21). Such concurrent measures are collected in panel surveys. However, as the elapsed time between two panel waves can be substantial and panel surveys also collect information for the baseline, still retrospective information is collected in most panel studies. Naturally, this information will be also affected by recall errors. Research in cognitive psychology (Roediger 2008) and in survey methodology (Belli, Bilgen, and Al-Baghal 2013; Tourangeau, Rips, and Rasinski 2000) has shown that the extent of such errors depends on characteristics of the event, the ability and response strategy of the

individual, and the mode of data collection. However, when modelled, it is mostly assumed that the response error is based on random noise.

Using administrative data to validate the responses for welfare receipt in two panel waves, in this study we can assess the amount of response error in the reported spells of welfare receipt. We can also assess, whether the response error is distorted and correlated with individual characteristics or whether it follows a random distribution.

However, even if the information is distorted by the response error, the crucial point from a statistical point of view is: Do errors in autobiographical reporting bias the results of the subsequent statistical analyses to such a degree that different conclusions would be drawn? Therefore, we compare time-to-event models based on administrative data with models based on respondent reports.

4.2 Previous research

Response error in event histories has received much attention in the literature as response error is one of the major causes of measurement error. There are two primary methods to assess measurement error in surveys. The first method is a test-retest design, where information for the same time frame is collected twice at two separate occasions from the same individual. The second method is the use of external data, where the survey responses of individuals are validated with entries from mostly administrative records from the same individual. Bound, Brown, and Mathiowetz (2001) give an extensive overview of studies on measurement errors in labour-market surveys. The overview and additional studies (Bollinger and David 1997; Bruckmeier, Müller, and Riphahn 2014; Kyyrä and Wilke 2014) show that labour market related events and events related to welfare receipt are likely to be misreported in surveys. However, most studies validated the accuracy of the survey response for the time of the interview and not for retrospective events. Few studies have analyzed the effect of these kind of errors on the resulting estimates of time-to-event models.

One of these studies is by Pierret (2001). He used data from the National Longitudinal Survey of Youth 1979. In an experimental subsample, respondents were asked about employment and welfare events in the previous year and again a year later about the same events in the previous two years. Thus, a test-retest situation for the two year period could be analysed. He found that the quality of recall was lower. Shorter spells of welfare receipt and employment were more likely to be underreported. To evaluate the impact of recall error on a time-to-event model on leaving welfare, Pierret (2001) compared the model results for the experimental sample with model results for the main sample, employing the same

model specifications in both samples. He found differences, yet the differences were not significant. Since the number of welfare recipients was very small in the experimental sample, this could also be due to a lack of statistical power.

Jäckle (2008) compared spell information from the British Household Panel study with spell information derived from administrative data. Assessing correlates for the recall error, she found that respondents with lower educational qualifications were more likely to report longer welfare spells. The results showed that measurement errors did bias estimates from event history data. The durations of long spells tended to be underestimated and that estimates were in general smaller, when using survey data. The primary aim of the study was to clarify how dependent interviewing[1] (DI) and independent interviewing (INDI) influence the response quality for a range of labor and welfare related benefits. The use of DI increased the quality of the reporting for longer spells, but not for shorter spells. In sum, the use of DI decreased the bias, when estimating the event history models.

In the study by Pyy-Martikainen and Rendtel (2009), spell information from five waves of a Finnish labor market panel survey was validated with register data. They found that shorter spells were more likely to be underreported. Investigating the recall error for unemployment error in detail, they found that the error was correlated across panel waves, was correlated with spell properties and was correlated with explanatory variables for the spell duration. Spell duration, sex, age and the education levels were strongly correlated with the recall error. They also found a substantial heaping of spell-defining events on interview months. Analysing the impact of the error on model coefficients using different model specifications, they found sizeable biases for selected variables. The bias caused both an overestimation and an attenuation of effect estimates.

Using Swedish data, Pina-Sanchez, Koskinen, and Plewis (2013) evaluated the impact of recall error on event history models explaining the duration of unemployment. Information on unemployment was collected for a one year recall period. Administrative data could be used to validate the survey reports. In this study, model variables were not correlated with the recall error and the authors established that the error caused a considerable attenuation of effect estimates. They argue that the recall error for unemployment assistance is thus non-differential. However, the number of observations in this study was small and the set of explanatory variables was sparse.

1 In most longitudinal studies, a substantial heaping of transitions can be observed on the seam between successive panel waves. This phenomenon is known as seam effect or seam bias. Dependent interviewing is seen as a tool to reduce its extent (Moore et al. 2009). With dependent interviewing, information of prior waves is used to remind subjects of their previous response.

4.3 Data

In this study, we use data of the German panel survey "Labor Market and Social Security" (PASS). The aim of this study is to assess the extent of recall error for welfare receipt. We will focus on recall error for unemployment benefit II, which is the most common type of welfare in Germany. Welfare benefit and unemployment benefit II will be used synonymously for the remainder of the text.

PASS is one of the largest household panel studies in Germany. Starting in 2006, this ongoing mixed-mode panel survey (CATI or CAPI) includes yearly information about 10 000 households with 15 000 respondents of the general population aged over 15 (Trappmann et al. 2013). The panel survey has been primarily used to study the effects of the German welfare reforms in 2005. Since then welfare benefits are granted on a benefit unit level. A benefit unit consists of at least one adult plus potential spouse plus any potential dependent children they are living with. A benefit unit is not necessarily, but mostly, congruent with the specific household. PASS is a dual frame survey. On arm of the frame is formed by administrative records of welfare recipients that were provided by the German Federal Labor Agency. A commercial listing of housing served as the second frame for a population sample. The dual frames of PASS were needed for the comparison of welfare recipients with non-recipients. We use data of the first two panel waves of PASS. The response rate (AAPOR-1) for the initial survey (12/2006–7/2007) was 26.7 %, in the second wave (1/2008–7/2008), 56.6 % of the first wave respondents could be reinterviewed (Gebhardt et al. 2009).

Each head of household answered a household interview, containing questions on the household situation and its receipt of welfare benefits. If anyone in the household ever received welfare benefits in the reference period, the beginning and the end of each spell were collected by an event-occurrence approach. Each member of the household aged over fourteen was individually interviewed with questions on e.g. education, personal income and migration.

4.3.1 Administrative data and linkage

For the validation of survey responses, external data is needed. For most validation studies, administrative data is used (Schnell 2014). Two types of record validation studies are common. In reverse record studies, the administrative data base is used as sampling frame (Marquis 1978). In forward record studies, records from surveys are linked to an administrative data bases after sampling has been done. As PASS is a dual frame study, this study is a reverse record study for the recipient sample and a forward record study for the population sample.

In order to validate the survey response for welfare receipt, we use a database containing all spells of welfare benefit receipts for individuals and benefit units in Germany, called "LeistungsHistorik Grundsicherung" (Geschäftsbereich ITM 2009). For research purposes limited access to this database according to the local privacy protection laws was provided by the Research Data Centre of the Social Security Administration. The database is considered as covering all cases in the population. Nevertheless, the database may contain a few errors. But since the data used in our analysis is used for processing claims and the administration of payout of the welfare benefit, the validity of the administrative spell data for welfare receipt is presumably high (Jacobebbinghaus and Seth 2007; Köhler and Thomsen 2009).

Case selection

Figure 4.1 describes the case selection for the analysis reported here. As the information on welfare receipt is collected on the household level, we restrict our analyses to the responses of heads of households. To ensure recall periods of the same length, only households participating in both waves were used. To avoid linkage errors, households with more than one registered benefit unit are discarded as the information in the household interview might refer to different benefit units. Due to the same reasoning, households had to be excluded that had changed in their composition during the data collection period.

According to German data protection laws, administrative data was only linked with the survey data if respondents gave their permission for linkage during the interview.

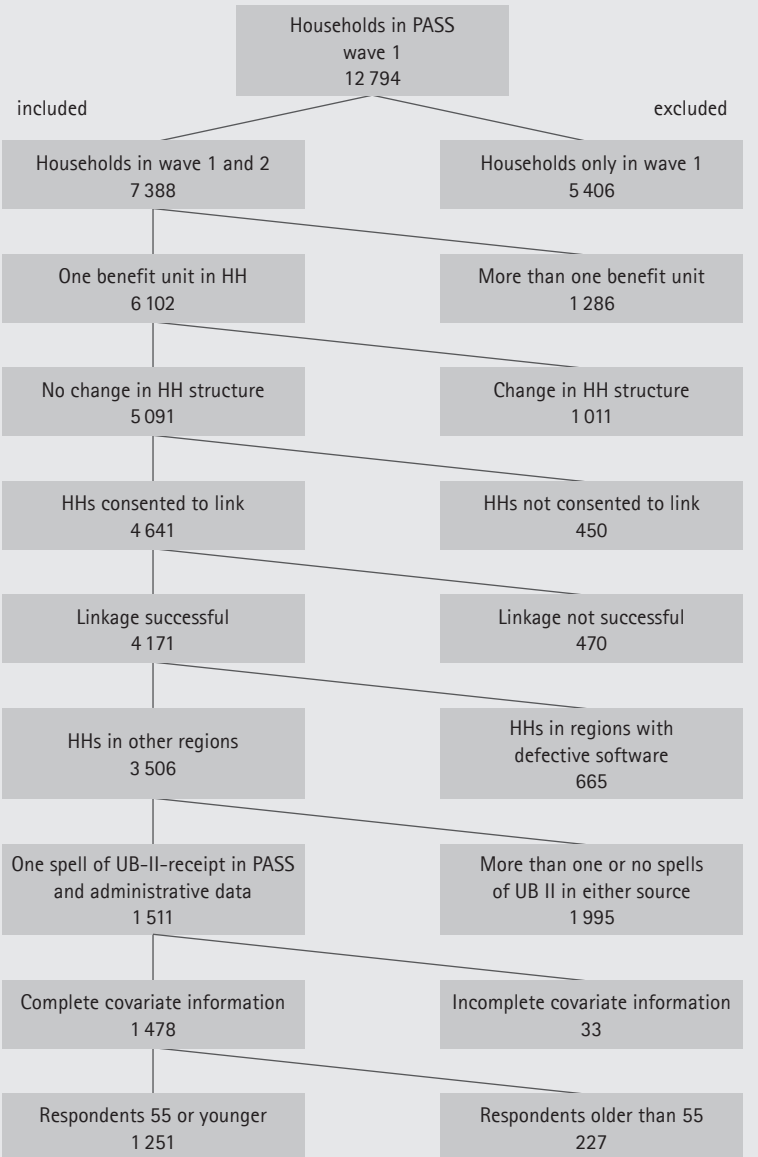
Due to technical errors in the administrative database, benefit units in some geographical regions had to be excluded.[2] Due to missing administrative records or missing identifiers, despite the use of a special record-linkage record program (MTB, (Schnell, Bachteler, and Reiher 2005)) about 10% of the consenting households could not be linked and are excluded.

To avoid spell matching problems (Luks and Brady 2003; Pigeot-Kübler and Schnell 2006), only respondents with a single spell of welfare receipt are included. If respondents with multiple spells are included, it would not be possible to accurately identify the recall error for specific welfare spells. Only respondents with complete information on all covariates used in the model are included. Because people aged over 55 often end welfare receipt by early retiring and

2 Therefore, this exclusion can be considered as missing at random, since the technical problems are not related to any characteristic of the respondents. For further details, see (Geschäftsbereich ITM 2009).

hence employ a different exit strategy than the rest of the population (Achatz and Trappmann 2011), we excluded this group. An inclusion of this group would have distorted the time-to-event models. Finally, 1 251 respondents remained for our analyses.

Figure 4.1: Steps of case selection for the study



4.4 Hypotheses

Time-to-event models use information on spell duration and censoring status. If spell duration or censoring status are misreported and if these errors are correlated with explanatory variables, estimates of time-to-event models are likely to be biased. The extent of recall error varies by the response difficulty of the question, the saliency of the event, cognitive abilities and response strategy of the respondent (Friedman 2004; Roediger 2008). Based on previous research, twelve hypotheses were formed.

Numerous studies (e.g. Rubin and Wenzel (1996)) beginning with Ebbinghaus (1885) have shown that with increasing recall period the human memory is less likely to provide accurate information:

H1 Recall error increases with the length of the recall period.

Furthermore, research strongly suggests that the timing of a past event is reconstructed by relating it to other known past events (Friedman 2007). Hence, if an event happened near an important public or private event, a so called anchor, it is dated more accurately (Loftus and Marburger (1983), Talarico and Rubin (2003)):

H2 Recall error is lower for spells that begin on temporal anchors.

Trivially, the reporting quality for right-censored spells should also be more accurate:

H3 Recall error is lower for ongoing spells.

The occurrence of spells for other types of social assistance increases the difficulty, since the respondent has to discriminate different spells and temporal boundaries (Belli, Bilgen, and Al-Baghal 2013; Thompson et al. 1996):

H4 Recall error increases with additional spells of social assistance.

Non-native speakers require increased efforts for interpreting and and processing the question (Tourangeau, Rips, and Rasinski 2000). Therefore, we expect:

H5 Recall error is higher for non-native speakers.

Persons with higher cognitive abilities seem to retrieve autobiographic memories more easily (Friedman 2004; Roediger 2008), therefore:

H6 Recall error is lower for persons with higher cognitive abilities.

Although age of the respondents may influence cognitive abilities (Thompson et al. 1996), in our case, we expect no influence of age on response quality since we include only persons who are 55 or younger:

H7 Recall error is not related to age within the age range considered in this study.

Besides difficulty and ability, *satisficing behavior* influences the quality of the response. The concept of satisficing was developed by Krosnick and Alwin (1987). The satisficing respondent uses a mental shortcut in his answering process, if the answer is hard to obtain. Dating processes are mentally exhaustive (Burt et al. 2000). Consequently the satisficing respondent is expected to give a reasonable but not an accurate response.

H8 Recall error increases with the amount of satisficing.

Not only the duration but also the censoring status might be affected by survey response error. Respondents can misreport their current welfare receipt status; therefore their censoring status will be misclassified. Since misclassification of the present receipt can not be explained by memory effects, an influence of ability and difficulty is not expected:

H9 Misclassification of the current status of welfare receipt is not related to cognitive abilities, language comprehension and number of spells.

However, satisficing behavior might increase the propensity to misclassify the current status of welfare receipt:

H10 A higher degree of satisficing increases the propensity of misclassification of the current status of welfare receipt.

Since response quality may depend on the survey mode (Biemer 2001), it should be controlled when testing such hypothesis.[3]

3 Data collection in PASS is conducted in a mixed-mode design with Computer Assisted Personal Interviews (CAPI) and Computer Assisted Telephone Interviews (CATI). The mode was not randomly assigned: Fieldwork started with CATI, but harder to reach respondents were reallocated to CAPI.

4.5 Methods

4.5.1 Definition of the recall error

The length of a spell is defined by the time between a beginning and an end. In this study, we can identify the time of the beginning and end in survey and administrative data for each respondent. Is_{bi} is the survey response for the time of the beginning and Ir_{bi} the time of the beginning derived from the register data, i.e. the administrative records, for each respondent i. Concordantly, Is_{ei} is the survey response for the time of the end point and Ir_{ei} the time of the end point derived from the administrative records. With the spell defining events one can create the spell durations for survey Ts_i and register Tr_i. Thus, we can compare the dates of the defining events and the resulting length of the spell for each person and calculate the size of the respective recall error:

$$e_{bi} = Is_{bi} - Ir_{bi}$$
$$e_{ei} = Is_{ei} - Ir_{ei}$$

It has been well documented that in comparison with the time an event actually happened, recent events are dated too far back in the past. This is known as backward telescoping (Janssen, Chessa, and Murre 2006). In contrast events further in the past are placed more, yet less accurately, in the present. This is called forward telescoping in survey research (Huttenlocher, Hedges, and Bradburn 1990; Sudman and Bradburn 1973). The occurrence of telescoping in survey response leads to the assumptions that the expected value of e_b is larger than the expected value of e_e. In addition the standard deviation of the error of spell beginnings should be larger than for spell ends. Therefore, the distributions of the observed recall errors should be contrary to the classical assumptions for measurement error, since the hypothesized error distributions would lead to a shortening of reported spell durations in comparison with the spells in the register. Assuming a normal distribution, a two sample t-test is used to compare the means of the errors for spell beginnings and ends. An one sample t-test is used to test whether the error for spell durations differs from zero. A F-test for the equality of variance (Brown and Forsythe 1974) is used to compare the variances for the errors in spell ends and spell beginnings. Hypotheses on causes of response error are tested with logistic regression models.

4.5.2 Operationalizations

Not all hypotheses can be directly tested. Thus, surrogate information has to be used in some instances. Respondents that communicate mainly in a foreign language have lower language comprehension (Esser 2006). Hence, this information is used as a proxy for lower language comprehension. Since no direct measures of cognitive abilities are employed in PASS, we use levels of formal education as proxy information. Levels of formal education were correlated with response error in previous studies. Satisficing was measured by the relative proportion of rounded answers in quantitative questions and the proportion of "Don't know" or "refused" responses. Non-differentiation was measured by the average degree of entropy over four item batteries. These indicators are commonly used as indicators for satisficing (Krosnick 1999). Additionally, placing the spell beginning at the seam between wave 1 and wave 2 was also used as indicator for lacking response effort. The definitions of all variables used are found in table A3.

We fit three logistic regression models each for the occurrence of recall error in spell beginnings (Model 1 – Model 3) and spell ends (Model 4 – Model 6). The dependent variable is defined in three different ways. In the first and fourth model, the indicator for response error is set to one, if the beginning or end of the spell differs at all between survey and administrative records. In the second and fifth model the response error is set to one, if the beginning or end of the spell is larger than one month. In the third and sixth model the response error is set to one, if the beginning or end of the spell is larger than two month. This approach was chosen in order to test for different mechanisms underlying the response error.

Misclassification of the censoring status is the second type of investigated response error. We fit three logistic regression models for the types misclassification (Model 7 – Model 9). In the first model (M7) the dependent variable is set to one, if the censoring status differs between survey and records. In the next model (M8) only those misclassifications are considered where the spell is censored in the survey but not in the register. This can be seen as underreporting, as current receipt is not reported. In a last model (M9) we consider misclassifications where the spell is censored in the register but not in the survey. This can be seen as overreporting. For the modelling of the recall errors, the information for welfare spell characteristics is based on register data, but the covariate information for personal characteristics is taken from the survey.

4.5.3 Effects on time-to-event analysis

In order to investigate the impact of response error on coefficients of time-to-event models, we specify different proportional hazard models for the exit out of UB II (Model 10 – Model 13) that are based on different combinations of survey and administrative data. The model using only spell information from the register (M10) is considered as the benchmark model. With the results of this, we compare the model using only survey data for the definition of the welfare spells (M11). To disentangle the effect of response error with regard to censoring and spell duration, one model (M12) uses censoring information from the survey and duration information from the register and another model (M13) uses censoring information from the register and duration information from the survey. All models are adjusted for covariates that according to Achatz and Trappmann (2011) are assumed to be associated with leaving UB-II-receipt. Summary statistics and definitions of the covariates are shown in table A3. Proportional hazard assumptions are checked graphically by using log-log plots (Collett 2003).

4.6 Results

4.6.1 The distribution of recall error

Analysing the recall error for welfare events, we find that 43 % of the spell beginnings and 16 % of the spell ends were reported with error. The mean of the error for spell beginnings is 1.63 months and is significantly larger than the mean of the error for spell ends –0.39 ($t = -5.45$, $p < .01$). This means that spell beginnings are telescoped into the present and spell end are telescoped in the past. This response behaviour causes a significant shortening of the reported spell length in general ($Ts_i - Tr_i < 0$: $mean = -2.03$, $SD = 7.26$, $t = -9.89$, $p < .01$). The variance of the error for spell beginnings ($SD^2_{e_b} = 6.90$) is significantly larger than the variance of the error for spell ends ($SD^2_{e_e} = 2.57$) ($\sigma^2_{e_b} > \sigma^2_{e_e}$: $F = 7.00$, $p < .01$). The accuracy for spell ends is better than for spell beginnings. This confirms our previously stated relations for the error distributions.

4.6.2 Explaining response error

The results for our multivariate models can be seen for spell beginnings in table 4.1 and for spell ends in table 4.2. The results of the models for the misclassification are shown in table 4.3.

We do not find evidence that H1 is supported. The recall error is not significantly associated with the length of the recall period. Our results give support to H2. The indicator for temporal anchor is significantly and negatively correlated with recall error in all three models (M1–M3). If the welfare spell began in January 2005, respondents are less likely to misdate the spell beginning. Testing H3, we find no significant associations between the indicator for an on-going welfare spell and errors in spell beginnings. For spell ends, respondents are significantly less likely to misdate the end, if the spell is on-going. As the results for H3 differ between beginnings and ends, H3 can only be partially rejected. For H4, no evidence can be found. We only find a significant association between the additional spells of unemployment assistance and recall error in Model 1. However, larger recall errors for spell beginnings and recall errors for spell ends are not associated with the presence of additional spells. Results for H5 are mixed. Significant associations are seen in the models for larger errors for spell beginnings (M2, M3, table 4.1). For spell ends, the household language is not associated with the recall errors. H6 can not be supported, even if all results do not point in the same direction. Vocational training is significantly and negatively associated with recall errors for spell beginnings for all three specifications of the error. For a university degree and levels of schooling, however, no associations with recall error for spell ends can be found. Recall errors for spell ends are also not correlated with levels of vocational training or education. Thus, there is little evidence that indicators for cognitive ability are associated with less recall error. H7 is not supported in our analysis as respondents in higher age groups seem to to report with less errors. For spell beginnings, the indicators for higher age groups are negatively and significantly associated with recall error. For spell ends, the indicator for the highest age group is significantly and negatively associated with recall error. Our data do also not support H8 in a substantive matter. Some of the indicators for satisficing (non-differentiation in M3, rounding in M6, proportion of item-nonresponse in M4) are associated with the recall errors for spell beginnings or spell ends. As we also control for the interview mode, we find that respondents that were interviewed personally and not by telephone were more likely to misreport.

Analysing the models for misclassification in table 4.3, we find support for H9. Misclassification of the current status of welfare receipt does not seem to be associated with indicators for task difficulty or respondent ability. Evidence regarding H10 is mixed. While the degree of rounding is significantly ssociated with the different expressions of misclassification, the indicators for nondifferentiation and item-nonresponse are not associated with the misreporting of the censoring status.

Table 4.1: Logistic regression: Error in spell beginnings. Odds ratios and p-values

	(1)		(2)		(3)	
	$\|e\_b\| > 0$ Months		$\|e\_b\| > 1$ Months		$\|e\_b\| > 2$ Months	
	OR	p-value	OR	p-value	OR	p-value
Duration between Spell Begin(Register) and subsequent interview	1.04	0.18	1.04	0.15	0.96	0.07
Indicator: Spell begin register 01/2005	0.04	0.00	0.10	0.00	0.22	0.00
Censored register	0.86	0.58	1.15	0.61	0.88	0.63
Length welfare spell register	1.03	0.13	1.02	0.45	1.04	0.04
Prior Unemployment benefit	2.10	0.01	1.35	0.19	1.30	0.23
Foreign HH-language	1.47	0.07	1.60	0.02	1.79	0.00
Vocational training	0.76	0.11	0.72	0.05	0.69	0.02
Higher vocational training	0.70	0.13	0.61	0.03	0.66	0.07
University degree	1.27	0.48	0.94	0.86	0.76	0.39
Intermediate secondary degree	1.19	0.31	1.15	0.38	1.11	0.51
Upper secondary degree	0.63	0.06	0.67	0.09	0.77	0.26
$35 \leq$ Age < 45	0.67	0.02	0.65	0.01	0.67	0.01
$45 \leq$ Age ≤ 55	0.53	0.00	0.61	0.00	0.64	0.01
Average Entropy	0.85	0.19	0.83	0.13	0.80	0.05
Proportion of rounded variables	1.03	0.92	1.27	0.44	1.63	0.11
Proportion of item non-response	1.67	0.66	3.10	0.31	3.37	0.27
CAPI	1.49	0.02	1.44	0.03	1.63	0.00
N	1 251.00		1 251.00		1 251.00	
Psd. (R^2)	0.28		0.18		0.13	
n (%) (Dep. Var.)	537 (43 %)		450 (36 %)		387 (31 %)	

To summarize the results, we observe that regarding the ability and difficulty, the specific anchor points for both spell beginning and spell end substantially increase the probability to date correctly. The length of the recall period is associated with the occurrence of minor errors further back in time. The language proficiency seem to influence the accuracy only for events further back in time. Older people tend to give more accurate answers for all events. This might be due to higher degrees of social desirability in younger people. Regarding the surrogates for satisficing, rounding behavior seems to be connected to the response quality. Respondents with higher rounding scores show a higher propensity to date inaccurately or misclassify, even if the coefficients do not always cross significance. However, non-differentiation and the degree of item-nonresponse have no significant effect. We find strong mode effects for errors regarding the spell beginning and the spell end. Overall, respondents interviewed face-to-face respond less accurately. This might be explained by social desirability. Respondents interviewed in person could be less inclined to divulge their true history of welfare receipt. This could also cause the positive association between CAPI and the misclassification of current welfare

receipt. However, the mode was not randomly allocated. Hence, the effect might be caused by selection processes in the mode assignment. Respondents harder to reach were more likely to be transferred to the CAPI field.

Table 4.2: Logistic regression: Error in spell ends. Odds ratios and p-values

	(4)		(5)		(6)	
	$\|e\_e\| > 0$ Months		$\|e\_e\| > 1$ Months		$\|e\_e\| > 2$ Months	
	OR	p-value	OR	p-value	OR	p-value
Duration between Spell End(Register) and subsequent interview	0.99	0.79	1.06	0.15	1.07	0.16
Indicator: Spell begin register 01/2005	1.58	0.12	0.76	0.36	1.17	0.65
Censored register	0.03	0.00	0.12	0.00	0.23	0.00
Length welfare spell register	0.97	0.14	1.02	0.40	1.00	0.88
Prior Unemployment benefit	1.54	0.23	0.96	0.91	1.41	0.38
Foreign HH-language	1.24	0.52	0.91	0.80	0.74	0.48
Vocational training	1.45	0.17	1.97	0.02	1.60	0.15
Higher vocational training	1.55	0.20	1.80	0.12	1.24	0.60
University degree	1.48	0.38	1.69	0.30	1.39	0.54
Intermediate secondary degree	0.85	0.51	1.12	0.67	1.12	0.69
Upper secondary degree	0.77	0.43	0.74	0.42	1.10	0.81
$35 \leq$ Age < 45	0.76	0.27	0.73	0.23	0.68	0.20
$45 \leq$ Age ≤ 55	0.57	0.03	0.58	0.06	0.52	0.04
Entry at Seam	0.98	0.98	1.75	0.41	2.51	0.17
Average Entropy	1.06	0.81	0.88	0.52	0.83	0.37
Proportion of rounded variables	2.03	0.11	2.14	0.11	2.47	0.08
Proportion of item non-response	0.03	0.07	0.23	0.47	0.06	0.24
CAPI	1.82	0.02	1.65	0.06	1.20	0.54
N	1 251.00		1 251.00		1 251.00	
Psd. (R^2)	0.39		0.19		0.14	
n (%) (Dep. Var.)	195 (16%)		109 (9%)		82 (6%)	

We do not find support for most of our hypotheses. The occurrence of larger recall errors for welfare receipt seems to be associated with different response processes than the occurrence of smaller errors. An explanation could be that the occurrence of smaller errors is associated with classical predictors like the difficulty of the question and the ability of the respondents, however the occurrence of larger response errors and of misclassifications, could be associated with social desirability.

Table 4.3: Logistic regression: Misclassification of the censoring status. Odds ratios and p-values

	(7)		(8)		(9)	
	Mis-classification: Total		Mis-classification: Overreporting		Mis-classification: Underreporting	
	OR	p-value	OR	p-value	OR	p-value
Indicator: Spell begin register 01/2005	1.14	0.67	2.87	0.11	0.47	0.06
Length welfare spell register					1.09	0.01
Prior Unemployment benefit	1.04	0.93			1.40	0.51
Foreign HH-language	1.16	0.70	1.75	0.37	0.80	0.65
Vocational training	1.07	0.83	0.39	0.13	1.68	0.18
Higher vocational training	0.79	0.60	0.19	0.14	1.41	0.52
University degree	0.47	0.29	0.29	0.30	0.65	0.62
Intermediate secondary degree	0.95	0.87	1.03	0.96	0.97	0.94
Upper secondary degree	1.35	0.46	1.57	0.56	1.57	0.35
$35 \leq Age < 45$	1.09	0.79	0.99	0.99	0.99	0.97
$45 \leq Age \leq 55$	0.87	0.68	1.34	0.66	0.72	0.43
Entry at Seam†	0.66	0.69			0.97	0.98
Average Entropy	0.89	0.60	0.91	0.78	0.94	0.81
Proportion of rounded variables	3.15	0.04	5.61	0.10	3.41	0.07
Proportion of item non-response	0.07	0.30	0.46	0.84	0.02	0.23
CAPI	1.37	0.29	0.62	0.47	1.55	0.19
N	1 251.00		1 251.00		1 251.00	
Psd. (R^2)	0.02		0.07		0.05	
n (%) (Dep. Var.)	63 (5%)		16 (2%)		47 (3%)	

† Variables are perfect negative predictors for Model 9 and are excluded.

4.6.3 Impact on coefficients of time-to-event analysis

Having modeled the probability for and extent of selected survey errors, we follow up on the relevant question, whether the errors are influencing the coefficients of the specific time-to-event model. In table 4.4 four proportional hazard models for the risk of leaving UB II are presented. With regard to the labor market related contents of the model, the results are in concordance with the current state of knowledge regarding the receipt and leaving of UB II (Buhr, Lietzmann, and Voges (2010), Achatz and Trappmann (2011)). A significant negative association is observed for single mothers in relation to the reference group. Couples without children have a higher probability to leave the welfare receipt. Respondents with a migrational history have lower chances of leaving welfare. Second generation migrants have no higher risk than the rest of the population. Health related problems are considerable negative risk factors. Respondents caring for next-of kin have lower chances to exit the receipt of UB II. Having vocational or higher vocational training or an university

degree increases considerably the chances to leave. Respondents in higher age groups have a lower exit probability than the younger reference group.

In order to quantify the impact of the response error and to subsequently disentangle the effect of the error components, the coefficients of the different models are compared. In general, models with the same censoring status (M10 with M12) tend to be more similar than models with the same information for spell length (M11 with M13). The direction of the response bias on the model parameters is not homogenous.

When analyzing the bias for each of the coefficients , the largest difference can be seen for the trait of single parent man (register: $\hat{\beta} = 0.71$, survey: $\hat{\beta} = 0.20$). As this trait is the smallest group in our analysis sample (1.7 %), the bias does not lead to a substantive shift in the interpretation due to the large confidence interval of the coefficient.

However, for two risk factors one could come to slightly different conclusions, if a hypotheses test on a 5% level (2-sided) would be used. Based on the survey model (M11) one would reject the null hypotheses for care for next-of-kin as a factor influencing the exit ($\hat{\beta}_s = -0.75$; 95% CI_s : −1,42 − 0,08), based on the register model (M10) one would retain the null hypotheses ($\hat{\beta}_r = -0.52$; 95% CI_r : −1,19 0,15). The same can be observed for vocational training. One would reject the null hypotheses using survey data ($\hat{\beta}_s = 0{,}38$; 95% CI_s : 0,06 0,69) but would retain the null hypotheses when applying register data ($\hat{\beta}_r = 0{,}24$; 95% CI_r : −0,08 0,57). Thus, the response error causes no attenuation but an overestimation of effect estimates.

For the indicators of the different age groups the impact of recall error can be seen. Younger respondents report less accurately and report shorter spells than older respondents. Keeping the register status for current status of UB-II-receipt constant, one can observe slightly stronger coefficients for age when comparing M10 ($45 \leq \text{Age} \leq 55$: $\hat{\beta} = -0.88$; $35 \leq \text{Age} < 45$: $\hat{\beta} = -0.76$) to M12 ($45 \leq \text{Age} \leq 55$: $\hat{\beta} = -0.93$; $35 \leq \text{Age} < 45$: $\hat{\beta} = -0.79$). But despite the large impact of the age groups on recall quality, the effect on the coefficients of the proportional hazards model is not substantial.

The previously described effects can also be visually assessed in graph A1. Coefficients shift according to the type of included survey information. The direction of the the shift is not homogenous. The bias leads to attenuation and increase of the estimated effects. The width of the confidence intervals is not affected.

Disseminating the error due to recall and misclassification and calculating the extent of the bias for each coefficients (see table A4), we find that for 12 of 15 variables the deviation of the coefficients due to misclassification is larger than the impact of recall error due to misdating the spell defining events. For three of

15 variables, the bias due to recall error exceeds the bias caused by misclassification. Despite the low number of misclassified events in relation to mistimed spells, the influence of misdating seems to be smaller than the influence of misclassification on the coefficients of the proportional hazard model.

Table 4.4: Proportional hazard models: Parameter estimates and 95% confidence intervals for the risk of leaving UB II

	(10) Register			(11) Survey			(12) Censoring Register – length Survey			(13) Censoring Survey – length Register		
	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI
Single man	*ref.*			*ref.*			*ref.*			*ref.*		
Single woman	0.28	-0.13	0.69	0.08	-0.33	0.48	0.25	-0.16	0.66	0.13	-0.27	0.53
Single parent woman	-0.61	-1.02	-0.21	-0.53	-0.90	-0.17	-0.65	-1.05	-0.24	-0.53	-0.89	-0.17
Single parent man	0.71	-0.07	1.50	0.20	-0.63	1.04	0.52	-0.26	1.31	0.40	-0.44	1.23
Couple without children	1.05	0.61	1.49	0.92	0.50	1.34	1.02	0.58	1.46	0.95	0.53	1.37
Couple with children	0.31	-0.06	0.68	0.29	-0.05	0.64	0.28	-0.09	0.65	0.30	-0.04	0.65
Western Germany	-0.01	-0.28	0.26	0.02	-0.24	0.27	-0.02	-0.28	0.25	0.02	-0.24	0.27
1. Gen migration	-0.75	-1.18	-0.32	-0.67	-1.07	-0.27	-0.75	-1.18	-0.32	-0.70	-1.10	-0.30
2. Gen migration	0.16	-0.26	0.57	0.18	-0.22	0.57	0.22	-0.20	0.64	0.14	-0.25	0.54
Health related problems	-0.58	-0.87	-0.29	-0.48	-0.75	-0.21	-0.56	-0.86	-0.27	-0.51	-0.78	-0.24
Care for next-of-kin	-0.52	-1.19	0.15	-0.75	-1.42	-0.08	-0.62	-1.29	0.05	-0.69	-1.36	-0.02
No vocational training	*ref.*			*ref.*			*ref.*			*ref.*		
Vocational training	0.24	-0.08	0.57	0.38	0.06	0.69	0.25	-0.08	0.58	0.40	0.09	0.71
Higher vocational training	0.65	0.26	1.04	0.68	0.31	1.06	0.62	0.22	1.01	0.77	0.40	1.14
University degree	1.06	0.60	1.52	0.93	0.48	1.38	0.97	0.51	1.42	1.08	0.63	1.53
Age < 35	*ref.*			*ref.*			*ref.*			*ref.*		
35 ≤ Age < 45	-0.76	-1.08	-0.44	-0.70	-0.99	-0.40	-0.79	-1.11	-0.47	-0.66	-0.96	-0.36
45 ≤ Age ≤ 55	-0.88	-1.23	-0.54	-0.91	-1.23	-0.58	-0.93	-1.28	-0.59	-0.85	-1.17	-0.52
AIC	3 209.28			3 629.16			3 180.08			3 586.27		
No. of events n = 1 251	240			271			240			271		

4.7 Discussion

Labour-market spells based on survey responses can be affected by response errors. In order to quantify and analyse the effects of the response error, in this study we validated survey responses on welfare receipt with administrative data that was provided by the German federal labor agency. The combination of survey and administrative data provided a rare opportunity to assess factors explaining the occurrence of recall error and investigate their impact on time-to-event analysis.

We found that recall error due to misdating is common and is much more severe for spell beginnings than for spell ends. Spell beginnings seem to be telescoped forward, which causes shorter reported welfare spells. We found that the recall error was correlated with a range of predictors. This is in line with the results of Pyy-Martikainen and Rendtel (2009). Using a wider approach to model response error than in previous studies, it seems that the occurrence of recall error is affected by the availability of temporal anchors and cognitive skills. No clear association can be seen for indicators for satisficing. Only rounding as an indicator for satisficing seems to influences more current events like spell ends. However, we found that age and interview mode affect the recall quality. Older people and respondents interviewed by telephone show a higher recall quality for welfare receipt. It is possible that due to the stigmatizing status of the receipt of UB II (Booth and Scherschel 2010), younger people and respondents interviewed in person are more reluctant to divulge their true history of welfare receipt. Social desirability seems to have greater impact on the quality of the response for welfare receipt than previously suspected.

We could find no significant determinants for the misclassifications of the censoring status. This could be due to the low number of observations. Other studies that used broader analyses samples by Bruckmeier, Müller, and Riphahn (2014, 2015) found structural correlates for the misreporting of current status of welfare receipt.

In this study the response error influence the coefficients of time-to-event model. The direction and strength of the bias are not homogenous across the variables. Our results are consistent with the findings of Pyy-Martikainen and Rendtel (2009) and Pierret (2001). A larger impact of response error could only be determined for coefficients that are based on a small number of cases. Still, for two variables, the response error caused the coefficients to cross significance thresholds. We do not find an overall attenuation of effect estimates as found by Jäckle (2008) and Pina-Sanchez, Koskinen, and Plewis (2013).

The estimation of event history models is based on the spell lengths and the censoring information of the individuals. Disentangling the influence of errors in

the two concepts on effect estimates, we find that the common errors in spell durations caused by misdating have less an impact on the coefficients of the proportional hazard model than the impact caused by rare misclassifications of the censoring status. This finding is relevant for survey practice, as it seems to be more important to measure the current status correctly than to implement measures that might reduce the amount of recall error.

This study has a range of limitations. We could only use respondents that could be linked to administrative data. This is a selective subsample. In the PASS study respondents that are older and report a higher income are more likely to consent (Beste 2011). However, we found similar associations for leaving welfare receipt as in studies using the complete PASS data. The analysis are also restricted to cases with one spell of welfare receipt in each data source. This was necessary to establish a firm definition and quantification for the response error. Having had relaxed the selection criteria, an exact quantification of the response error would not have been possible, since the necessary spell-matching procedures would have added an uncontrollable layer of uncertainty. Thus, the impact of omissions and merging of welfare spells was not assessed and should be the subject of further work. It has been shown that the the interviewer can influence misreporting (Schober and Conrad 1997). We conducted sensitivity checks by recalculating Models 1 to 10 with multilevel logistic regressions and nested the respondents in interviewers. A significant interviewer effect could only be observed in one model (M1) and coefficients in any model remained unchanged. Therefore, we omitted the interviewer level and did not include explanatory variables on the interviewer level.

In this study, we analysed the response error for welfare receipt. This study provides further evidence that in many instances the survey responses for labourmarket events are affected by misreporting and that response errors are differential. Thus, the impact of the errors on statistical analyses remains unknown in the absence of validation data. Researchers should abstain from far-reaching claims from when using survey data.

5 Dependent interviewing and sub-optimal responding[1]

Johannes Eggs and Annette Jäckle

Abstract

With proactive dependent interviewing (PDI) respondents are reminded of the answer they gave in the previous interview, before being asked about their current status. PDI is used in panel surveys to assist respondent recall and reduce spurious changes in responses over time. PDI may however provide scope for new errors if respondents falsely accept the previous information as still being an accurate description of their current situation. In this paper we use data from the German Labour Market and Social Security panel study, in which an error was made with the preload data for a PDI question about receipt of welfare benefit. The survey data were linked to individual administrative records on receipt of welfare benefit. A large proportion of respondents accepted the false preload. This behaviour seems mainly driven by the difficulty of the response task: respondents with a more complex history of receipt according to the records were more likely to confirm the false preload. Personality also seemed related to the probability of confirming. Predictors of satisficing, indicators of satisficing on other items in the survey, and characteristics of the survey and interviewer were not predictive of confirming the false preload.

Keywords: measurement error, validation, record linkage, panel survey, welfare benefit, satisficing

5.1 Introduction

With Proactive Dependent Interviewing (PDI), respondents are reminded of the answer to a survey question they gave in a previous interview, before being asked about their current situation (Mathiowetz and McGonagle 2000). For example, "Last time we interviewed you, you told us that you were working as a pharmacist. Is this still the case?" Dependent interviewing questions are implemented by preloading each respondent's answer from the previous interview into the computerized questionnaire script. Variants of dependent interviewing are nowadays used in most longitudinal panel studies (Schoeni et al. 2013). PDI is commonly used to collect information about labour market status and employment characteristics such as industry and occupation (e.g. in the UK Household Longitudinal Study,

1 This chapter is identical with Eggs and Jäckle (2015).

Current Population Survey, National Longitudinal Survey of Youth 1997 (NLSY97), Health and Retirement Study, English Longitudinal Study of Ageing, Survey of Labour and Income Dynamics). In this paper, we examine the risk that respondents confirm answers from the previous interview, regardless of whether they are accurate or not.

PDI is used for two main reasons (Jäckle 2009). First, PDI questions can be used to determine routing in the questionnaire and to omit redundant questions. For example, if the respondent is still working for the same employer and in the same occupation as at the previous interview, other characteristics of the job may not have to be collected again. Thus, PDI reduces respondent burden, may shorten the interview and facilitates the flow of the interview (Jäckle 2008; Sala, Uhrig, and Lynn 2011). Second, PDI increases the longitudinal consistency of responses across interviews. When questions are asked independently, without reference to previous answers, respondents may for various reasons report a different status in one interview from the next, even if their actual status has not changed (Moore et al. 2009). PDI reduces spurious changes in responses over time, by reducing measurement error in each interview (Lynn et al. 2012).

However, the use of PDI can have disadvantages. Concern is voiced that respondents may falsely confirm a previous status as still applying, as they rely on recognizing the previous information instead of retrieving information from memory (Hoogendoorn 2004). Dependent interviewing could thus lead to spurious stability replacing the original problem of spurious change. Also, inaccurate responses from previous interviews may be confirmed by respondents as still applying, such that errors are carried forward into future interviews (Conrad, Rips, and Fricker 2009). Thus, PDI might provide new sources of measurement error, if respondents falsely confirm information from previous interviews.

In this study we use data from the German panel survey "Labour Market and Social Security" (PASS), where preload information regarding welfare receipt was falsely processed for a subgroup of respondents in one panel wave. We use the survey data linked to individual level administrative records on welfare receipt to address the following questions:

1. To what extent do respondents confirm previous information when that is false? How much of the apparent false confirmation is in fact due to false reporting at the previous wave?
2. What are the mechanisms causing false confirmation?
3. Which socio-demographic characteristics are associated with false confirmation?
4. What are the implications of false confirmation for measurement error?

5.2 Theoretical background on false confirmation

False confirmation, and measurement error in general, is caused by sub-optimal responding (Thomas 2014). Sub-optimal responding occurs if individuals are not sufficiently motivated to invest the necessary cognitive resources to respond optimally, or if other non-motivational factors related to the question design or survey implementation interfere. Errors can occur in any step of the response process described by Tourangeau, Rips, and Rasinski (2000): comprehension of the question and response options, retrieval of relevant information from memory, judgment of the retrieved information to form a conclusion, and formulating a response or selecting a response option.

With proactive DI, the respondent has to compare the information they are reminded of with information retrieved from memory and judge both sets of information. Even if respondents are motivated to provide an accurate response, there are several factors that could lead to false confirmation of previous information. Respondents may fail to understand the question or response options. For example, they may be confused about the type of welfare income they are being asked about. Respondents may have trouble recalling relevant information, which could be because they never encoded the information in memory, or due to memory decay, or they may have difficulty judging the retrieved information against the information they are reminded of. In these cases respondents may believe the information from the previous interview to be correct and therefore confirm it. Finally, respondents may inadvertently select an inaccurate response option.

If respondents are not sufficiently motivated to provide an accurate response, they may satisfice by choosing a cognitive shortcut (Krosnick 1999). There are several satisficing strategies that could lead to false confirmation of previous information. Firstly, respondents may minimize effort by stopping the search for a response at the first plausible endpoint, which is simply confirming the previous information. Alternatively, respondents may be susceptible to a general tendency to agree with, rather than reject, information presented to them (confirmation bias, see Nickerson (1998). Similarly, respondents tend to agree to questions out of an inner impulse or in order to be liked or to avoid a conflict or an argument with the authority respectively the interviewer (acquiescence, see Johanson and Osborn (2004, p. 536) and Tourangeau, Rips, and Rasinski (2000, p. 5)). The likelihood that respondents satisfice by selecting the first plausible response or acquiescing is thought to be higher with respondents who are less motivated to participate in the survey (Krosnick 1999). Respondents with lower cognitive abilities have to invest more mental resources to retrieve and formulate an accurate answer and are therefore also more likely to satisfice (Krosnick 1999). More difficult tasks require more cognitive resources and

thus an increased difficulty also increases the risk of satisficing (Meisenberg and Williams 2008).

The likelihood of acquiescing is also related to personality (Kieruj and Moors 2013) and survey procedures: more experienced interviewers elicited higher rates of acquiescence than inexperienced interviewers in a study by Olson and Bilgen (2011) telephone interviews produced higher levels of acquiescence than personal interviews in a study by Leeuw (2005). The effect of interviewer age and sex is inconclusive (Davis et al. 2010), however respondent age and sex were related to acquiescence in a study by Vaerenbergh and Thomas (2013).

Cognitive ability and task difficulty may also be related to sub-optimal responding among respondents who are motivated to provide accurate responses (Knäuper et al. 1997). Respondents with higher ability may be more likely to accurately remember information about welfare receipt, and find it easier to accurately compare the retrieved information with the information from the previous interview. Similarly, if the task is more difficult, respondents are more likely to have trouble accurately recalling and judging information.

In sum, sub-optimal responding may lead respondents to confirm information from previous interviews even if it is not correct. This could be due to motivational problems or other factors influencing the response process. Overall, we expect the likelihood that respondents falsely confirm previous information to be higher among respondents who are less motivated to provide accurate information, respondents with lower cognitive ability and if the task set by the survey question is more difficult. In addition, we expect that some personal characteristics and characteristics of the survey may influence the likelihood that respondents falsely confirm previous information.

5.3 Previous studies on false conrfimation and misreporting of benefits

The extent to which respondents falsely confirm information presented to them in PDI questions is not known. However there is a previous study that examined responses when the preloaded information was wrong. Aughinbaugh and Gardecki (2008) used data from the NLSY97, where the preload information about receipt of a certain type of welfare income was not drawn from the previous wave interview, but from two waves before. A sub-sample of 610 respondents had reported a different receipt status in the following interview. Thus these respondents were reminded that they had received/not received the welfare income at the date of the previous interview, when in fact they had reported the opposite. The authors found that only one third of these respondents corrected the information presented

to them in the PDI question. Respondents with higher scores on an intelligence measure and respondents who were rated as being more honest by the interviewer were more likely to correct the false preload information. A limitation of this study is that the true status of welfare receipt was unknown. For respondents who had misreported their receipt status at the previous wave, the preload information from two waves earlier was in fact correct and respondents would rightly have confirmed the preload. We use the unique opportunity presented by the combination of an error in preload data and linked administrative records, to identify respondents for whom the preload was truly wrong, to examine their reactions to the preload, and to check the implications for measurement error.

Misreporting of welfare receipt is related to the probability of actual receipt, and thus with a range of socio-economic indicators. In a study by Bruckmeier, Müller, and Riphahn (2014) that used data from the same survey and linked administrative records as we use in this study, recipients that were more like non-recipients were more likely to underreport receipt, than recipients whose eligibility was certain. For example, respondents where another household member was in work or who had higher levels of household savings were more likely to underreport receipt. Respondent characteristics related to misreporting receipt might also be associated with the risk of falsely confirming. We therefore also examine whether the types of respondents who are more likely to underreport receipt, are also more likely to falsely confirm information presented to them in PDI questions.

5.4 The panel survey and validation data

The data for this study are from the German panel survey "Labour Market and Social Security (PASS)". The survey was established to study the impact of major welfare reforms, called the "Hartz reforms" that introduced a new type of welfare scheme called unemployment benefit II (UB II). PASS was designed to assess the dynamics of welfare receipt and to investigate how the welfare reforms influence the social situation of affected households and the persons living in them. PASS was set up as a household survey, since UB II provides economic resources that are means tested at the level of the benefit unit. A benefit unit consists of at least one adult plus their spouse (if applicable) plus any dependent children living with them. A benefit unit is in most cases congruent with the household. The panel study is conducted by the Institute for Employment Research and is funded by the German Federal Ministry for Employment and Social Affairs.

5.4.1 Survey design

In order to compare recipients of UB II with non-recipients, PASS was set up as a dual-frame survey. It consists of a recipient sample and a sample drawn from the general population. The recipient sample was selected from a register of recipients of UB II held by the German Federal Employment Agency. 300 primary sampling units (PSUs) were drawn from postcodes with selection probabilities depending proportionally on the size of the population. Within each PSU, benefit units were drawn. The population sample was based on a commercial database of household addresses, where addresses were sampled within PSUs. The population sample was stratified disproportionately by socio-economic status such that households with low status were oversampled. Subsequently, refreshment samples were drawn every year. The refreshment samples consist of households that are first time recipients of UB II. Sizes of the refreshment samples vary around 1 000 households covering around 1 400 individuals aged 15 years or older.

Prior to the first survey interview, each household receives an advance letter that informs the household about the study and includes a leaflet describing the data security protocol. To collect information about the household, the head of the household is asked to complete a household interview containing among others questions on household composition and receipt of UB II. For the recipient sample the head of the household is defined as the person that applied for UB II. For the population sample, the head of the household is defined as the person that is most familiar with the overall situation of the household. After the household interview, every member of the household aged fifteen or older is asked to complete a personal interview. Proxy interviews for currently unavailable members of the household are not allowed.

PASS uses a mixed mode design whereby data are collected using either computer-assisted telephone interviews (CATI) or computer-assisted personal interviews (CAPI). In wave 1 households were first approached in CATI, non-respondents and households for whom no valid telephone numbers were known were followed up with CAPI. From wave 2 onwards households are first approached in the mode in which they were last interviewed. Refreshment samples are contacted first by CAPI. The first time a household is interviewed, each household member who completes the personal interview receives a conditional incentive of 10 Euros. In subsequent panel waves, the incentive is posted unconditionally together with the advance letter that informs respondents of the upcoming interview. In order to assess socio-economic dynamics, households are interviewed annually. In wave 1 PASS had household response rates of 28.7 % for the recipient sample and 24.7 % for the population sample (RR1 according to The American Association for Public

Opinion Research 2011). For an overview of the PASS panel, see Trappmann et al. (2013).

5.4.2 Administrative data and linkage

The administrative data used to validate survey reports are from the Integrated Employment Biographies (IEB) held by the Research Data Centre of the German Federal Employment Agency. It contains exact start and end dates of all spells of UB-II-receipt. This information is of high quality as it is directly produced by the software that administers benefit claims and payments (Jacobebbinghaus and Seth 2007; Köhler and Thomsen 2009). The IEB is a person level dataset. Spells that refer to a benefit unit are therefore recorded for each person in that unit.

The linkage between PASS survey data and IEB administrative data requires informed consent of respondents. Respondents who have not given consent to data linkage are asked again in the following wave. Among respondents interviewed at wave 4 (the wave we focus on in this study), 81% had given consent to linkage at some point. The recipient sample was selected from the IEB data and therefore linkage was trivial. Respondents in the population sample were linked by their name and address, gender and date of birth using error tolerant procedures based on Jaro (1989).

5.4.3 Dependent interviewing and preload error

The survey uses proactive dependent interviewing to collect information on UB-II-receipt. As UB II is a means tested welfare programme that is paid to households, the information is collected in the household questionnaire. The head of the household is asked:

> *In the last interview in ‹MONTH/YEAR› you stated that the household you were living in then was receiving unemployment benefit 2 at the time. Until when was this benefit received without interruption? Please report the month and the year.*

Dependent interviewing relies on preload information. For this question, the preload is whether or not the household was receiving UB II at the time of the previous interview. When preparing the preload information for wave 4 an error occurred: households that reported a terminated UB II spell, but no current receipt at the wave 3 interview, were coded as still receiving UB II at the time of the interview. In the PDI question these households were reminded that they had received UB II at the time of the interview and asked until when it had continued,

when in fact they had reported that receipt had ended by then. If the respondent said that the preload information was wrong, the spell was treated as having ended at the previous interview date and the respondent was asked whether they had had any other spells of receipt since. That is, respondents were not explicitly asked to confirm the preload, but if they disputed the preload data this was treated as a valid response. We use the expression "confirmed the preload" somewhat loosely to refer to respondents who did not contradict the preload. The preload error occurred for 393 households; 73.7 % from the recipient sample, 11.1 % from the population sample, 15.2 % from the refreshment samples. These households form the base for our analyses and 354 were successfully linked to administrative data.

5.5 Predictors of sub-optimal responding

In section 2 we argued that sub-optimal responding is related to the cognitive ability of the respondent, the difficulty of the response task, the motivation of the respondent and acquiescence. The following is a discussion of the indicators we use for each of these dimensions.

As proxy measures for *cognitive ability* we use education and age. We expect respondents with higher education to be less likely to confirm the false preload and therefore use a dichotomous indicator that is set to one if the respondent holds an intermediate or higher degree. As cognitive ability decreases with age we also expect older respondents to be more likely to falsely confirm the preload. However as our study sample consists of individuals below 67 (the age cut-off for UB II eligibility) we expect this association to be weak.

How *difficult* the task of reporting on UB-II-receipt is for the respondent depends on the complexity of their history of receipt. Respondents who have had multiple spells of receipt will find it more difficult to accurately recall details of any one particular spell (Eisenhower, Mathiowetz, and Morganstein 1991). The administrative records of 354 households could be used to derive two indicators of the complexity of the respondent's history: the number of spells of UB-II-receipt and the elapsed time since receipt ended. We use the number of welfare spells for the time period of 12 months around the date of the wave 3 interview. We expect that the number is positively related to the confirmation of the false preload, as the increased complexity of the respondent's history makes it more likely that the respondent will make errors in recalling information or that they will not expend the necessary cognitive resources to accurately assess the possibility of welfare receipt at the time of the last interview. The elapsed time measures the time between the end of the last spell of UB-II-receipt and the date of the last interview. We expect that the elapsed time is negatively associated with the false confirmation. If more

time has passed between welfare receipt and interview date, it should be easier for the respondent to remember correctly, whether welfare was received at the time of the last interview. Thus respondents should be less likely to make errors of recall and judgment and should need less effort to report accurately, reducing the probability of sub-optimal responding.

We further use interviewer observations as proxies for the combined effect of respondent cognitive ability and difficulty of the response task (questions in the appendix). Interviewers were asked on a 5-point scale, whether the respondent had difficulty remembering dates. The variable was coded as 1 if the interviewer judged that the respondent had difficulty or strong difficulty remembering dates. We expect the interviewer judgment of whether respondents had difficulty recalling information to be positively associated with confirming the false preload, as respondent difficulty could be due to low cognitive ability or a complex history of receipt, or both, which would increase the likelihood of suboptimal responding. The indicators of ability (education, age) and complexity of the respondent's history (number of spells in the records, elapsed time since end of receipt) are correlated to some extent with the interviewer assessments of whether the respondent had difficulty recalling dates of events. The largest correlation is between difficulty dating events and education (-0.14, $p = 0.02$), suggesting that the interviewer observations do measure additional aspects related to ability and difficulty.

The *motivation* of respondents is measured by observations made by the interviewer, as well as indicators of satisficing on other items in the survey (Hoogendoorn 2004). Interviewers were asked on a 5-point scale whether they believed that the respondent was interested in the interview. The variable was coded as 1, if the respondent had shown no or little interest. A similar strategy was chosen by Aughinbaugh and Gardecki (2008). We further use the amount of rounding, non-differentiation and don't know/refused answers by the respondent, which are commonly used proxies for satisficing (Krosnick et al. 2002). Dichotomous indicators are formed that were coded as 1 if the respondent rounded in more than 50% of the numerical questions in the household questionnaire (where a response was classified as rounded if it was a multiple of 50 euros), used constantly the same response option in at least one of three item batteries, respectively had more than 1% "don't know/refused" answers in the personal questionnaire. On average each respondent received six numerical questions and 103 questions in the survey. We expect low motivation and the indicators of satisficing on other items to be positively associated with confirming the false preload.

Additional indicators related to *acquiescence* include personality traits and characteristics of the survey and interviewers. Acquiescence is related to

agreeableness (Knowles and Nathan 1997). Agreeableness is one dimension of the "'Big Five"' personality traits. The Big Five are broad dimensions that depict the range of personalities (John and Srivastava 1999). The personality traits are measured by a German version of the Big Five item battery (Rammstedt and John 2005). These dimensions are the traits of extroversion, agreeableness, conscientiousness, neuroticism, and openness. Factor scores were calculated for each dimension via confirmatory factor analysis in line with Rammstedt and John (2005). We focus on agreeableness and expect that a higher agreeableness score is positively related with confirming the false preload. The Big Five item battery was only measured one wave after the preload error occurred. It has however been argued that acquiescence is a stable personality trait (Kieruj and Moors 2013). Hence, the later data collection should not distort the hypothesized relation between the constructs. However, cases are lost due to panel attrition from wave 4 to wave 5. Survey and interviewer-specific characteristics can also influence acquiescence. We expect telephone interviewing (versus face-to-face) and interviewer experience to be positively associated with confirming the false preload.

Finally, previous research has shown that the risk of measurement error in reporting welfare receipt is associated with socio-economic factors (Bollinger and David 1997; Bruckmeier, Müller, and Riphahn 2014). Bruckmeier, Müller, and Riphahn (2014) showed that women, singles, younger individuals, individuals in higher income categories, with larger amounts of savings and shorter spells of welfare receipt were more likely to misreport. They concluded that respondents that were less likely to receive welfare were more likely to underreport. The authors also used data from the PASS panel survey. Hence, we derived similar indicators as in this earlier study to test whether the indicators related to underreporting are also associated with the risk of confirming the false preload.

5.6 Results

To what extent do respondents confirm previous information when that is false?
For our analyses we use the interviews of 393 heads of households, who at wave 4 received a question with false preload information regarding their welfare receipt at the time of the wave 3 interview. Of these, 30.1 % contradicted the interviewer, stating that the preloaded information was false. That is, 69.9 % of respondents did not correct the preload. Instead they either reported that the spell had ended between the wave 3 interview and the wave 4 interview (46.8 %), or was still ongoing at date of the wave 4 interview (17.8 %), or that the spell had ended and a new one had started (5.3 %).

How much of the apparent false confirmation is due to false reporting at the previous wave?

All respondents included in our analysis sample reported at the previous interview that they were not currently receiving UB II. However, some of these respondents may have underreported receipt. Welfare receipt can be considered a sensitive item that is generally underreported in social surveys (Bound, Brown, and Mathiowetz 2001). In the PASS survey, welfare receipt is underreported by about 10–15 % (Kreuter, Müller, and Trappmann 2010). Therefore for some respondents in our analysis sample, the apparently false preload indicating receipt at the time of the previous interview may in fact have been correct and these households would have been correct in confirming the preload. We can identify households that underreported welfare receipt at the previous interview using the register data. Table 5.1 documents the extent to which respondents confirmed the preload, by whether the preload was in fact correct. Of the 354 households that could be linked, 74 (20.9 %) had received UB II at the time of the last interview according to the record data. That is, their preload indicating receipt was in fact correct. Of these households 68 (91.9 %) confirmed the preload and only a minority continued to underreport. In contrast, among the 280 households where the preload really was wrong, only 64.3 % confirmed the preload. The probability of confirming the preload was therefore significantly higher if the preload was in fact correct ($P < 0.001$). However of the overall confirmation rate of 70.1 %, only 19.2 percentage points were due to respondents who underreported receipt at the previous wave (calculated as the probability of confirming, conditional on the preload being correct, multiplied by the probability of the preload being correct: .919 x .209 = .192). The remaining 50.9 percentage points were respondents who confirmed a preload that really was wrong robability of confirming, conditional on the preload being wrong, multiplied by the probability of the preload being wrong: .643 x .791 = .509). The high proportion of respondents who confirmed the preload is thus mainly driven by false confirmation rather than misreporting at the previous wave.

What are the mechanisms causing false confirmation?

For the subsequent analyses we focus on the 280 households where the preload really was wrong according to the records, and respondents were reminded of receipt when in fact they had not been receiving UB II at the date of the previous interview. The descriptive statistics for this subgroup are shown in appendix tables A5 and A6.

Table 5.1: Probability of confirming preload, by whether preload was correct

Validation against records:	Confirmed preload					
	Yes		No		Total	
	n	row %	n	row %	n	col %
Preload correct	68	91.9	6	8.1	74	20.9
Preload wrong	180	64.3	100	35.7	280	79.1
Total	248	70.1	106	29.9	354	100.0

Notes: $\chi^2 = 21.84$, $P < 0.001$

To test which mechanisms might explain why respondents confirm false preload information, we first test the bivariate associations between each of the predictors of sub-optimal responding (as discussed in section 5.5) and the probability of confirming the preload (tables 5.2 and 5.3). We split continuous variables at the mean or into quintiles and use χ^2–tests to test for significant associations. We then estimate multilevel logistic models and calculate average marginal effects for the probability of confirming the false preload (table 5.4). The 280 respondents are nested in 170 interviewers; 79 of the interviewers conducted only one interview with a respondent from the analysis sample, while 91 interviewers conducted two or more interviews. We include the interviewer level to estimate standard errors of interviewer level variables appropriately; we do however not interpret interviewer effects, due to the small number of respondents per interviewer. As the Big Five personality traits were collected a year after the preload error, and hence some observations are lost to attrition, we estimate separate models excluding (Model 1 in table 5.4) and including the Big Five traits (Models 2 and 3 in table 5.4).

Our measures of respondent *cognitive ability* were not significant predictors of the probability of confirming the false preload. While there was a tendency for respondents with lower education to be more likely to confirm the preload than respondents with higher education, this difference was not significant in the bivariate tests (table 5.2) or in the logistic regression models (table 5.4). Similarly, while there were some differences between age quintiles in the probability of confirming (table 3), there was no clear pattern in the effects and the probability of confirming did not appear to increase with age as expected.

The measures of task difficulty derived from the administrative records were strong predictors of the probability of confirming the false preload. Respondents with two or more spells of UB-II-receipt in the 12-month window around the wave 3 interview were 24 percentage points more likely to confirm the false preload according to the bivariate test (table 5.2, $p = 0.002$), than respondents with one or no spell. Controlling for other characteristics, the average marginal effect estimated from the logistic regression model (Model 1 in table 4) suggests that each additional spell of UB-II-receipt increased the probability of confirming the false preload by

18.1 % ($p < 0.001$). Similarly, respondents for whom the length of time between the end of the last UB II spell and the date of the wave 3 interview was shorter than the average of 6.8 months, were 20.3 percentage points more likely to confirm the false preload according to the bivariate tests (table 5.2, $p < 0.001$) than respondents whose elapsed time was longer than average. Examining the probability of confirming the preload by quintiles of the elapsed time shows a clear linear relationship (table 5.3): the probability of confirming was highest amongst those where the elapsed time was only 1 to 3 months (80.4 %), and monotonically fell to 48.2 % among the group with the longest elapsed time of 14–39 months ($p = 0.002$). These results are confirmed by the estimates from the logistic regression (Model 1 in table 5.4) according to which each additional month between the end of the spell and the date of interview decreased the probability of confirming the preload by 1.1 % ($p < 0.05$).

Table 5.2: Percent confirming false preload, by predictors of sub-optimal responding (binary predictors)

	Value of binary predictor		Test of proportions	
	0	1	(p-value)	n
Higher education	68.9	59.9	0.122	276
Respondent age > 55	63.8	63.8	0.996	276
Number of UB II spells in records > 2	60.4	84.4	0.002	280
Months since last UB-II-receipt in records > 6.8 months	76.1	55.8	0.000	280
Difficulty dating events	62.7	75.0	0.322	260
Interview not interesting	61.5	64.7	0.591	258
Rounding in more than 50 % of questions	66.3	55.4	0.108	276
Non-differentiation in 1+ item batteries	64.8	61.9	0.627	276
Item non-response > 1 %	63.2	67.6	0.605	276
CAPI (No = CATI)	61.9	69.8	0.202	280
Female interviewer	72.0	58.6	0.021	280
Interviewer experience > 3 months	68.1	59.2	0.122	280
Agreeableness score > 0	64.2	67.0	0.675	203
Extraversion score > 0	65.0	66.0	0.879	203
Openness score > 0	70.1	63.2	0.313	202
Neuroticism score > 0	60.5	71.9	0.090	203
Conscientiousness score > 0	66.7	64.5	0.744	203

Notes: Continuous variables split at the mean.

The interviewer assessment of whether the respondent had difficulty recalling dates of events was not significantly associated with the probability of confirming the preload. While respondents who were judged to have had difficulty tended to be more likely to confirm the preload (table 5.2) the difference was not significant and not confirmed by the logistic regression.

Table 5.3: Percent confirming false preload, by predictors of sub-optimal responding (continuous predictors by quintiles)

	Percent confirmed preload	n	χ^2 (p-Value)
Respondent age 20–32	64.3	56	
Respondent age 33–40	53.6	56	
Respondent age 41–48	60.3	63	
Respondent age 49–56	80.0	50	
Respondent age 57–67	62.8	51	0.073
1–3 months since last UB-II-receipt in records	80.4	56	
4–6 months since last UB-II-receipt in records	75.0	56	
7–8 months since last UB-II-receipt in records	60.7	56	
9–13 months since last UB-II-receipt in records	57.1	56	
14–39 months since last UB-II-receipt in records	48.2	56	0.002
Interviewer experience 1–2 years	64.9	94	
Interviewer experience 3 years	72.7	66	
Interviewer experience 4 years	70.0	40	
Interviewer experience 5 years	51.4	35	
Interviewer experience 6+ years	55.6	45	0.156

The indicators of respondent motivation were also not associated with the probability of confirming the preload. According to the bivariate tests (table 5.2) and the regression estimates (table 5.4) there were no differences in the probability of confirming regardless of whether or not the interviewer judged that the respondent had shown little interest in the survey, and whether or not the respondent had rounded, non-differentiated or given don't know or refusal responses to other items in the questionnaire.

Characteristics of the interviewer and survey that may be related to acquiescence were also not associated with the probability of confirming the false preload. Although respondents interviewed by men were 13.4 percentage points more likely to confirm the preload than respondents interviewed by women in the bivariate tests (table 5.2, $p = 0.021$), interviewer sex was not significant in the logistic regression model (table 5.4). Interviewer experience was not related to the probability of confirming the preload in the bivariate tests (whether split at the mean in table 5.2, or split by quintile in table 5.3) or in the logistic regression model. The survey mode was also not significantly associated with the probability of confirming in any of the tests.

Table 5.4: Average marginal effects of random effects logistic models for confirming false preload

Pr (confirmed false preload)	(1)		(2)		(3)	
	AME	se	AME	se	AME	se
Higher education	-0.126	0.077	-0.079	0.103	-0.087	0.109
Respondent age	0.004	0.003	0.004	0.004	0.004	0.004
Number of UB II spells in records	0.181***	0.043	0.155**	0.056	0.155**	0.056
Months since last UB-II-receipt in record	-0.011*	0.005	-0.015*	0.007	-0.015*	0.007
Interviewer: diffculty dating events	0.207	0.159	0.551	0.297	0.595	0.313
Interviewer: interview not interesting	0.049	0.074	0.023	0.090	0.010	0.092
Rounding in > 50% of questions	-0.075	0.085	-0.039	0.113	-0.025	0.112
Non-differentiation in 1+ item batteries	0.030	0.104	-0.054	0.144	-0.093	0.149
Item non-response > 1%	-0.008	0.077	-0.003	0.107	0.020	0.109
CAPI (vs. CATI)	0.113	0.097	0.010	0.141	0.028	0.137
Female interviewer	-0.101	0.088	-0.198	0.123	-0.190	0.119
Interviewer experience in years	0.004	0.020	0.015	0.027	0.013	0.026
Agreeableness			0.139*	0.068	0.108	0.067
Extroversion					0.015	0.058
Openness					-0.019	0.061
Neuroticism					0.094	0.061
Conscientiousness					-0.052	0.078
Rho	0.39		0.56		0.57	
Observations	242		177		176	
AIC	353.4		265.3		281.4	
Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$						

Respondent personality was associated with the probability of confirming the false preload. In the bivariate tests, where the indicators for personality traits were dichotomized at the mean, agreeableness was not associated with the probability of confirming. However controlling for other characteristics and including the agreeableness score as a continuous variable in the logistic regression (Model 2 in table 5.4), each additional point on the agreeableness score (which ranged from -1.70 to 1.61) increased the probability of confirming the false preload by 13.9% ($p < 0.05$). However when the other four personality traits, extroversion, openness, neuroticism and conscientiousness, were included in the model (Model 3), none of the traits were significant predictors of confirming the preload. As the Big Five measures were collected in a subsequent wave and cases were lost due to attrition, we estimated additional models using the independent variables from Model 1 and the estimation samples from Models 2 and 3 to check for selectivity in the results due to attrition. There were no relevant shifts in the results (not shown) and thus we assume that the results are robust to the sample selection due to attrition. Using the estimation sample for Model 3 to rerun Models 1 and 2 further suggests that model fit did not improve much by adding personality traits: the Akaike

Information Criterion (AIC) changed from 263.0 in Model 1 to 264.0 in Model 2 when agreeableness was added, and to 281.4 when the remaining Big Five traits were added. This suggests that while personality may have had some effect leading to acquiescence, the complexity of the respondent's history was the main driver of confirming the false preload.

In sum, while a large proportion of respondents confirmed the false preload, this behaviour does not seem to be driven by lack of respondent motivation. Indicators of motivation and satisficing on other items, and indicators of respondent cognitive ability were not predictive of confirming the preload. Instead, the measures of the difficulty of the response task derived from administrative records were strong predictors: those respondents for whom the task of recalling information about any one particular spell would have been more difficult were more likely to confirm the false preload. In addition, respondents who scored higher on agreeableness were more likely to confirm the false preload.

Who is at risk of falsely confirming previous information?

Previous research has shown that specific socio-economic groups are more likely to misreport their welfare receipt status. Using the indicators that predicted underreporting of UB-II-receipt in the study by Bruckmeier, Müller, and Riphahn (2014), we tested whether the same factors increased the risk of confirming the false preload. The predictors included the respondent's sex, age, whether they had a disability, whether they were an immigrant, education, household type, whether anyone in the household was in regular employment, monthly net household income, value of household savings, whether they owned their home, the number of months of receipt of UB II over the life of the panel, and location (East or West Germany). We estimated multilevel logistic regression models for the probability of confirming the false preload. Respondents were nested in interviewers. The results are presented in table A7 in the appendix. The results show no significant associations between the socio-economic indicators and confirming the false preload. The confirmation of false preloads therefore seems to be driven by different factors than underreporting of receipt.

What are the implications of the respondent behaviour for measurement error?

The false preload reminded respondents that they had received UB II at the time of the previous interview, although at the time they had reported that receipt had ended. For respondents who confirmed this false information, the error in receipt status may therefore be carried over to the current interview. We therefore also examine what impact confirming the false information from wave 3 had on measurement error in receipt status at the time of the wave 4 interview (table 5.5).

Among all respondents for whom the preload error was made, the wave 4 receipt status was wrong for 11.2%. As expected, the error rate was higher for respondents who confirmed the preload (14.4%), than for those who did not confirm (3.8%). We would expect the error due to confirmation of the false preload to mainly consist of overreporting: respondents who confirmed the false information that they were receiving UB II at the time of the wave 3 interview would be likely to overreport receipt at the wave 4 interview. Surprisingly however, while 29 of the misreporters overreported receipt, 10 underreported. That is, these respondents reported that the wave 3 spell had ended and failed to report a subsequent spell that, according to the records, was ongoing at the time of the wave 4 interview. A second surprising result is that the error rates were lower when respondents who had misreported their wave 3 status, such that the preload was actually correct, were excluded. Excluding these cases the wave 4 status was wrong for 8.0% of respondents, with all but one being respondents who had confirmed the preload. This suggests that respondents who misreported at wave 3 were likely to again misreport at wave 4. In sum, respondents who confirmed the false preload were more likely to report their wave 4 status with error, than respondents who did not confirm the preload.

Table 5.5: Impact of confirming preload on measurement error

		Error in wave 4 status	
	Preload confirmed	Percent	n
Preload error (N = 354)	Yes	14.4	243
	No	3.8	106
	Total	11.2~	349
Preload wrong according to records (N = 280)	Yes	12.0	175
	No	1.0	100
	Total	8.0†	275

Notes: 5 households missing due to don't know/refusal answer about current receipt.
~ 10 underreporters and 29 overreporters; † 9 underreporters and 13 overreporters.

5.7 Discussion

One of the main concerns against using proactive dependent interviewing is that reminding respondents of an answer they gave in a previous interview, before asking about their current status, offers respondents the opportunity to satisfice: respondents might say that the previous answer still applies, regardless of whether their situation has in fact changed. If respondents falsely confirm previous information as still applying, PDI may lead to underreporting of change.

In this study we present novel evidence on the risk that respondents confirm false information from previous interviews. We use a unique data source combining

responses from a panel survey, where the preload data for a PDI question contained errors, with linked individual-level administrative records. We exploit the linked administrative records to identify measurement error in the survey reports, and to derive indicators not affected by measurement error that describe the respondent's history. Using the combined data we examine the extent to which respondents confirm the false preload, which mechanisms lead respondents to confirm, and the implications for measurement error.

While a large proportion of respondents confirmed the false preload, this behaviour seems mainly driven by recall difficulties among respondents with complex histories, rather than by satisficing behaviours. Overall, 69.9 % of respondents confirmed the preload. However using the linked administrative data we were able to identify that the preload, that mistakenly reminded respondents of UB-II-receipt at the time of the previous interview, was in fact correct for some respondents who had underreported receipt at the previous wave. Respondents for whom the preload was in fact correct were more likely to confirm the preload, than respondents for whom the preload really was wrong. Nonetheless, the confirmation rate among respondents where the preload really was wrong was still high at 64.3 %.

To examine the mechanisms that lead respondents to confirm the false preload, we tested a range of factors that are related to sub-optimal responding. Our results suggest that the confirmation bias was not related to respondent motivation or ability: the probability of confirming the false preload was not related to interviewer observations of respondent interest in the survey, indicators of satisficing on other items in the questionnaire, age, education, or interviewer observations about whether the respondent had recall difficulties. The probability of confirming was also not associated with characteristics of the survey and interviewer (sex, experience and mode of interview) that were related to acquiescence in other studies. Instead the complexity of the respondent's history of welfare receipt was a strong predictor of confirming the false preload. Respondents who, according to the administrative records, had had a larger number of spells of receipt, or for whom the spell had ended close to the date of the previous interview, were more likely to confirm the false preload. This suggests that respondents who would have had difficulty recalling information about any one particular spell were more likely to think that the preload information was plausible and therefore confirm it. The respondent's personality also appeared to have an effect: respondents who scored higher on the agreeableness score were more likely to confirm. However the effect disappeared once other personality traits were controlled for.

The finding that interviewers' assessments of the respondents' motivation and cognitive difficulties were not associated with the probability of confirming

could in part be due to measurement problems with the interviewer observations. Previous studies have found mixed results as to the usefulness of interviewer observations. For example, Feldman, Hyman, and Hart (1951) found little reliability in interviewer assessments of respondents' intelligence. However, Aughinbaugh and Gardecki (2008) found that respondents rated as being more honest were more likely to correct the false preload and Barret, Sloan, and Wright (2006) found that interviewer assessments of the respondent's cognition was positively related to several indicators of data quality.

We found no associations between the probability of confirming the false preload and socio-economic characteristics that are commonly associated with underreporting welfare receipt. This suggests that underreporting and confirmation of false preload information are driven by different mechanisms: respondents who are similar to non-recipients in their socio-economic characteristics are more likely to underreport receipt (Bruckmeier, Müller, and Riphahn 2014). This is akin to the common result that those who overreport voting tend to have characteristics similar to voters (Ansolabehere and Hersh 2012). Confirming false preload information however seems to be driven by the complexity of the respondent's history that makes it difficult to report accurately. That is, confirmation is not driven by factors related to group identity or membership, but by the respondent's actual experiences. Testing for links between respondent experiences and reporting errors requires exogenous information about experiences that are not themselves affected by reporting error. We were fortunate to have access to the administrative records as an exogenous source of information about respondents' histories.

We also used the administrative records to examine the implications of confirming the false preload for measurement error. While a majority confirmed the false preload, the current receipt status was wrong for only 11.2 % of respondents who had confirmed. The error rate was higher among respondents who had also misreported their status at the previous interview.

Our study has several limitations that threaten the internal and external validity. First, there are sizable intra-interviewer correlations in the probability of confirming the false preload. However, as the maximum number of interviews per interviewer is seven and a large number of interviewers conducted only one interview in the analysis sample, a meaningful interpretation of the intrainterviewer correlation is not feasible (Hox 2010). The interviewer effects might also be confounded with area effects for CAPI, although only 41 % of all cases were interviewed via CAPI. Second, the results are specific to a sample who had recently reported welfare receipt. Although the preload information was wrong, it was plausible for these respondents, which may explain the high rates of confirming. In Aughinbaugh and Gardecki (2008) study, where preload errors were also made for respondents who had reported

receipt in either of the previous two interviews, the confirmation rates were similarly high. For non-recipients a false preload indicating receipt would be implausible and they would be less likely to confirm it as a response. Confirmation rates are therefore likely to be much lower in general population samples. Investigating the risks of false confirmation in a general population sample would ideally require an experimental design allocating randomized preloads to respondents, where the responses and preloads can be linked to administrative records. Third, the results are specific to the German welfare programme UB II, to the question wording and the reference period. Nonetheless Aughinbaugh and Gardecki (2008) reported similar confirmation rates for a different outcome and with different question wording and reference period. Fourth, individuals that agreed to the record linkage are a selective subsample (Beste 2011). In the PASS study respondents that are older and report a higher income are more likely to consent.

In sum, our study suggests that respondents do not react to the information presented to them in PDI questions by satisficing. The gains achieved by PDI in reducing underreporting are likely to outweigh the potential costs of false confirmation. This corresponds to conclusions drawn by Lynn et al. (2012) who experimentally contrasted dependent interviewing with independent questions on benefit receipt, where the responses were also linked to administrative records. Their results showed that PDI substantially reduced underreporting, but did not increase overreporting of receipt. Our results nonetheless reinforce the need for strict quality control of preload answers used for dependent interviewing questions.

6 Concluding remarks

Survey data can be influenced by a range of errors than can distort the reliability and validity of the results. A prominent framework to assess errors in surveys is the Total Survey Error framework. One of the prominent error sources is measurement error. Measurement error can impact the quality of survey data. This is especially true for longitudinal data, as measurement error can influence measures of change, which is one of the primary reasons for conducting a panel survey. In contrast to most of the previous studies on this topic, the set of studies of this thesis focused on the longitudinal aspects of measurement error. All studies analyse the measurement error for unemployment benefit II for the data of a German panel study. To define measurement error, the survey information for UB II was linked on subject level with the entries from administrative records for UB-II-receipt. This allowed the assessment of the measurement error for the first five panel waves of PASS. Thus, it was possible to conduct validation studies for larger time spans than in any previous research.

The first study focused on the development of the measurement error over time. A significant decrease for the underreporting was found over subsequent panel waves, while the degree of overreporting was found to be stable over time. This is partially caused by selective attrition, as misreporters are more likely to refrain from the panel participation. Still, respondents were less likely to underreport in later panel waves and thus data quality increases over time. This can also be found for the length of benefit receipt. While in earlier waves respondents report shortened times of welfare receipt, in later waves the times in receipt are almost completely reported.

The second study focused on the effects of the measurement error on fixed-effects models. The research question is directly linked to the results of the first study, in which a decrease of underreporting over time was found. This decrease of measurement error causes false transitions into UB-II-receipt, as previous underreporters report a false uptake of UB II. For the analysis sample, close to 50 % of all transitions into UB II could not be validated and it was found that common assumptions regarding longitudinal measurement error are violated. When analysing the effect of measurement error on the effect estimates of fixed-effects models, it was found that measurement error caused a substantive overestimation of the effect of UB II on subjective health for men. The measurement error biased the effects away from the zero for men. For women, the measurement error did not severely bias the model coefficients. When implementing simple models to correct for the error, it was found that no approach reduced the bias for both the male and female subsamples.

The third study focused on the influence of measurement error on spell-defining events. Thus, it focused not only the measurement error for current receipt, but also on the quality of recall for past time points that define the beginning and end of a welfare spell. For the UB-II-spells, it was found that the respondents telescope spell beginnings to the present and telescope spell ends into the past. Thus, the resulting survey spells are significantly shorter than their respective counterparts found in administrative data. When modeling the found recall error, significant associations were found for variables related to the amount of elapsed time between time of event and time of interview. Regarding the analysis of the misclassification of the censoring status of a given spell, no clear results were found. When analyzing how the errors affect the coefficients of a time-to-event model, it was found that the coefficients were not biased considerably. When disentangling the contribution of the two types of error to the bias, it was found that the misclassification of the censoring status contributed the major part to the bias, while the recall error contributed the minor part. As the entry at the seam is one of the reasons for the reporting of shorter welfare spells, this finding relates back to the results of study one. The errors in the beginning caused by underreporting in early waves are less important than the measurement error for the last known status.

In the first study, it was argued that dependent interviewing is one reason, why measurement error decreases over time. However, dependent interviewing might also increase measurement error, if false information is taken forward from one wave to the next. This question was analysed in the fourth study. In wave four of the PASS study, a non-random subset of respondents was given a question that stated that they received UB II at the time of the last interview, when in fact they had stated that they did not. Using the administrative record to check their actual status of receipt for the time of the the last interview, it was found that the majority confirmed the preload, even if they did not receive UB II at the time of the last interview. A substantial proportion of those respondents even prolonged their receipt up to the current interview. Thus, it seems to be likely that with dependent interviewing respondents are likely to take forward false reporting. However, the best predictors for this behaviour were derived from administrative records. If a respondent had more UB-II-spells and the closer a respondent's UB-II-spell ended to the time of interview, the more likely a respondent was to confirm the false preload.

Results from all four studies suggest that measurement error for welfare receipt is highly differential. The measurement error for welfare receipt is not based on a random process, but based on personal characteristics and the welfare histories of the respondents. Results from the first, second and fourth study provide further evidence that non-recipients that are more similar to UB-II-recipients

are more likely to overreport and that UB-II-recipients that are more similar to non-recipients are more likely to underreport. For underreporting, similar results were found by Bollinger and David (2005) and by Bruckmeier, Müller, and Riphahn (2014). For overreporting, the studies provided first detailed analyses regarding the correlations between personal characteristics and overreporting.

The results provide a direct loop back to substantive research when analyzing the impact of welfare receipt as the established pattern of misreporting can cause an overestimation of the difference between recipients and non-recipients of UB II. Such an overestimation of the difference can be seen for the results of the first and second study of this thesis. The results of the first study also imply that the overestimation does not necessarily decrease over panel waves. However, while an overestimation seems to be likely, it is not necessarily the case as can also be seen for the results of study two.

It has been criticized that the effect of survey errors on regression models that are used for substantive research is not analysed enough (Bound, Brown, and Mathiowetz 2001; Groves and Lyberg 2010). In this thesis, the impact of measurement error on model coefficient was evaluated exemplarily for two popular classes of regression models for longitudinal data. It is possible that the impact might be completely different using other models or just other model specifications. The results of the studies provide hence no information, how the measurement error might influence the results of other models or different model specifications. Still, this information is valuable, as it provides additional evidence how the measurement might influence results of analyses in general. The studies also raise the awareness that the measurement error does not follow classical assumptions and that it might bias the results of research in any direction. It has to be mentioned that the results of this thesis do not imply that the survey responses of the UB-II-receipt can not be analysed using PASS. PASS is the only study that provides a sufficient number of cases for relevant subgroups of UB-II-recipients and this results also suggest that substantive bias does not necessarily arise.

I would suggest that a range of sensitivity checks should be conducted using different models or variable specifications when analysing welfare receipt with survey data. This applies in particularly when analysing transitions into UB-II-receipt that were shown to be artefacts in many instances. Transitions out of UB-II-receipt are reported with less error. As misreporting mostly affects the first panel waves, one possible check would be, if one recodes the complete recipient sample as being recipients at the time of the first interview. If sensitivity checks are not possible, at least one should mention that measurement error might have influenced the results and keep a healthy skepticism when interpreting the outcomes of empirical models. This applies not only for the use of PASS data. As

welfare receipt has been shown to be misreported in any survey that analysed the reporting of welfare receipt, it is hard to fathom that other German surveys are not affected by misreporting. Using only administrative records is no feasible alternative, as the records contain only a limited set of variables and the quality of the administrative records is not uniformly high.

The results of these studies indicate that the measurement error is not only differential, but also that error patterns differ between sub-populations. Similar conclusions were stated by Kim and Tamborini (2014) for the measurement error in the reporting of income and by Kyyrä and Wilke (2014) for unemployment.

These findings have repercussions for measurement error models that aim to decrease the bias. They rely on assumptions about the error distribution and are mostly applied on the complete analysis sample. As the assumptions are violated and the error distributions differ between groups of respondents, the use of models might even increase the bias. Thus, for variables like welfare receipt, where the error is not caused by a random process, the use of measurement error models does not seem to be a feasible approach.

The results of this thesis also provide information regarding the administration of longitudinal surveys. The results show that the implementation of asymmetric dependent interviewing can help to reduce measurement error over time. Dependent interviewing can also cause measurement error, if it is not properly administered.

6.1 Further research

In this thesis and in prior studies, correlation structures for the measurement error were determined. A better understanding of the error-generating process could help finding ways and means to reduce the error. However, the underlying causal pathway for the misreporting of the respondent is still not clear, even if the welfare history and the closeness to the labour market seem to play a role. While a rich welfare history might increase the difficulty to report correctly, the correlation of labour market closeness is harder to explain. This could be explained by the influence of psychological traits, response behaviour and social stigma. The association of these factors with misreporting has not been studied in detail. Such information could be used to reduce misreporting by altering the administration of the survey. Possible approaches could be changes in the wording of the question or even the use of register information in the question to prime the respondents. However, the last approach would have to be carefully weighted regarding possible concerns regarding data protection and survey participation.

In this thesis, the measurement error for welfare receipt was assessed using data from up to five consecutive panel waves. It is planned that the linked data

will be made available for additional panel waves. This increases the potential of the data even further. One possible research topic could be the comparison of the measurement error for the entry samples over time. The entry samples are similar in their patterns regarding misreporting for the time of the first interview. However, it would be interesting to see, how similar to the results of the first study the evolvement of the measurement error over time will be.

Additionally, a re-analysis of the studies of this thesis could provide evidence about the robustness of the results. With the availability of additional linked panel waves, it would be possible to drop early panel waves and conduct re-analyses for substantive research questions with record information. This could be a similar research design as implemented in studies two and three in this thesis. It could be expected that if the amount of measurement error decreases over panel waves, the impact of measurement error should also be smaller if the analysis sample is restricted on later panel waves.

It would be also interesting to see, whether the patterns for consent have changed over time. PASS would allow such analyses as in every wave a new entry sample into UB-II-receipt is drawn and thus in every year a new sample of respondents is asked for their consent to the record-linkage. Available studies by Beste (2011) and Sakshaug and Kreuter (2012) use only the survey data of the first panel wave. Beste (2011) found an association between income and migration status and the propensity to consent. In both studies only a minor bias due to non-consent was found. While the consent rates remain fairly stable across panel waves, it might be possible that due to a raising awareness regarding data protection the patterns for consent have changed. If this is the case, the bias due to non-consent might have changed as well.

In this thesis, register data was used to define the gold standard and used only the measurement of one construct, the welfare receipt. However, welfare receipt is not the only element of intersection between the data sources. Yet, the quality of the measurement varies substantially between each construct. For welfare receipt, the data quality is higher in the administrative data, for education the survey information is more reliable (Kruppe, Matthes, and Unger 2014). A systematic comparison could provide a feeling of quality for each of the comparable indicators and enable researchers to act accordingly.

Despite these possibilities, I would also argue that the "simple" comparison of indicators has a limited usefulness as such research does not necessarily help applied researchers, when analysing a specific research question as knowledge of error might not help to reduce the bias in the analyses. Using the linked data is also not always an option, as the use of the linked data is restricted due to data protection issues. Also, due to non-consent and failed linkage, analyses are only

possible for a (large) subset of respondents. Yet, as the lack of information for non-consenters is basically a missing data problem, one way to use the linked data might be the application of multiple imputation (Little and Rubin 2002). With multiple imputation, it might be possible to generate synthetic sets of variables for the register information for non-consenters. Going one step further, it might also be possible to generate sets of synthetic variables for the full sample, which might reduce the bias due to measurement error. This approach would also maintain data confidentiality and the synthetic variables could be made available for the scientific community with less restrictions. Another way to enhance the analytical possibilities might be the use of two-phase or two-stage design methods (Cain and Breslow 1988). With a two-phase study design, the full data can be analysed, even if the information for variables is missing for the largest part of the data. Such an approach could be employed to use survey data when analysing register data, even if the survey information is available only for a small fraction of the administrative records. Both methods could be possibly applied in order to broaden the usability of the linked data. The main downside of both approaches is that they are highly technical, which decreases their usability for the applied researcher.

Bibliography

AAPOR The American Association for Public Opinion Research (2009). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* 6th ed. AAPOR.

Achatz, J. and M. Trappmann (2011). *Arbeitsmarktvermittelte Abgänge aus der Grundsicherung – der Einfluss von personen- und haushaltsgebundenen Barrieren.* IAB-Discussion Paper 02/2011. Institut für Arbeitsmarkt- und Berufsforschung.

Allison, P. D. (2009). *Fixed Effects Regression Models.* Los Angeles: Sage.

Alwin, D. F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement.* Hoboken: Wiley.

Angrist, J. and J.-S. Pischke (2008). *Mostly harmless econometrics.* Princeton: Princeton University Press.

Ansolabehere, S. and E. Hersh (2012). 'Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate.' In: *Political Analysis* 20.4, pp. 434–479.

Antoni, M. and A. Bethmann (2014). *PASS-Befragungsdaten verknüpft mit administrativen Daten des IAB (PASS-ADIAB) 1975–2011.* FDZ-Datenreport 03/2014. Institut für Arbeitsmarkt- und Berufsforschung.

Aughinbaugh, A. and R. M. Gardecki (2008). 'Assessing Acquiescence among Participants in the National Longitudinal Survey of Youth 1997'. Unpublished.

Barret, K., M. Sloan, and D. Wright (2006). *Interviewer Perceptions of Interview Quality.* Proceedings of the ASA – Survey Methods Section.

Battistin, E. and A. Chesher (2014). 'Treatment effect estimation with covariate measurement error'. In: *Journal of Econometrics* 178.2, pp. 707–715.

Belli, R. F., I. Bilgen, and T. Al-Baghal (2013). 'Memory, communication, and data quality in calendar interviews'. In: *Public Opinion Quarterly* 77.1, pp. 194–219.

Berg, M., R. Cramer, C. Dickmann, R. Gilberg, B. Jesske, M. Kleudgen, A. Bethmann, B. Fuchs, M. Trappmann, and A. Wurdack (2012). *Codebuch und Dokumentation des ‚Panel Arbeitsmarkt und soziale Sicherung' (PASS) Band I: Datenreport Welle 5.* FDZ-Datenreport 06/2012. Institut für Arbeitsmarkt- und Berufsforschung.

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

Beste, J., A. Bethmann, and M. Trappmann (2010). *Arbeitsmotivation und Konzessionsbereitschaft: ALG-II-Bezug ist nur selten ein Ruhekissen.* IAB-Kurzbericht 15/2010. Institut für Arbeitsmarkt- und Berufsforschung.

Bethmann, A., B. Fuchs, and A. Wurdack (2013). *User Guide „Panel Study Labour Market and Social Security" (PASS). Wave 6.* FDZ-Methodenreport 07/2013. Institut für Arbeitsmarkt- und Berufsforchung.

Biemer, P. (2001). 'Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing'. In: *Journal of Offcial Statistics* 17.2, pp. 295–320.

– (2011). *Latent Class Analysis of Survey Error.* Hoboken: Wiley.

Bollinger, C. R. and M. H. David (1997). 'Modeling Discrete Choice With Response Error: Food Stamp Participation'. In: *Journal of the American Statistical Association* 92.439, pp. 827–835.

– (2001). 'Estimation with Response Error and Nonresponse: Food-Stamp Participation in the SIPP'. In: *Journal of Business & Economic Statistics* 19.2, pp. 129–141.

– (2005). 'I didn't tell, and i won't tell: dynamic response error in the SIPP'. In: *Journal of Applied Econometrics* 20.4, pp. 563–569.

Booth, M. and K. Scherschel (2010). *The impact of activating labor market policies on labor market orientations and institutions.* Working Papers 9. Economic Sociology Jena.

Bound, J., C. Brown, and N. Mathiowetz (2001). 'Measurement Error in Survey data'. In: *Handbook of Econometrics.* Ed. by J. J. Heckman and E. Leamer. Vol. 5. Amsterdam: Elsevier, pp. 3705 3843.

Bound, J. and A. B. Krueger (1991). 'The extent of measurement error in longitudinal earnings data: Do two wrongs make it right?' In: *Journal of Labor Economics* 9.1, pp. 1–24.

Brown, M. B. and A. B. Forsythe (1974). 'Robust Tests for the Equality of Variances'. In: *Journal of the American Statistical Association* 69.346, pp. 364–367.

Bruckmeier, K., J. Eggs, C. Himsel, M. Trappmann, and U. Walwei (2013). *Steinig und lang – der Weg aus dem Leistungsbezug.* IAB-Kurzbericht 14/2013. Institut für Arbeitsmarkt- und Berufsforschung.

Bruckmeier, K., G. Müller, and R. T. Riphahn (2014). 'Who misreports welfare receipt in surveys?' In: *Applied Economic Letters* 21.12, pp. 812–816.

– (2015). 'Survey misreporting of welfare receipt – respondent, interviewer, and interview characteristics'. In: *Economics Letters* 129, pp. 103–107.

Buhr, P., T. Lietzmann, and W. Voges (2010). 'Lange Wege aus Hartz IV? zur Dynamik von Mindestsicherung unter dem Bundessozialhilfegesetz und dem SGB II'. In: *ZeS Report* 15, pp. 1–6.

Burt, C. D., S. Kemp, M. Conway, and J. M. Grady (2000). 'Ordering autobiographical experiences'. In: *Memory* 8.5, pp. 323–332.

Cain, K. C. and N. Breslow (1988). 'Logistic-regression analysis and efficient design for 2-stage studies'. In: *American Journal of Epidemiology* 128.6, pp. 1198–1207.

Callegaro, M. (2008). 'Seam Effects in Longitudinal Surveys'. In: *Journal of Offcial Statistics* 24.3, pp. 387–409.

Cannell, C. F., K Marquis, and A. Laurent (1977). 'A Summary of Studies of Interviewing Methodology'. In: *Vital and health statistics: Series 2.* 2nd ser. 69. Ed. by C. Cannell.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Craniceanu (2006). *Measurement error in nonlinear models.* 2nd ed. New York: Chapman & Hall.

Chowdhury, G. and S. Nickell (1985). 'Hourly Earnings in the United States: Another Look at Unionization, Schooling, Sickness, and Unemployment Using PSID Data'. In: *Journal of Labor Economics* 3.1, pp. 38

Collett, D. (2003). *Modelling Survival Data in Medical Research.* 2nd ed. New York: Chapman & Hall.

Conrad, F. G., L. J. Rips, and S. S. Fricker (2009). 'Seam Effects in Quantitative Responses'. In: *Journal of Official Statistics* 25.3, pp. 339–361.

Czajka, J. L. (1983). 'Subannual Income Estimation'. In: *Technical, Conceptual and Administrative Lessons of the Income Survey Development Program.* Ed. by M. David. New York: Social Science Research Council.

– (2013). 'Can administrative records be used to reduce nonresponse bias?' In: *The Annals of the American Academy of Political and Social Science* 645, pp. 171–184.

Davis, R. E., M. P. Couper, N. K. Janz, C. H. Caldwell, and K. Resnicow (2010). 'Interviewer effects in public health surveys'. In: *Health Education Research* 25.1, pp. 14–26.

Ebbinghaus, H. (1885). *Über das Gedächtnis – Untersuchungen zur experimentellen Psychologie.* Leipzig: Duncker.

Eggs, J. (2013). *Unemployment benefit II, unemployment, and health.* IAB-Discussion Paper 12/2013. Institut für Arbeitsmarkt- und Berufsforschung.

– (2015). *Measurement error for welfare receipt and its impact on panel models.* Proceedings of Statistics Canada Symposium 2014 forthcoming.

Eggs, J. and A. Jäckle (2015). 'Dependent Interviewing and Sub-Optimal Responding'. In: *Survey Research Methods* 9.1, pp. 15–29.

Eichhorst, W., M. Grienberger-Zingerle, and R. Konle-Seidl (2010). 'Activating Labor Market and Social Policies in Germany: From Status Protection to Basic Income Support'. In: *German Policy Studies* 6.1, pp. 65–106.

Eisenhower, D., N. A. Mathiowetz, and D. Morganstein (1991). 'Recall Error: Sources and Bias Reduction Techniques'. In: *Measurement Errors in Surveys.* Ed. by P. B. Biemer, R. M. Grove, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. Wiley, pp. 127–144.

Esser, H. (2006). 'Migration, Sprache und Integration'. In: *AKI-Forschungsbilanz* 4.

Feldman, J. J., H. Hyman, and C. W. Hart (1951). 'A Field Study of Interviewer Effects on the Quality of Survey Data'. In: *Public Opinion Quarterly* 15.4, pp. 734–761.

Freedman, L. S., J. M. Commins, J. E. Moler, W. Willett, L. F. Tinker, A. F. Subar, D. Spiegelman, D. Rhodes, N. Potischman, M. L. Neuhouser, A. J. Moshfegh, V. Kipnis, L. Arab, and R. L. Prentice (2015). 'Pooled results from 5 validation studies of dietary self-report instruments using recovery biomarkers for potassium and sodium intake.' In: *American Journal of Epidemiology* 181.7, pp. 473–487.

Freeman, R. B. (1984). 'Longitudinal Analyses of the Effects of Trade Unions'. In: *Journal of Labor Economics* 2.1, pp. 1–26.

Frick, J. R., J. Goebel, E. Schechtman, G. G. Wagner, and S. Yitzhaki (2006). 'Using analysis of Gini (ANOGI) for detecting whether two subsamples represent the same universe the German Socio-Economic Panel Study (SOEP) experience'. In: *Sociological Methods & Research* 34.4, pp. 427–468.

Friedman, W. J. (2004). 'Time in Autobiographical Memory'. In: *Social Cognition* 22.5, pp. 591–605.

(2007). 'The role of reminding in long-term memory for temporal order'. In: *Memory & Cognition* 35.1, pp. 66–72.

Gebhardt, D., G. Müller, A. Bethmann, M. Trappmann, B. Christoph, C. Gayer, B. Müller, A. Tisch, B. Siflinger, H. Kiesl, B. Huyer-May, J. Achatz, C. Wenzig, H. Rudolph, T. Graf, and A. Biedermann (2009). *Codebuch und Dokumentation des ‚Panel Arbeitsmarkt und soziale Sicherung' (PASS) – Band I: Einführung und Überblick*. FDZ-Datenreport 06/2009. Institut für Arbeitsmarkt- und Berufsforschung.

Geschäftsbereich ITM (2009). *LHG. Version 5.01.* Benutzerhandbuch. Institut für Arbeitsmarkt- und Berufsforschung.

– (2014). *Benutzerhandbuch LHG Leistungshistorik Grundsicherung Version 08.00. Benutzerhandbuch.* Institut für Arbeitsmarkt- und Berufsforschung.

Gray, P. (1955). 'The Memory Factor in Social Surveys'. In: *Journal of the American Statistical Association* 50.270, pp. 344–363.

Groen, J. A. (2012). 'Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures'. In: *Journal of Offcial Statistics* 28.2, pp. 173–198.

Groves, R., F. Fowler, M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology.* Hoboken: Wiley.

– (2009). *Survey Methodology.* 2nd ed. Hoboken: Wiley.

Groves, R. M. (1991). 'Measurement Error across the Disciplines'. In: *Measurement Error ins Surveys.* Ed. by P. Biemer, R. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. Wiley.

Groves, R. M. and L. Lyberg (2010). 'Total survey error: Past, present, and future'. In: *Public Opinion Quarterly* 74.5, pp. 849–879.

Hernan, M. A. and J. M. Robins (2014). *Causal Inference.* Working text. Chapman & Hall.

Hoogendoorn, A. W. (2004). 'A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing'. In: *Journal of Official Statistics* 20.2, pp. 219–232.

Hox, J. J. (2010). *Multilevel Analysis – Techniques and Applications.* Routledge.

Huttenlocher, J., L. V. Hedges, and N. M. Bradburn (1990). 'Reports of elapsed time: bounding and rounding processes in estimation'. In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 16.2, pp. 196–213.

Ioannidis, J. (2013). 'Implausible results in human nutrition research'. In: *British Medical Journal* 347, pp. 725–727.

Jacobebbinghaus, P. and S. Seth (2007). 'The German integrated employment biographies sample IEBS'. In: *Schmollers Jahrbuch* 127.2, pp. 335–342.

Janssen, S. M., A. G. Chessa, and J. M. J. Murre (2006). 'Memory for time: how people date events'. In: *Memory & Cognition* 34.1, pp. 138–147.

Jaro, M. (1989). 'Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida'. In: *Journal of the American Statistical Association* 84.406, pp. 351–357.

Jäckle, A. (2008). *Measurement Error and Data Collection Methods: Effects on Estimates from Event History Data.* ISER Working Paper Series 2008-13. Institute for Social & Economic Research.

– (2009). 'Dependent Interviewing: A Framework and Application to Current Research'. In: Methodology of Longitudinal Surveys. Ed. by P. Lynn. Chichester: Wiley, pp. 93–110.

Jäckle, A., J. Eggs, and M. Trappmann (2015). 'Will respondents eventually get it right? Changes in measurement error across waves of a panel survey using dependent interviewing: Results from a five-wave Validation study'. Unpublished Manuscript.

Jäckle, A. and P. Lynn (2007). 'Dependent Interviewing and Seam Effects in Work History Data'. In: *Journal of Offcial Statistics* 23.4, pp. 529–551.

Johanson, G. A. and C. J. Osborn (2004). 'Acquiescence as differential person functioning'. In: *Assessment & Evaluation in Higher Education* 29.5, pp. 535–548.

John, O. P. and S. Srivastava (1999). 'The Big Five trait taxonomy: History, measurement, and theoretical perspectives'. In: *Handbook of personality: Theory and research.* Ed. by L. A. Pervin and O. P. John. New York: Guilford.

Kapteyn, A. and J. Y. Ypma (2007). 'Measurement Error and Misclassification: A Comparison of Survey and Administrative Data'. In: *Journal of Labor Economics* 25.3, pp. 513–551.

Küchenhoff, H., S. M. Mwalili, and E. Lesaffre (2006). 'A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX'. In: *Biometrics* 62.1, pp. 85

Kieruj, N. D. and G. Moors (2013). 'Response style behavior: Question format dependent or personal style?' In: *Quality & Quantity* 47.1, pp. 193–211.

Kim, C. and C. R. Tamborini (2014). 'Response Error in Earnings: An analysis of the survey of Income and Program Participation Matched with administrative Data'. In: *Sociological Methods & Research* 43.39, pp. 36–72.

Knäuper, B., R. F. Belli, D. H. Hill, and A. R. Herzog (1997). 'Question Difficulty and Respondents' Cognitive ability: The effect on data quality'. In: *Journal of Official Statistics* 13.1, pp. 181–199.

Knowles, E. S. and K. T. Nathan (1997). 'Acquiescent responding in Self-Reports: Cognitive Style or Social Concern'. In: *Journal of Research in Personality* 31.2, pp. 293–301.

Köhler, M. and U. Thomsen (2009). 'Data Integration and Consolidation of Administrative Data from various Sources. The case of Germans' Employment Histories'. In: *Historical Social Research* 34.3, pp. 215–229.

Korbmacher, J. and M. Schröder (2013). 'Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer'. In: *Survey Research Methods* 7.2, pp. 115–131.

Kreuter, F., G. Müller, and M. Trappmann (2010). 'Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data'. In: *Public Opinion Quarterly* 74.5, pp. 880–906.

Krosnick, J. A. (1999). 'Survey research'. In: *Annual Review of Psychology* 50, pp. 537–567.

Krosnick, J. A. and D. F. Alwin (1987). 'An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement'. In: *The Public Opinion Quarterly* 51.2, pp. 201–219.

Krosnick, J. A., A. L. Holbrook, M. K. Berent, R. T. Carson, M. W. Hanemann, R. J. Kopp, R. C. Mitchell, S. Presser, P. Ruud, K. V. Smith, W. R. Moody, M. C. Green, and M. Conway (2002). 'The Impact of „No Opinion" Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice?' In: *Public Opinion Quarterly* 66.3, pp. 371–403.

Krumpal, I. (2011). 'Determinants of social desirability bias in sensitive surveys'. In: *Quality & Quantity* 47.4, pp. 20205–2047.

Kruppe, T., B. Matthes, and S. Unger (2014). *Effectiveness of data correction rules in process-produced data: The case of educational attainment*. IAB-Discussion Paper 15/2014. Institut für Arbeitsmarkt- und Berufsforschung.

Kyyrä, T. and R. A. Wilke (2014). 'On the reliability of retrospective unemployment information in european household panel data'. In: *Empirical Economics* 46.4, pp. 1473–1493.

Leeuw, E. de (2005). 'To Mix or Not to Mix Data Collection Modes in Surveys'. In: *Journal of Offcial Statistics* 21.2, pp. 233–255.

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data.* 2nd ed. Wiley.

Loftus, E. F. and W. Marburger (1983). 'Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events'. In: *Memory & Cognition* 11.2, pp. 114–120.

Luks, S. and H. E. Brady (2003). 'Defining Welfare Spells: Coping with problems of survey responses and administrative data'. In: *Evaluation Review* 27.4, pp. 395–420.

Lynn, P. (2009). 'Methods for Longitudinal Surveys'. In: *Methodology of Longitudinal Surveys.* Ed. by P. Lynn. Wiley, pp. 1–20.

Lynn, P., A. Jäckle, S. P. Jenkins, and E. Sala (2006). 'The effects of dependent interviewing on responses to questions on income sources'. In: *Journal of Offcial Statistics* 22.3, pp. 357–384.

– (2012). 'The Impact of Interviewing Method on Measurement Error in Panel Survey Measures of Benefit Receipt: Evidence from a Validation Study'. In: *Journal of the Royal Statistical Society, Series A* 175.2, pp. 289–308.

Marquis, K. H. (1978). *Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations.* Tech. rep. Rand Corporation.

Mathiowetz, N. A. and K. A. McGonagle (2000). 'An assessment of the current state of dependent interviewing in household surveys'. In: *Journal of Offcial Statistics* 16.4, pp. 401–418.

May, M. and J. Schwanholz (2013). 'Vom gerechten Weg abgekommen? Bewertungen von Hartz IV durch die Bevölkerung'. In: *Zeitschrift für Sozialreform* 59.2, pp. 197–225.

Meisenberg, G. and A. Williams (2008). 'Are acquiescent and extreme response styles related to low intelligence and education?' In: *Personality and Individual Differences* 44.7, pp. 1539–1550.

Millimet, D. L. (2011). 'The Elephant in the Corner: A Cautionary Tale about Measurement Error in Treatment Effects Models'. In: *Advances in Econometrics* 27, pp. 1–39.

Moore, J., N. Bates, J. Pascale, and A. Okon (2009). 'Tackling Seam Bias Through Questionnaire Design'. In: *Methodology of Longitudinal Surveys.* Ed. by P. Lynn. Wiley, pp. 73–92.

Neter, J. and J. Waksberg (1964). 'A Study of Response Errors in Expenditures Data from Household Interviews'. In: *Journal of the American Statistical Association* 59.305, pp. 18–55.

Nickerson, R. S. (1998). 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.' In: *Review of General Psychology* 2.2, pp. 175–220.

Novick, M. R. (1966). 'The Axioms and Principal Results of Classical Test Theory'. In: *Journal of Mathematical Psychology:* 3.1, pp. 1–18.

Olson, K. and I. Bilgen (2011). 'The Role Of Interviewer Experience on Acquiescence.' In: *Public Opinion Quarterly* 75.1, pp. 99–114.

Paull, G. (2002). *Biases in the reporting of labour market dynamics.* IFS Working Papers W02/10. Institute for Fiscal Studies.

Pierret, C. R. (2001). 'Event History Data and Survey Recall: An Analysis of the National Longitudinal Survey of Youth 1979 Recall Experiment'. In: *The Journal of Human Resources* 36.3, pp. 439–466.

Pigeot-Kübler, I. and R. Schnell (2006). *Errors in autobiographical memory and their effects in time-to-event analysis.* Proposal for a research grant to the DFG (unpublished manuscript).

Pina-Sanchez, J., J. Koskinen, and I. Plewis (2013). 'Implications of Retrospective Measurement Error in Event History Analysis'. In: *Metodologia de Encuestas* 15, pp. 5–25.

Pyy-Martikainen, M. and U. Rendtel (2009). 'Measurement Errors in Retrospective Reports of Event Histories – A Validation Study with Finnish Register Data'. In: *Survey Research Methods* 3.3, pp. 139 155.

Rammstedt, B. and O. P. John (2005). 'Kurzversion des Big Five Inventory (BFIK): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit'. In: *Diagnostica* 51.4, pp. 195–206.

Rendtel, U. (2012). *The Fade-Away Effect of initial nonresponse in panel surveys: Empirical results for EU-SILC.* Freie Universität Berlin.

Rendtel, U., L. Nordberg, M. Jäntti, J. Hanisch, and E. Basic (2004). *Report on quality of income data.* CHINTEX Working Paper 12.

Roediger, H. L. (2008). 'Relativity of remembering: why the laws of memory vanished'. In: *Annual Review of Psychology* 59, pp. 225–254.

Rubin, D. C. and A. E. Wenzel (1996). 'One Hundred Years of Forgetting: A Quantitative Description of Retention'. In: *Psychological Review* 103.4, pp. 734–760.

Sakshaug, J. and F. Kreuter (2012). 'Assessing the magnitude of non-consent biases in linked survey and administrative data'. In: *Survey Research Methods* 6.2, pp. 113–122.

Sala, E., S. N. Uhrig, and P. Lynn (2011). 'It Is Time Computers Do Clever Things!": The Impact of Dependent Interviewing on Interviewer Burden.' In: *Field Methods* 23.1, pp. 3–23.

Schneeweiß, H. and H.-J. Mittag (1986). *Lineare Modelle mit fehlerbehafteten Daten*. Heidelberg: Physica-Verlag.

Schnell, R. (2014). 'Linking Surveys and Administrative Data'. In: *Improving Survey Methods.* Ed. by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. Routledge.

Schnell, R., T. Bachteler, and J. Reiher (2005). 'MTB: Ein Record-Linkage-Programm für die empirische Sozialwissenschaft'. In: *ZA-Information* 56, pp. 93–103.

Schnell, R., P. B. Hill, and E. Esser (2008). *Methoden der empirischen Sozialforschung.* 6th ed. München: Oldenbourg.

Schober, M. F. and F. G. Conrad (1997). 'Does conversational interviewing reduce survey measurement error?' In: *Public Opinion Quarterly* 61.4, pp. 576–602.

Schoeni, R. F., F. Stafford, K. A. McGonagle, and P. Andreski (2013). 'Response Rates in National Panel Surveys.' In: *The Annals of the American Acadamy of Political and Social Science* 645, pp. 60–87.

Schwarz, N (2007). 'Cognitive aspects of survey methodology'. In: *Applied Cognitive Psychology* 21.2, pp. 277–287.

Soubelet, A. and T. A. Salthouse (2011). 'Influence of Social Desirability on Age Differences in Self-Reports of Mood and Personality'. In: *Journal of Personality* 79.4, pp. 741–762.

Statistik der Bundesagentur für Arbeit (2013a). *Arbeitsmarkt in Zahlen – Erwerbstätige Arbeitslosengeld II-Bezieher September 2012.*

– (2013b). *Arbeitsmarkt in Zahlen. Statistik der Grundsicherung für Arbeitsuchende – Bedarfsgemeinschaften und deren Mitglieder November 2012.*

Sudman, S. and N. Bradburn (1973). 'Effects of time and memory factors on response in surveys'. In: *Journal of the American Statistical Association* 68.344, pp. 805–815.

Talarico, J. M. and D. C. Rubin (2003). 'Confidence, not consistency, characterizes flashbulb memories'. In: *Psychological Science* 14.5, pp. 455–461.

Thomas, R. K. (2014). 'Fast and Furious ... ... or Much Ado About Nothing? Sub-Optimal Respondent Behavior and Data Quality.' In: *Journal of Advertising Research* 54.1, pp. 17–31.

Thompson, C., J. Skowronski, S. Laarsen, and A. Betz (1996). *Autobiographical Memory: Remembering What and Remembering When.* Mahwah: Erlbaum. Tourangeau, R., L. J. Rips, and K. Rasinski (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Tourangeau, R. and T. Yan (2007). 'Sensitive questions in surveys.' In: *Psychological Bulletin* 133.5, pp. 859–883.

Trappmann, M., J. Beste, A. Bethmann, and G. Müller (2013). 'The PASS panel survey after six waves'. In: *Journal for Labour Market Research* 46.4, pp. 1–7.

Trivellato, U. (1999). 'Issues in the Design and Analysis of Panel Studies: A Cursory Review'. In: *Quality & Quantity* 33.3, pp. 339–352.

Vaerenbergh, Y. van and T. D. Thomas (2013). 'Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies.' In: *International Journal of Public Opinion Research* 25.2, pp. 192–217.

Warren, J. R. and A. Halpern-Manners (2012). 'Panel Conditioning in Longitudinal Social Science Surveys'. In: *Sociological Methods & Research* 41.4, pp. 491–534.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.

Appendix

Appendix for Chapter 2

The material deprivation score is based on the following questions. Respondents who answer "No" are asked a follow-up question to determine whether this is for financial reasons. The summed deprivation score counts the number of items the household does not have for financial reasons.

1. Do you have an apartment with at least as many rooms as persons living there?
2. Do you have an apartment without damp walls or floors?
3. Do you have a separate bathroom with bathtUB or shower in your apartment?
4. Do you have a toilet inside your apartment?
5. Do you have a garden, a balcony or a terrace?
6. Do you have sufficient winter clothing [for each household member]?
7. Do you have a car?
8. Do you have a television?
9. Do you have a video recorder or DVD player?
10. Do you have a computer with internet access?
11. Do you have a washing machine?

And which of the following things do you/or does your household do?

12. Buy new clothing once in a while [for each family member], even if the old clothes are not yet worn-out?
13. Do you have a hot meal at least once per day?
14. A holiday away from home for at least one week a year [for each member of the family, which however need not be taken jointly]?
15. Invite friends for dinner at home once per month?
16. Eat out at a restaurant [with the family] once a month?
17. You [Each member of the family] can go to the cinema, a theatre or a concert at least once a month?
18. Saving a fixed amount a month?
19. Replacing worn but still useable furniture with new?
20. Pay for unexpected expenses with one's own money, e.g. to replace a broken washing machine.
21. Having medical treatment which is not fully covered by your health insurance, such as dentures or glasses [if you/one of the family members need them]?
22. Always pay the rent for the apartment and/or the interest on the house or apartment one lives in on time?
23. Always pay the gas, heating and electricity bill on time

Appendix for Chapter 3

Table A1: Fixed effect linear regressions for a subjective health score: Parameter estimates and 95% confidence intervals for men for different model specifications

	Complete sample			Linked sample			with register information			with averaged values			without t_1			using sample frame information			Balanced panel		
	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI	$\hat{\beta}$	95%	CI
Unemployed = 1	-0.099***	-0.137	-0.062	-0.094***	-0.133	-0.054	-0.115***	-0.154	-0.077	-0.100***	-0.159	-0.041	-0.070*	-0.126	-0.015	-0.105***	-0.143	-0.066	-0.080**	-0.138	-0.022
UB II = 1	-0.038*	-0.074	-0.002	-0.051**	-0.088	-0.013	-0.005	-0.045	0.035	-0.088**	-0.147	-0.029	-0.064*	-0.113	-0.014	-0.032	-0.065	0.001	-0.092**	-0.152	-0.033
2. wave	-0.050***	-0.078	-0.021	-0.052***	-0.082	-0.021	-0.052***	-0.082	-0.021							-0.055***	-0.086	-0.024	-0.023	-0.071	0.025
3. wave	-0.062**	-0.105	-0.018	-0.060*	-0.106	-0.014	-0.059*	-0.105	-0.012	-0.028	-0.067	0.011	-0.006	-0.050	0.038	-0.063**	-0.109	-0.017	0.015	-0.059	0.090
4. wave	-0.084**	-0.147	-0.020	-0.086*	-0.154	-0.018	-0.086*	-0.154	-0.018	-0.050	-0.127	0.028	-0.050	-0.136	0.036	-0.090**	-0.158	-0.023	0.015	-0.097	0.127
5. wave	-0.119**	-0.201	-0.038	-0.117**	-0.204	-0.031	-0.117**	-0.203	-0.030	-0.083	-0.198	0.032	-0.084	-0.208	0.041	-0.122**	-0.208	-0.035	0.014	-0.131	0.160
Partner = 1	0.012	-0.024	0.048	0.014	-0.025	0.053	0.015	-0.024	0.054	-0.003	-0.070	0.064	0.007	-0.043	0.057	0.014	-0.025	0.053	0.003	-0.066	0.071
Log(HHincome)	0.014	-0.005	0.033	0.013	-0.007	0.033	0.014	-0.007	0.034	0.024	-0.011	0.058	0.002	-0.025	0.029	0.013	-0.007	0.033	0.017	-0.016	0.050
Age	0.017	-0.002	0.036	0.017	-0.003	0.037	0.017	-0.003	0.037	0.019	-0.018	0.056	0.024	-0.016	0.064	0.017	-0.003	0.037	-0.015	-0.050	0.019
_cons	-0.721	-1.519	0.077	-0.689	-1.530	0.151	-0.717	-1.558	0.125	-0.897	-2.479	0.686	-0.964	-2.670	0.742	-0.688	-1.527	0.151	0.636	-0.865	2.137
N	17926			15354			15354			9734			11051			15354			5969		
N_g	5837			4941			4941			4333			4751			4941			1315		

* p < 0.05, ** p < 0.01, *** p < 0.001
Source: Linked PASS-IEB data 2006–2011.

Table A2: Fixed effect linear regression for a subjective health score: Parameter estimates and 95 % confidence intervals for women for different model specifications

	Complete sample			Linked sample			with register information			with averaged values			without t_1			using sample frame information			Balanced panel		
	$\hat{\beta}$	95 %	CI	$\hat{\beta}$	95 %	CI	$\hat{\beta}$	95 %	CI	$\hat{\beta}$	95 %	CI	$\hat{\beta}$	95 %	CI	$\hat{\beta}$	95 %	CI	$\hat{\beta}$	95 %	CI
Unemployed = 1	-0.091***	-0.129	-0.052	-0.099***	-0.140	-0.059	-0.096***	-0.136	-0.057	-0.083**	-0.143	-0.024	-0.098***	-0.152	-0.045	-0.091***	-0.130	-0.051	-0.087**	-0.144	-0.030
UB II = 1	0.013	-0.022	0.048	0.011	-0.026	0.048	0.004	-0.037	0.044	0.008	-0.045	0.062	-0.001	-0.053	0.051	-0.011	-0.045	0.022	0.001	-0.052	0.054
2. wave	-0.027	-0.057	0.002	-0.021	-0.052	0.010	-0.021	-0.051	0.010							-0.022	-0.053	0.010	-0.026	-0.068	0.015
3. wave	-0.033	-0.078	0.011	-0.020	-0.066	0.026	-0.020	-0.066	0.026	-0.027	-0.061	0.007	-0.007	-0.047	0.034	-0.022	-0.068	0.025	-0.036	-0.097	0.026
4. wave	-0.052	-0.115	0.011	-0.043	-0.109	0.022	-0.043	-0.109	0.022	-0.054	-0.120	0.011	-0.047	-0.121	0.027	-0.045	-0.111	0.020	-0.072	-0.155	0.011
5. wave	-0.078	-0.160	0.003	-0.061	-0.145	0.023	-0.061	-0.145	0.023	-0.092	-0.188	0.005	-0.083	-0.190	0.023	-0.064	-0.149	0.021	-0.120*	-0.226	-0.015
Partner = 1	0.013	-0.022	0.048	0.008	-0.029	0.046	0.008	-0.030	0.045	-0.010	-0.070	0.050	0.013	-0.035	0.062	0.007	-0.030	0.045	0.009	-0.049	0.067
Log(HHincome)	-0.004	-0.028	0.021	-0.003	-0.028	0.022	-0.003	-0.028	0.022	-0.011	-0.048	0.026	-0.015	-0.046	0.016	-0.003	-0.029	0.022	-0.005	-0.041	0.030
Age	0.004	-0.015	0.022	-0.001	-0.020	0.018	-0.001	-0.020	0.018	0.018	-0.012	0.048	0.012	-0.021	0.045	-0.001	-0.020	0.018	0.014	-0.009	0.038
_cons	-0.108	-0.891	0.676	0.079	-0.718	0.877	0.083	-0.715	0.882	-0.673	-1.987	0.642	-0.400	-1.798	0.998	0.097	-0.700	0.894	-0.555	-1.602	0.493
N	18370			16100			16100			10371			12068			16100			6862		
N_g	6290			5446			5446			4629			5198			5446			1661		

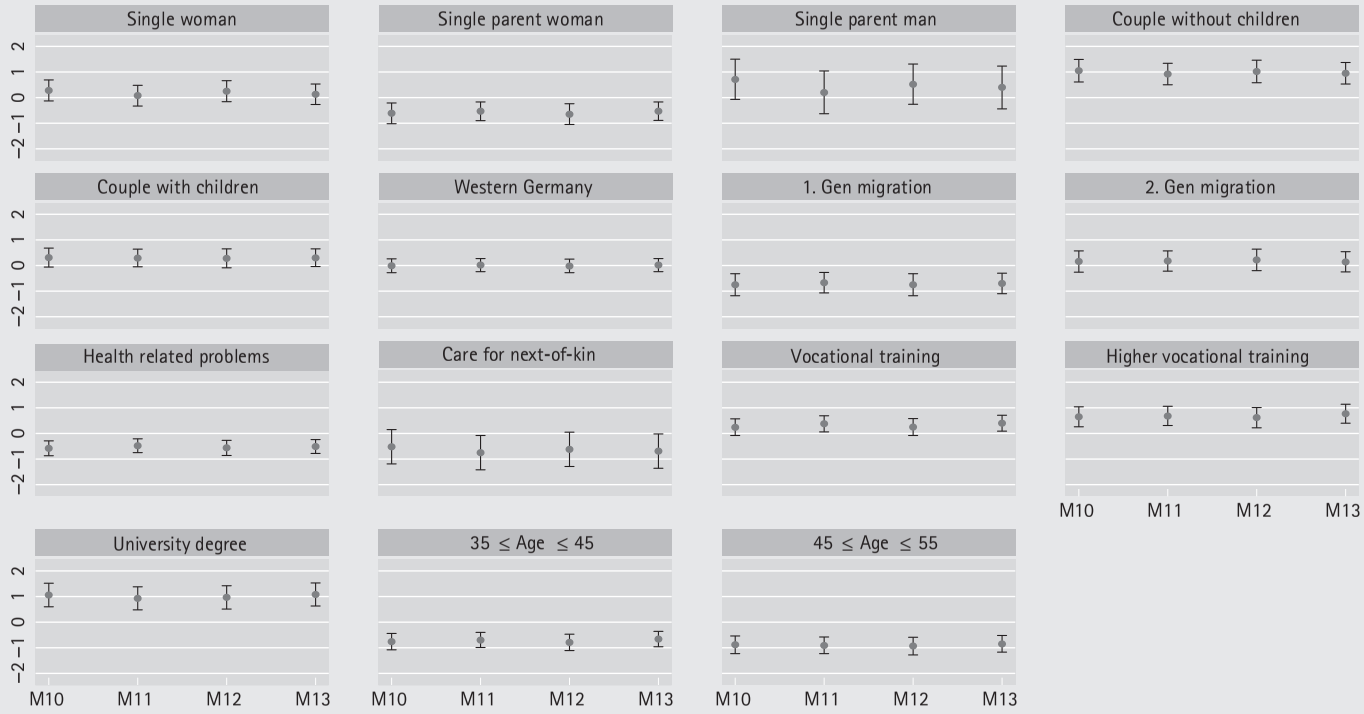
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Source: Linked PASS-IEB data 2006–2011.

Appendix for Chapter 4

Table A3: Variables: Definitions and descriptive statistics

Variable	Type	Definition	mean (SD), proportion
Elapsed time beginning	metric	Time between subsequent interview and date of spell beginning recorded in register.	23.49 (5.43)
Elapsed time end	metric	Time between subsequent interview and date of spell end recorded in register.	1.85 (2.36)
Begin 1/05 Reg	binary	1 if UB II spell began in January 2005, 0 spell began later.	0.62
Right-censored in register	binary	1 if UB II spell was ongoing at the time of the interview, 0 otherwise.	0.81
Prior UB I Spell	binary	1 if prior UB I spell was recorded in the register, 0 no prior UB I spell in register.	0.10
Foreign HH-language	binary	1 if language predominantly spoken in respondent household is not German, 0 otherwise.	0.13
No vocational training	binary	1 if respondent has no vocational training, 0 otherwise.	0.32
Vocational training	binary	1 if highest qualification is vocational training, 0 otherwise.	0.45
Higher vocational training	binary	1 if highest qualification is higher vocational training, 0 otherwise.	0.15
University degree	binary	1 if highest qualification is a university degree, 0 otherwise.	0.08
Age < 35	binary	1 if yes, 0 otherwise.	0.36
35 ≤ Age < 45	binary	1 if yes, 0 otherwise.	0.35
45 < Age ≤ 55	binary	1 if yes, 0 otherwise.	0.30
Woman	binary	1 if woman, 0 man.	0.57
Entry at seam	binary	1 if person placed transition into UB II incorrectly at the seam, 0 otherwise.	0.02
Average Entropy	continuous	Shannon's entropy was calculated for each of four item batteries. The results for each item battery was standardized and then averaged.	-0.05 (0.69)
Don't know/Refusals	Proportion	The proportion of refused/don't know answers on all answered questions was calculated.	0.034 (0.062)
Rounding	binary	The proportion of rounded answers on all answered numerical questions was calculated.	0.25 (0.22)
Western Germany	binary	1 if current residence in Western Germany, 0 otherwise.	0.64
CAPI	binary	1 CAPI interview mode, 0 CATI interview mode.	0.25
Single man	binary	1 if respondent single man, 0 otherwise.	0.27
Single woman	binary	1 if respondent single woman, 0 otherwise.	0.14
Single parent woman	binary	1 if respondent single parent woman, 0 otherwise.	0.28
Single parent man	binary	1 if respondent single parent man, 0 otherwise.	0.02
Couple without children	binary	1 if respondent lives with a partner without children, 0 otherwise.	0.08
Couple with children	binary	1 if respondent lives with a partner with children, 0 otherwise.	0.20
1. Gen migration	binary	1 if respondent immigrated, 0 otherwise.	0.18
2. Gen migration	binary	1 if parents of respondent immigrated, 0 otherwise.	0.09
Health related problems	binary	1 if respondent reported to have health related problems, 0 otherwise.	0.41
Care for next-of-kin	binary	1 if respondent reported to give care to next-of-kin, 0 otherwise.	0.06

Figure A1: Proportional hazard models: Parameter estimates and 95% confidence intervals for the risk of leaving UB II



M10: Register M11: Survey

M12: Cens.Reg Length.Sur M13:Cens.Sur Length.Reg

Table A4: Bias in coeffcients due types of response errors

	M10–M11 Total bias	M10–M12 Bias due misdating	M10–M13 Bias due censoring
Single man	*ref.*		
Single woman	0.2	0.03	0.15
Single parent woman	-0.08	0.04	-0.08
Single parent man	0.51	0.19	0.31
Couple without children	0.13	0.03	0.1
Couple with children	0.02	0.03	0.01
Western Germany	-0.03	0.01	-0.03
1. Gen migration	-0.08	0	-0.05
2. Gen migration	-0.02	-0.06	0.02
Health related problems	-0.1	-0.02	-0.07
Care for next-of-kin	0.23	0.1	0.17
No vocational training	*ref.*		
Vocational training	-0.14	-0.01	-0.16
Higher vocational training	-0.03	0.03	-0.12
University degree	0.13	0.09	-0.02
Age < 35	*ref.*		
35 ≤ Age < 45	-0.06	0.03	-0.1
45 ≤ Age ≤ 55	0.03	0.05	-0.03

Appendix for Chapter 5

Interviewer observations, asked at the end of each personal interview:
In your opinion: How difficult was it for the respondent to date certain events?
1 Very difficult
2
3
4
5 Not difficult at all

In your opinion: How interesting was the interview for the respondent?
1 Not interesting at all
2
3
4
5 Very interesting

Table A5: Summary statistics for respondents with false preload (continuous variables)

	mean	sd	min	max	count
Respondent age	44.21	12.23	20.00	67.00	276
Number of UB II spells in records	1.63	0.94	0.00	5.00	280
Months since last UB-II-receipt in records	9.11	6.41	0.07	39.63	280
Extroversion	-0.05	0.89	-2.20	1.45	203
Openness	0.15	0.81	-2.12	1.40	202
Neuroticism	-0.03	0.81	-1.56	2.10	203
Conscientiousness	0.06	0.70	-2.09	1.12	203
Agreeableness	-0.08	0.72	-1.70	1.61	202
Interviewer experience in years	3.64	2.31	1.00	19.0	280

Notes: The Big Five personality traits were collected in wave 5 and hence some observations were lost due to attrition.

Table A6: Summary statistics for respondents with false preload (categorical variables)

	Percent	count
Female respondent	55.1	276
Higher education	56.9	276
Interviewer observation: diffculty dating events	6.2	260
Interviewer observation: interview not interesting	52.7	258
Rounding in more than 50 % of questions	23.6	276
Non-differentiation in 1+ item batteries	35.1	276
Item non-response > 1 %	13.4	276
CAPI (vs. CATI)	30.7	280
Female interviewer	57.9	280

Table A7: Average marginal effects of random effects logistic models for confirming the preload (socio-economic characteristics)

Pr(confirmed false preload)	AME	se
Female respondent	-0.023	0.074
Respondent age	0.001	0.003
Disability	0.056	0.095
Migrated	-0.020	0.120
Omitted: no schooling		
Lower secondary degree	-0.074	0.164
Higher secondary degree	-0.095	0.165
Vocational education	-0.050	0.096
Young children in household (age ≤ 4)	-0.177	
Omitted: single person		
Household without children	0.043	0.114
Single Parent	0.104	0.114
Household with children	0.098	0.107
Other	0.051	0.223
Regular employed person in HH	-0.128	0.082
Omitted: HH income < 500 Euro		
HH income 500–749 Euro	0.011	0.223
HH income 750–99 Euro	0.113	0.238
HH income = 1 000 Euro	0.082	0.226
Omitted: no HH savings		
HH savings 1 000 Euro	0.023	0.081
HH savings 1 000–2 499 Euro	-0.043	0.110
HH savings 2 500–4 999 Euro	0.143	0.147
HH savings ≥ 5 000 Euro	-0.107	0.115
HH owns home	-0.153	0.099
Omitted: UB-II-receipt 12 months		
UB-II-receipt 12–25 months	-0.081	0.157
UB-II-receipt > 25 months	0.167	0.137
Eastern Germany	0.094	0.073
N	262	
Rho	0.542	
AIC	353.775	

Notes: Multilevel Logistic Regression; Average marginal effects; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; HH = household

Abstract

Survey data serve the purpose of acquiring information for use in politics, business and science. Nevertheless, such data may be inaccurate, which could mean that the information obtained is distorted. One example is the measurement error in survey data.

The work looks at measurement errors in longitudinal surveys. Longitudinal data can be used to observe changes within units of analysis over time. However, if units of analysis change over the course of time, the measurement error may also change accordingly. To date, less research has been devoted to the structure and impact of measurement errors in longitudinal surveys than those in cross-sectional studies. The work aimed to complement the findings available in this research field. To this end, the measurement error for a central unit of a panel study was examined for up to five consecutive interviews conducted at different points in time. The study focused on the following questions:

- How does the measurement error change over time and how can these changes be explained?
- To what extent does the measurement error affect the regression models for longitudinal data and can this impact be rectified by means of simple error models?
- To what extent does the measurement error affect event history analyses?
- To what extent do incorrectly formulated questions affect the measurement error?

In order to analyse the measurement error in longitudinal surveys, the datasets of the 'Labour Market and Social Security' PASS panel study were merged with the register data of the Federal Employment Agency at the personal level. The measurement error was defined by comparing the entries for the criterion of basic income support for job seekers (Unemployment Benefit II). The register data were deemed to be correct and thus regarded as the true value.

The study clearly shows that the number of persons drawing Unemployment Benefit II is more likely to be under- than overreported. It also demonstrates that the total measurement error declines over the individual panel waves. Firstly, this means that data quality improves over time, and secondly, that a number of incorrect transitions into Unemployment Benefit II recipiency are recorded. A further result is that the measurement error correlates to a large number of variables and over time. Therefore, the fact that a measurement error occurs cannot automatically be attributed to a random process but is more likely to be due to a combination of recipient history, the associated social stigma and individual personality traits.

Other findings included the fact that the structure of the error may lead to the distinctions between recipients and non-recipients being overestimated; however, the error does not necessarily affect the estimated results of models in longitudinal surveys. A comparison of the different correction procedures for the measurement error ascertains that the various methods do not reduce the distortion across groups. With regard to administering panel studies, it would appear that the method of 'dependent interviewing', in other words, using information obtained in previous interviews for the current one, may be conducive to increasing data quality over time. However, in cases where incorrect information is input, the study also shows that respondents tend to corroborate it, even going so far as to adopt this flawed information.

Kurzfassung

Umfragedaten dienen der Informationsgewinnung für Politik, Wirtschaft und Wissenschaft. Allerdings können Daten fehlerbehaftet sein, was dazu führen kann, dass die gewonnenen Informationen Verzerrungen aufweisen. Ein solcher Fehler in Umfragedaten ist der Messfehler.

Die Arbeit beschäftigt sich mit Messfehlern im Längsschnitt. Längsschnittdaten können dazu genutzt werden, Veränderungen innerhalb von Untersuchungseinheiten über die Zeit zu beobachten. Wenn sich allerdings Untersuchungseinheiten über die Zeit verändern können, so kann sich auch der Messfehler über die Zeit verändern. Struktur und Auswirkungen von Messfehlern im Längsschnitt wurden bislang deutlich seltener analysiert als Messfehler im Querschnitt. Dieses Forschungsfeld etwas zu ergänzen, war das Ziel der Arbeit. Dazu wurde der Messfehler für einen zentralen Sachverhalt einer Panelstudie für bis zu fünf aufeinanderfolgende Interviewzeitpunkte untersucht. Der Fokus liegt dabei auf folgenden Fragestellungen:

- Wie verändert sich der Messfehler über die Zeit und wie lassen sich diese Veränderungen erklären?
- In welchem Maße hat der Messfehler Auswirkungen auf Regressionsmodelle für Längsschnittdaten und lassen sich diese Auswirkungen durch einfache Fehlermodelle korrigieren?
- In welchem Maß hat der Messfehler Auswirkungen auf Ereigniszeitanalysen?
- In welchem Maß beeinflussen falsch gestellte Fragen den Messfehler?

Zur Analyse des Messfehlers im Längsschnitt wurden auf Personenebene Umfragedaten der Panelstudie „Arbeitsmarkt und Soziale Sicherheit" (PASS) mit Registerdaten der Bundesagentur für Arbeit zusammengespielt. Durch den Vergleich der Einträge für das Merkmal zur Grundsicherung für Arbeitsuchende (Arbeitslosengeld II) konnte der Messfehler definiert werden. Es wurde davon ausgegangen, dass die Registerdaten fehlerfrei sind und somit den wahren Wert abbilden.

Die Untersuchung macht deutlich, dass der Bezug von Arbeitslosengeld II häufiger unter- als überberichtet wird. Auch zeigt sich, dass der gesamte Messfehler über die einzelnen Panelwellen zurückgeht. Dies bedeutet einerseits, dass sich die Datenqualität über die Zeit verbessert, und andererseits, dass eine Vielzahl von falschen Übergängen in den Arbeitslosengeld-II-Bezug in den Befragungsdaten entstehen. Ein weiteres Ergebnis ist, dass der Messfehler mit einer Vielzahl von Variablen und über die Zeit korreliert. Die Entstehung des Messfehlers ist somit nicht zwangsläufig auf einen Zufallsprozess zurückzuführen, sondern gründet wahrscheinlich auf einem Zusammenspiel zwischen der Bezugshistorie, empfundenem sozialem Stigma und Persönlichkeitseigenschaften. Weitere Ergebnisse sind,

dass die Struktur des Fehlers dazu führen kann, dass die Unterschiede zwischen Beziehern und Nicht-Beziehern überschätzt werden können; allerdings wirkt sich der Fehler nicht zwangsläufig auf Schätzergebnisse von Modellen im Längsschnitt aus. Bei dem Vergleich verschiedener Korrekturverfahren für den Messfehler wird festgestellt, dass die Verfahren die Verzerrung nicht über Gruppen hinweg vermindern. Hinsichtlich der Durchführung von Panelstudien zeigt sich, dass die Nutzung von „Dependent Interviewing", also dem Nutzen von Informationen aus vorherigen Interviews für das derzeitige Interview, einen Teil dazu beitragen kann, dass sich die Datenqualität über die Zeit erhöht. Wenn falsche Informationen eingespielt werden, zeigt sich allerdings auch, dass Befragte diese in der Regel bestätigen und dazu tendieren, die falsche Information fortzuschreiben.

Measurement error is a common phenomenon in the empirical sciences. Longitudinal data can especially be affected by it, as measurement error can influence measures of change, which is one of the primary reasons for collecting longitudinal data in panel surveys. However, measurement error in longitudinal data is rarely analysed.

In this series of papers, the measurement error for welfare receipt is analysed for up to five consecutive panel waves by linking panel survey data with administrative data on the individual level. Results from all four studies suggest that measurement error for welfare receipt is highly differential. The measurement error for welfare receipt is not based on a random process, but based on personal characteristics and the welfare histories of the respondents.

W. Bertelsmann Verlag

ISBN 978-3-7639-4111-7

Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit
Weddigenstraße 20-22 ▪ 90478 Nürnberg ▪ www.iab.de