Assessing Survey Data Quality Making Use of Administrative Data

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

Barbara Ingrid Maria Felderer

Eingereicht am 25. September 2015

Assessing Survey Data Quality Making Use of Administrative Data

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

Barbara Ingrid Maria Felderer

Eingereicht am 25. September 2015

Erstgutachterin:	Prof. Dr. Frauke Kreuter
	Universität Mannheim
Zweitgutachter:	Prof. Dr. Joachim Winter
	Ludwig-Maximilians-Universität München
Tag der Disputation:	23. Dezember 2015

Abstract

Social research is in large parts based on the analysis of survey data. Therefore, the ability to draw valid inferences from survey data depends heavily on the quality of these data. The statistical field of Survey Methodology investigates how to construct high-quality surveys.

This thesis contributes to the field of survey methodology by analyzing correlates of survey nonresponse and measurement error. We are able to use the administrative data from the German federal employment agency to study these error sources. The data cover a huge part of the German population and hold rich information on employment histories, job seeking, labor market program participation, and receipt of welfare benefit. The samples of the surveys we are analyzing are drawn from administrative records so that we have useful information on survey respondents *and* nonrespondents. Respondents are asked to allow us to link their survey responses to the administrative data to enrich the data set. For all respondents who consented to the data linkage, we are able to compare some of their survey responses to their records in the administrative data which we use as a gold standard.

Using these data, three studies analyzing survey data quality are presented. In the first study, the administrative data are used as a rich sample frame to analyze the effects of two different kinds of incentives on nonresponse bias within the German household panel survey "Panel Labor Market and Social Security" (PASS) conducted by the German Institute for Employment Research. Our question is whether nonresponse bias differs in two incentive schemes, unconditional prepaid cash and conditional post-paid lottery ticket. Using the administrative data as a gold standard for all sample cases, respondents and nonrespondents, we find that the former incentive scheme is more effective than the latter.

In the second study, we analyze the possibility to increase response quality by asking respondents to think hard about the next question and giving them a special incentive during the interview. We study different indicators for the quality of answers to the survey questions: The amount of "don't know"s and "no answer"s, as well as heaping and rounding of numerical items. Moreover, we are able to validate some survey responses by comparing them to the administrative records. We find the request to think hard and the additional incentives to affect item nonresponse in some of our questions. However, there is no evidence for general patterns regarding any of our quality indicators.

The third study is concerned with item nonresponse and measurement error in income questions. In particular, we investigate whether these two errors depend on a person's income itself, and whether this relationship is different for different survey modes, telephone and web survey. We find that item nonresponse and measurement error are related to a person’s income. Moreover, we confirm the often reported finding of mean-reverting measurement error. While the relationship between item nonresponse and a person’s income is different for the web and telephone modes, the modes do not show different relationships between measurement error and a person’s income.

Zusammenfassung

Viele Studien der empirischen Sozialforschung überprüfen den Wahrheitsgehalt soziologischer Theorien durch die statistische Analyse von Befragungsdaten. Die Aussagekraft solcher sozialwissenschaftlicher Studien hängt somit entscheidend von der Qualität der zu Grunde liegenden Daten ab. Die statistische Teildisziplin der Survey-Methodologie untersucht, wie Umfragen so durchgeführt werden können, dass die Qualität der erhobenen Daten möglichst hoch ist.

Der zentrale Beitrag der vorliegenden Arbeit zur survey-methodologischen Forschung ist die Analyse der Einflussfaktoren von *Nonresponse* (Teilnahmeverweigerung) und *Messfehler*. Für die Analyse dieser beiden Fehlerquellen wurden Daten der Bundesagentur für Arbeit verwendet, welche einen Großteil der deutschen Bevölkerung erfassen und umfassende Informationen über die Erwerbshistorien, Arbeitssuchen, Teilnahme an Maßnahmen der Bundesagentur und den Bezug von Transferleistungen liefern. Die Stichproben für die im Rahmen dieser Dissertation analysierten Umfragen wurden aus den Daten der Bundesagentur gezogen, so dass Informationen sowohl über Respondenten als auch über Non-Respondenten verfügbar sind. Die Respondenten wiederum wurden um ihre Zustimmung zur Verbindung ihrer Umfrage- und offiziellen Daten gebeten. Für all diejenigen Respondenten, die dieser Verbindung zustimmten, können die im Rahmen der Umfrage gegebenen Antworten direkt mit den administrativen Daten der Bundesagentur, unserem *Goldstandard*, verglichen werden.

Die Arbeit präsentiert drei Studien, in welchen die Daten der Bundesagentur zur Untersuchung verschiedener Qualitätsaspekte der Befragungsdaten Verwendung finden. In der ersten Studie werden die administrativen Daten verwendet, um die Effekte zweier verschiedener Arten von Belohnungen (Incentives) auf den *Nonresponse Bias* in der deutschen Haushalts-Panelbefragung "Panel Arbeitsmarkt und soziale Sicherung" (PASS) zu untersuchen, welche vom Institut für Arbeitsmarkt- und Berufsforschung (IAB) durchgeführt wird. Die zentrale Frage ist, ob sich die durch Antwortverweigerung entstehende Verzerrung zwischen den beiden Belohnungsarten – im Voraus geleistete Barzahlung unabhängig von der Teilnahme vs. im Falle der Teilnahme versprochene Gabe eines Lotterie-Loses – unterscheidet. Die Analyse ergibt, dass die erstere Belohnungsart effektiver als die letztere ist.

In der zweiten Studie wird die Möglichkeit untersucht, die Qualität der Antworten durch zwei Arten von Maßnahmen zu erhöhen: Die zusätzliche Aufforderung, sich bei der nächsten Frage besonders anzustrengen, und eine zusätzliche Belohnung während des Interviews. Verschiedene Indikatoren der Antwortqualität werden hierzu untersucht: Der Anteil der Antworten "weiß nicht" und "keine Antwort" sowie die Anteile der Antworten, in denen sog. "Jubiläumszahlen" oder gerundete

Werte angegeben werden. Des Weiteren werden einige der Befragungsantworten mit Hilfe der administrativen Daten validiert. Es ergibt sich, dass die Aufforderung, sich anzustrengen, und die zusätzlichen Belohnungen für manche, aber nicht alle Fragen die Neigung zu den Antworten “weiß nicht” und “keine Antwort” beeinflussen. Allerdings zeigt sich kein klares Muster für die Qualitätsindikatoren.

Die dritte Studie befasst sich mit Antwortverweigerung und Messfehler in Einkommensfragen, speziell mit dem Zusammenhang dieser beiden Fehlerquellen und dem persönlichen Einkommen der Befragten, sowie der Frage, ob diese Zusammenhänge sich für unterschiedliche Erhebungsmodi – Web und Telefon – unterscheiden. Es zeigt sich, dass Antwortverweigerung und Messfehler mit dem persönlichen Einkommen zusammen hängen. Der häufig berichtete “mean-reverting”-Effekt des persönlichen Einkommens auf den Einkommensmessfehler wird bestätigt. Während sich der Zusammenhang zwischen Antwortverweigerung und persönlichem Einkommen zwischen den Erhebungsmodi unterscheidet, zeigen die Modi keinen Unterschied im Zusammenhang zwischen persönlichem Einkommen und Einkommensmessfehler.

Contents

List of Figures

List of Tables

1 Introduction

Survey data quality has many dimensions (Biemer et al., 2014; Biemer, 2010), and at least three dimensions can be distinguished (Biemer and Lyberg, 2003): accuracy, timeliness, and accessibility. A well known framework to discuss survey data accuracy is the total survey error framework (Groves and Lyberg, 2010; Groves et al., 2009). Total survey error (TSE) is defined as the difference of a population parameter and an estimation of this parameter based on a population survey. All steps in the survey process can contain errors that all sum up to total survey error. In general, two sides of the *survey life cycle* can be distinguished: first, the *measurement side* that starts by defining a construct to be measured, and then moves on with designing a question to measure this construct, collecting answers in interviews and editing the answers. Second, the *representation side*, which consists of defining a target population and a corresponding sample frame, drawing the sample, and surveying the sample of respondents. Groves and Lyberg (2010) point out that there are two inferential steps in all surveys, one in each of the two parts of the survey life cycle. The first inference is from the survey response of a single respondent to the measurement of an underlying construct. The second inference is from the group of survey respondents to the relevant target population.

Let us, for example, consider the problem of estimating the *general life satisfaction* of people living in Germany. The general life satisfaction is the underlying construct of primary interest which survey participants may be asked for by using, say, a 7-point happiness scale. At this stage, it is crucial that the scale really measures a respondent's underlying life satisfaction rather than, e.g., some temporary mood. The second inference is from the survey respondents to all people living in Germany. At this stage, the main challenge is to make sure that the survey respondents can be considered as a realization of a random sample from the German population.

The example illustrates that both inferential steps can be subject to a variety of errors. First, the measurement might not be valid for the underlying construct. A lack of *validity* is in our case given if the survey question fails to convey some general aspect of happiness. Second, the survey response might not equal the measurement, but show *measurement error*. This can be the case if a respondent is unable to choose the response option that matches her true happiness. Third, the sample frame might not cover all cases of the target population. *Coverage error* can, for example, occur if people who do not permanently live in Germany are part of the sample. Fourth, the sample might not be a random subset of the target population and lead to *sampling error*. This will for example be the case if listers do not follow random walk protocols. Fifth, survey respondents systematically differ from nonrespondents. Such *nonresponse error* can for example occur if only people who are satisfied with their lives respond to the survey. As errors caused in each step sum up to the TSE, leading to decreasing survey data quality, each step has to be implemented with caution.

All error sources can lead to severe bias in a survey estimate and/or increase the estimate's variance. A useful quantity that accounts for both aspects is the mean squared error (MSE). The MSE of a survey estimate is given by the squared bias plus the variance. The MSE is always a characteristic of a single estimate, and not a measure for the whole survey. Thus, the MSE can be large for some survey estimates while being small for others.

To compute the TSE, some sort of gold standard is needed to compare the survey estimate to, e.g., means and totals from census data. However, aggregate information for the target population does not suffice to decompose TSE to its error components. For example, underestimation of the mean age in a survey as compared to census data can be due to overcoverage of young cases in the target population, oversampling of young cases from the sample frame, higher response rates among younger sample cases, or underreporting of age in the survey. Only if the gold standard information is available for all cases of the target population on an individual level, the single components of TSE can be identified.

For this thesis, we are able to make use of information from administrative records from the German Federal Employment Agency (Bundesagentur für Arbeit) that allow us to decompose TSE to its single error sources. We primarily focus on two components of TSE: nonresponse and measurement error.

Nonresponse The fact that not all sampled cases are responding to a survey does not necessarily lead to a bias in survey estimates, but always affects the variance part of the MSE (Groves and Peytcheva, 2008). However, if nonrespondents differ from respondents systematically, estimates based on survey respondents will lead to biased estimates of the target population. Differences between respondents and nonrespondents can usually only be observed for some socio-demographics that are part of the sample frame, but not for key survey variables. As bias is not directly observable, but the response rates are, and because it is often assumed that high response rates lead to higher quality than low response rates, many effort is put into increasing response rates. Response rates are found to depend on many aspects of the survey (Groves et al., 2002): professional invitation letters, recruitment, interviewer behavior and interviewer-respondent interaction in personal surveys, the survey topic, the survey burden, the ease of the questionnaire, the number of call attempts, the use of incentives, and many more. All aspects that influence response rates might affect nonresponse bias as well. In this thesis, we will study nonresponse bias by comparing respondents and nonrespondents based on administrative data.

Survey measurement error is defined by Groves et al. (2009, p. 52) as

> *"a departure from the true value of the measurement as applied to a sample unit and the value provided".*

Differences between the true and reported values can exist for many reasons: a respondent might not understand the question, not know the answer, not be willing to search her memory for the right answer, or just decide to report a wrong value. Respondents might decide not to answer a question, or to give a wrong answer because the correct answer is socially not desired, or they do not want to share the information because it is too personal and private.

If respondents are not motivated to give the best answer they can, they can choose to give a satisficing answer, i.e., an answer that is acceptable or appears to be reasonable (Krosnick, 1991). Satisficing can take various forms, like rounding and heaping of continuous data, randomly choosing answer categories, non-differentiation, the choice of middle categories, and acquiescence. The propensity for a respondent to show satisficing behavior is found to be positively correlated with the difficulty of the task, and negatively correlated with the respondent's motivation and the respondent's ability to solve the task. Measurement error in general can be due to characteristics of the questions and characteristics of the respondent. Strategies to minimize measurement error should therefore be applied on both levels, i.e., questions need to be as clear, unambiguous, and easy to understand as possible, and the interview burden needs to be as low as possible to keep the respondent motivated. Like nonresponse error, measurement error will always affect the variance part of the MSE. If it is systematic, it will also have a negative impact on the bias part.

Two design features that are often discussed in terms of nonresponse and measurement error will be analyzed in this thesis: the use of incentives (Singer and Ye, 2013) and the choice of the survey mode. It is well known that incentives can effectively be used to increase response rates — cash incentives are more effective than in-kind incentives, prepaid ones are more effective than post-paid ones, and higher incentives are found to increase response rates more than lower ones — but it is not clear whether incentives also affect nonresponse *bias* and measurement error (Medway, 2012). Although many studies exist that analyze the effect of incentives on sample composition and satisficing behavior, studies using validation data are rare. The influence on survey response has often been analyzed by comparing response distributions between incentive groups. However, such an approach does not allow to attribute differences to one of the error sources: distributions might differ because different sample cases select themselves to the survey, or because measurement error differs between the groups. Different survey modes — e.g. personal interviews, telephone interviews, and web surveys — are known to attract different kinds of sample cases. There is also some evidence that measurement error differs by survey mode (Kreuter et al., 2008). This work contributes to a deeper understanding of the effect of incentives and survey mode on nonresponse and measurement error.

The remainder of the thesis is structured as follows: The data sources that are used throughout the thesis are introduced in Chapter 2. Chapter 3 is an evaluation of the effects of different incentive schemes on nonresponse bias in the large-scale German panel survey "Panel Arbeitsmarkt und soziale Sicherung".

Chapter 4 deals with the question whether survey measurement error can be decreased by within-survey incentives. Chapter 5 studies the effect of telephone vs. web mode on item nonresponse and measurement error in income questions.

Contributions by Advisors

The main parts of Chapter 3, "The Effect of Monetary Incentives on Nonresponse Bias in a Household Panel", are joint work with Gerrit Müller, Frauke Kreuter, and Joachim Winter. The incentive experiment that is analyzed in the Chapter was planned and conducted by the team of the "Panel Arbeitsmarkt und soziale Sicherung" of the Institut für Arbeitsmarkt- und Berufsforschung. The paper idea was put forward by Gerrit Müller, who also gave me advice when analyzing the data and writing the manuscript. Frauke Kreuter and Joachim Winter provided helpful suggestions.

The fourth Chapter, "The Influence of Within-Survey Incentives on Response Quality in a Web Survey", is joint work with Frauke Kreuter and Joachim Winter. It was Frauke's idea to design an experiment to ask respondents to give good answers, and she gave me the opportunity to field my experiment in an IAB survey. Setting up the experiment was done in close cooperation with Joachim Winter and Frauke Kreuter, and both advised me when writing the Chapter.

Chapter 5 "Item Nonresponse and Measurement Error in Income Questions" benefitted greatly from suggestions by Joachim Winter and Frauke Kreuter.

2 Data Sources

Throughout the thesis, survey data quality is evaluated by linking survey data to administrative records of the German Federal Employment Agency (FEA). All surveys were conducted by the Institute of Employment Research (IAB) and are described in the following Sections.

2.1 Administrative Data of the German Federal Employment Agency

The FEA holds the employers' notifications to the German pension fund (employment history (Beschäftigtenhistorik BeH)) and process-generated data from the FEA. These contain the benefit recipient history (Leistungshistorik Grundsicherung (LHG) and Leistungsempfängerhistorik (LeH)), records about participation in labor market programs (Maßnahmeteilnahmehistorik (MTH)), and job seeker history (Arbeitsuchendenhistorik (ASU)). The data are only available internally to researchers working for or associated with the IAB, in form of single-data products or combined in the integrated employment biographies (IEB). The latter is also available for external researchers in the form of a subsample of the IEB (Oberschachtsiek et al., 2009).

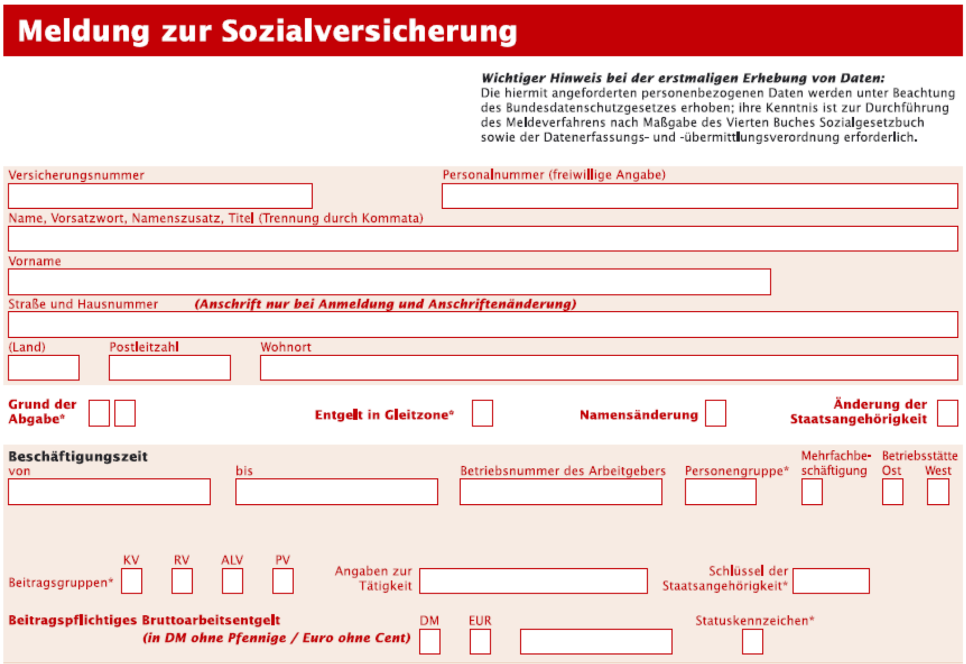
The IEB data contain

- individual day-to-day employment and wage spells from 1975 to present. The data do not contain information on civil servants or self-employed. In total, nearly 80% of the German labor force are covered (and almost 100% of the employees liable to social security).
- individual unemployment insurance spells since 1975.
- detailed information on participation in training programs and individual job seeking history since 2000.
- social security benefit information since 2005 on the level of benefit communities.

Our analyses focus on the BeH and LHG which are explained in more detail.

2.1.1 BeH

Every employee who holds a job subject to the social security system is covered by the BeH data set. Employers are obliged to regularly transmit detailed information about their employees (Meldung zur Sozialversicherung) to the German pension fund. The information is used to calculate social security contributions, pension claims, and unemployment insurance. The parts of the employer's notification that contain the information that is analyzed in this thesis can be found in Figure 2.1. Each notification includes personal information about the employee, i.e., the security number, name, and place of residence. Employees are sent a copy of the employer's notification to check the information and inform their employer about errors. Employers are asked to send their corrections to the social security system, and the administrative data are updated accordingly.



Meldung zur Sozialversicherung

Wichtiger Hinweis bei der erstmaligen Erhebung von Daten:
Die hiermit angeforderten personenbezogenen Daten werden unter Beachtung des Bundesdatenschutzgesetzes erhoben; ihre Kenntnis ist zur Durchführung des Meldeverfahrens nach Maßgabe des Vierten Buches Sozialgesetzbuch sowie der Datenerfassungs- und -übermittlungsverordnung erforderlich.

Versicherungsnummer

Personalnummer (freiwillige Angabe)

Name, Vorsatzwort, Namenszusatz, Titel (Trennung durch Kommata)

Vorname

Straße und Hausnummer ***(Anschrift nur bei Anmeldung und Anschriftenänderung)***

(Land)

Postleitzahl

Wohnort

Grund der Abgabe* □□ **Entgelt in Gleitzone*** □ **Namensänderung** □ **Änderung der Staatsangehörigkeit** □

Beschäftigungszeit

von

bis

Betriebsnummer des Arbeitgebers

Personengruppe*

Mehrfachbeschäftigung □

Betriebsstätte Ost □ West □

Beitragsgruppen* KV □ RV □ ALV □ PV □

Angaben zur Tätigkeit

Schlüssel der Staatsangehörigkeit*

Beitragspflichtiges Bruttoarbeitsentgelt
(in DM ohne Pfennige / Euro ohne Cent) DM □ EUR □

Statuskennzeichen* □

Figure 2.1: Employer's notification form to social security.

In the following, the items on the notification form that are of special interest for our analyses are described in more detail:

Grund der Abgabe denotes the reason for the notification. There are three main groups of reasons for notifications: registration of a new employment, notification of the end of an employment, and yearly notifications or special notifications. Begin and end of employments have to be notified to the pension fund immediately. The most common is the yearly notification that covers the whole year. It has to be sent to the pension fund for every employee in every year. In general, all changes in the employment situation that are important for the social-security system need to be transmitted to the German pension fund. We are able to construct full employment episodes from the notifications when taking the reasons for notification into account.

Beschäftigungszeit Each notification includes the exact start and end date of every employment. For the yearly notification, start and end date are January 1st and December 31, respectively. The maximum length of the spell is one year.

Personengruppe An indicator of the kind of job is also included in each notification. For us, this information is important as it indicates whether a respondent has a regular job, or is marginally employed.

Angabe zur Tätigkeit Additional information about the employment type is reported in a 9-digit number that includes keys (Tätigkeitsschlüssel) for the kind of job, information about educational degree, vocational training (Berufsausbildung) of the employee, and information on the contract. The contract information includes an indicator for subcontract work (Leiharbeit), and indicators for fixed-term contracts as well as full- or part time employment. We use this information to derive indicators for part-time and full-time employment.

Schlüssel der Staatsangehörigkeit The nationality of the employee can also be found in the notification.

Beitragspflichtiges Bruttoarbeitsentgelt The gross income for the whole employment spell is notified to the social security system. Usually, this includes monthly payments and bonus payments, if existent. However, in some cases employers decide to send extra notifications for bonus payments. Note that income notifications are censored by the social security contribution assessment ceiling (“Beitragsbemessungsgrenze"): any amount above this threshold is censored at the assessment ceiling which is different for every year and also differs between East and West Germany. The income reports are known to be very accurate as they are very important and misreporting is sanctioned. We use the gross income information to derive a respondent’s monthly income.

The notifications also contain additional information on the job and a unique establishment number. All notifications, as of 1975, are collected in the BeH data set. Every notification is thereby represented in one row of the data set. A short data example that is reduced to key variables can be found in Table 2.1. Note that a person can have different spells at the same time, as can be seen for the person with ID 2 in the example who has two mini jobs in 2013. Bonuses are reported separately from the monthly income in a special notification for person 2. The start and end dates of this notification correspond to the end date of the yearly employment notification.

ID	start	end	job	reason for notification	income in spell
1	17.08.2010	31.12.2010	training	yearly notification	55000
1	01.01.2011	30.06.2011	training	end of employment	58005
1	01.07.2011	31.12.2011	training	yearly notification	183.16
1	01.01.2012	31.12.2012	training	yearly notification	183.16
1	01.09.2013	31.12.2013	regular	yearly notification	70
2	15.03.2011	31.12.2011	regular	yearly notification	110
2	01.01.2012	31.07.2012	regular	maternity leave	112
2	31.12.2012	31.12.2012	regular	special payment	100
2	01.10.2013	31.12.2013	mini job	yearly notification	13.34
2	01.11.2013	31.12.2013	mini job	yearly notification	13.34

Table 2.1: Example of BeH data for two employees containing employment times, reason for notification, and income. Data are organized in longitudinal format including multiple spells for each employee.

2.1.2 LHG

The LHG data are a process-generated data of the FEA that contain start and end dates of all periods in which unemployment benefit II (Arbeitslosengeld II, UB II in the remainder) is received. The data set is available on individual level even though UB II is given at benefit community level ("Bedarfsgemeinschaft"). A benefit community consists of at least one person capable of work. Partners living in the same household are also part of the benefit community, and unmarried children under 25 years living in the same household can also be included if they can not meet their requirements through their own income or funds themselves. Children older than 25 constitute a separate benefit community even if they are living in the same household as their parents. In addition to the individual identification number that can be merged to the social security number, every benefit community has an unique identification number in the LHG data set. Thus, every individual in the LHG data set can be assigned to a benefit community.

Both, the BeH and LHG data sets are known to be very reliable (Jacobebbinghaus and Seth, 2007) as they are are not only collected for statistical purposes but to calculate pension claims and administer benefit claims and payments.

Many of the surveys that are conducted by the IAB are sampled from the administrative data of the FEA. This offers the great opportunity to link the survey data back to the administrative records to enrich survey responses by administrative data, or to compare survey responses and administrative records. However, data linkage is only allowed for respondents who explicitly agree to the linkage. More information on asking for linkage consent, correlates of consent, and possible consent bias can be found in Sakshaug and Kreuter (2012); Sakshaug et al. (2013); Sakshaug and Kreuter (2014); Sakshaug et al. (2012).

In the following Chapters, the linkage of survey and administrative data is exploited for assessing nonresponse bias and measurement error. The corresponding surveys are described in the following Sections.

2.2 Panel Study Labour Market and Social Security (Panel Arbeitsmarkt und Soziale Sicherung)

The panel "Labour Market and Social Security" (PASS) is an annual household panel conducted by the IAB. Its aim is to study the individual and social effects of the German "Hartz reforms" which came into effect in 2005 (see Trappmann et al., 2009). These reforms introduced the unemployment benefit II at the community level. PASS consists of two different samples to compare benefit recipients and non-recipients of the new UB II. About half of the sample from the first wave is sampled from the register of UB II recipients at the Federal Employment Agency. This is the *recipient sample* because all households had received benefits by the date of sampling (July 2006). The other half of the sample, called the *population sample*, is selected from a commercial database of residential addresses. In this sample, people with low socio-economic status are overrepresented. The first wave of data collection took place between December 2006 and July 2007. In every wave, refreshment households are sampled for the recipient sample. These households include at least one benefit community that received UB II in July of the respective year but did not receive UB II at any other time since July 2006. To counter panel mortality, a new sample was drawn for both subsamples before the fifth wave in 2011. The recipient refreshment sample consists of households with benefit units who received UB II in July 2010 (Jesske and Schulz, 2012).

Each sampled household in the survey receives a household questionnaire which is to be completed by the head of the household. Afterwards, each member of the household who is aged at least 15 receives a personal questionnaire or a questionnaire for elderly people. The household questionnaire contains questions about the household composition, dwelling, household income, and material deprivation, as well as received unemployment benefits. The personal questionnaire contains questions about the individual's employment status, employment history, and income. The interviews for PASS are collected in sequential mixed-mode design (CATI and

CAPI) whereas respondents can switch modes whenever they like. Interviews are conducted in German, Russian and Turkish.

In the third wave, there has been an incentive experiment that is described and analyzed in Chapter 3.

2.3 Work and Consumption in Germany (Arbeitsmarkt und Verbraucherverhalten in Deutschland)

The survey “Work and Consumption in Germany” was fielded by the IAB and in part funded by the LMU Munich. A gross sample of 24,236 eligible persons (aged 18 and older) was drawn from the administrative data (IEB, Version V09.00, IAB Nürnberg 2011) in June 2011 according to three non-overlapping strata. All sample cases had at least one security-contributing job in the last ten years. Cases from the first stratum further received UB II in the last five years, cases from the second stratum received unemployment insurance (UB I) in the last 10 years, and cases from the third stratum did neither receive income support nor income insurance, but held jobs with two or more different employers in the last 10 years.

Sample cases were randomly allocated to telephone or web mode: 12,400 cases were assigned to the telephone survey, while 11,836 cases were assigned to the web survey. Addresses for all sample cases were available on the sample frame. Additionally, for some of the cases telephone numbers were available, but no email addresses.

The telephone and web parts of the “Work and Consumption” survey cover the same topics while differing in some questions. Moreover, for both surveys the sample cases were contacted by invitation letters that were kept as similar as possible. Both surveys are explained in more detail in the next Paragraphs.

Telephone Survey (CATI) For 10,455 out of 12,400 cases, a telephone number was available on the sample frame. During fieldwork, about eleven percent of the phone numbers turned out to be invalid and could not be replaced by a working phone number from the public directories, or were ineligible. Among those persons who could be contacted (9,332), the target of 2,400 completed interviews could be met. During fieldwork, at least 20 contact attempts were made per case, at varying points in time such as weekdays and weekends, and different times of day. However, we did not attempt specific refusal conversion once a case expressed some mild form of refusal. Fieldwork was conducted in the months of August to October 2011. The average survey completion time was 21 minutes.

Web Survey The web survey was constructed to be very similar to the CATI survey, but differed in some questions. Considering layout and functionality of the web

	telephone survey	web survey
gross sample N	12,400	11,836
net sample (contactability)	9,332	10,525
completed interviews n	2,400	1,068
response rate (AAOPOR RR1)	19.35%	9.01%
cooperation rate (AAPOR CR1)	25.72%	10.15%

Table 2.2: Response rates across the CATI and web survey "Work and Consumption in Germany".

survey, we followed the guidelines for best practice given by Couper (2008). For instance, we presented only one question per screen, only used drop-boxes where applicable, and did not force the respondents to select an answer. The web survey was programmed and hosted by the LINK institute. Invitation letters were sent out including the URL of the web survey's website as well as personal user names and passwords. All sampled cases were also promised an incentive of €3 in form of an Amazon voucher for completing the interview. Respondents were asked to enter their email address at the end of the interview so that they could be sent the voucher after completion.

In total, we sent invitation letters to 11,836 sample cases, and 1,068 cases completed the survey. Overall, 1,311 letters were returned to sender due to an incorrect address. Thus, those respondents never received the invitation to participate. The response rate of this survey was 9%[1], the average completion time about 15 minutes. In order to reduce undercoverage and improve response rates, the invitation letters also offered the opportunity to call a toll-free number to conduct a shortened version of the interview on the telephone. Of the contacted cases, 312 people made use of the hotline. Out of those calls, 161 called to schedule an appointment for a telephone interview, resulting in an additional 132 interviews. The shortened telephone interviews are not part of the following analyses. We embedded an incentive experiment in the web survey of the "Work and Consumption in Germany" study that is described and analyzed in chapter 4.

Table 2.2 summarizes the sample sizes, response, and cooperation rates according to AAPOR standards (AAPOR, 2015) for both modes in the "Work and Consumption in Germany" study. Even though the web survey respondents received an incentive of €3, the response rate is much higher in the CATI than in the web survey. Lower response rates in web surveys, however, are usually found. Note that, while CATI sample members are defined to be contactable once an interviewer established contact to the respondent on the phone, we can only judge contactability of sample members

[1] There were 14 partial completions that will not be used for the following analyses. They are, however, part of some other studies.

of the web survey based on returning letters. Of course, it might be that letters were not returned even though they did not reach the target person. Therefore, the net sample size might be overestimated while the response rate is underestimated in the web survey. Thus, response rates should be compared with caution.

Item nonresponse and measurement error in the *income* question are compared between the two modes in Chapter 5.

3 The Effect of Monetary Incentives on Nonresponse Bias in a Household Panel

3.1 Introduction

It is well known that respondent incentives can increase response rates (James and Bolstein, 1990; Willimack et al., 1995; Church, 1993; Singer, 2002; Singer and Ye, 2013; Toepoel, 2012). It is also widely agreed that response rates per se are not a good indicator of survey quality and that other measures, such as nonresponse bias, should be taken into account (Groves and Peytcheva, 2008; Singer and Ye, 2013).

But much less clear is the effect of incentives on nonresponse bias. An increasing response rate does not necessarily decrease nonresponse bias. However, if incentives increase the response rates of some groups more than others, nonresponse bias is likely to be affected. An increase in response rates only for subgroups who are likely to participate anyway will increase nonresponse bias. Therefore, the desired effect of incentives is to bring people who are less likely to respond into the respondent pool. For example incentives that entice people with low socio-economic status or low levels of education, and ethnic minorities, who are typically underrepresented in surveys (Watson and Wooden, 2009) could be effective in reducing nonresponse bias.

Concerning the kind of incentive, literature shows that cash incentives increase response rates more than in-kind incentives, unconditional incentives have a higher effect than conditional ones and higher incentives increase response rates more than lower ones (Church, 1993; Singer et al., 1999; Ryu et al., 2006; Singer and Ye, 2013). The potential impact of incentives on underrepresented subgroups can be explained by leverage saliency theory (Groves et al., 2000). Under this theory, the decision to participate in a survey depends on various features of the survey, their relative importance to the sample case, and how salient they are made to the sample case. If an incentive is enough of a positive inducement to participate, overcoming the negative and less enticing features of the survey, it may pull people into the respondent pool who would not otherwise participate. The subgroups most affected by a monetary incentive are thought to be those for which money has a high importance.

Economic models of survey response stress that incentive payments may be perceived as compensation for the time and effort a respondent provides (Philipson, 1997); such models predict that a modest incentive should have a stronger effect on low income respondents (because their opportunity cost of time is lower).

Empirically, these predictions are broadly confirmed; for instance, Mack et al. (1998) find that incentives of $20 can disproportionately increase participation of respondents from poverty and Black households. Also Groves et al. (2006) find that people who are less interested in surveys can be brought into the pool by monetary incentives which serve as compensation for lack of interest. In a meta-analysis of incentive experiments in face-to-face and telephone surveys Singer et al. (1999) find some evidence that incentives can improve sample composition by increasing the response propensity for people who are otherwise underrepresented, such as low income people or non-whites. None of these studies examines the effects of incentives on nonresponse bias of survey statistics directly.

In this Chapter, we analyze the effects of incentives on the nonresponse bias of survey statistics using survey data from the German household panel PASS that was specifically designed to study people with low socio-economic status (in particular, welfare benefit recipients). In the third wave of this panel survey, an incentive experiment was conducted in which households were randomly given either a prepaid unconditional cash incentive or a post-paid conditional lottery ticket. The experimental groups differ by two characteristics of the incentives: conditional vs. unconditional and cash vs. in-kind incentive. Although the effects can not be separated, they are known to influence respondents in the same way (Church, 1993; Singer et al., 1999; Singer and Ye, 2013).

Our question here is whether prepaid unconditional monetary incentives affect nonresponse bias differently than a conditional post-paid lottery ticket. This is usually untestable because no valid information on nonrespondents is available and therefore a gold standard for comparison is missing. An important aspect of our study is that we have administrative data on the target variables for both respondents and nonrespondents and we can compute nonresponse bias directly by comparing respondents to the whole sample. We therefore address our research question first by comparing nonresponse bias for the two experimental groups using administrative data.

For most surveys this direct approach can not be applied as usually no administrative data are available for both respondents and nonrespondents. In the absence of such information, most studies have to rely on survey data on respondents to estimate nonresponse bias. To study this particular problem, we will in a second analysis approximate the nonresponse bias based on survey respondents, as proposed by Bethlehem (1988). We will then compare both methods as to the conclusions they draw about the effect of the two types of incentives on nonresponse bias. The analyses based on the administrative data will thereby serve as a gold standard for this comparison.

3.2 Data

3.2.1 Panel Data

For our analyses we use data from the first three waves of the German household panel survey PASS (see Chapter 2). We analyze the recipient and population samples jointly but use the sample indicator as a control in our statistical models when appropriate. Due to the overrepresentation of UB II recipients and low income people, we are able to analyze the effect of the two types of incentives on these groups who are often a small proportion of the sample in other surveys. Data are not weighted for our analyses because survey weights can not be linked to the administrative data on an individual level for all cases.

The incentive experiment was conducted in wave 3. In the first two waves, sample units received a thank-you card containing a stamp worth 55 cents with their advance letter, and responding households were given a German lottery ticket ("Aktion Mensch", worth €1.50 in the first, and "ARD-Fernsehlotterie" worth €5 in the second wave). In the experiment, panel households were randomly assigned to two treatment groups: one group was promised a lottery ticket (lottery group), and the other group was sent 10 Euros with their cover letter (cash group) (Büngeler et al., 2010). Note that the experiment compares a conditional in-kind incentive to an unconditional cash incentive. Also, the monetary values of each are different.

In total, 16,091 households were part of the PASS wave 3 sample, including 4,031 wave 3 refreshment households who were not part of the incentive experiment. We exclude 4,793 cases from our analyses who have only responded to one of the prior waves and focus on 7,267 panel households who responded to both waves 1 and 2. Of those, a randomly-selected 985 were part of another experiment and are omitted from our analyses. For this Chapter we analyze the remaining 6,282 households, 2,952 of which belong to the recipient sample and 3,330 to the population sample. Cases were randomly assigned to the experimental groups within these two sub samples. In total, 3,163 cases were part of the conditional lottery ticket incentive group and 3,119 cases were part of the conditional cash incentive group.

We are especially interested in the effect of incentives on estimates of personal income, which we split into terciles for our analyses (low, middle and high income). The income variable contains earnings from own employment only, i.e, social benefits are not included. In addition, we examine socio-demographic variables that usually affect response, like gender, nationality (foreign or German), employment status, UB II status, whether the person has a job of up to €400 per month that is not taxed and exempt from social insurance payments ("mini job") and age. Age is split into five categories: younger than 30, 30 – 39, 40 – 49, 50 – 59, and 60 or older.

3.2.2 Administrative Data

We use administrative data from the "Integrated Employment Biographies" (IEB) file provided by the Research Data Center of the Federal Employment Agency to compute the full sample's true value (see Chapter 2.1). This data set contains detailed employment and benefit records for all sample units, respondents and nonrespondents to the survey.

IEB data has been found to be very reliable concerning employment status, wages, and transfer payments (Jacobebbinghaus and Seth, 2007).

We only make use of variables that are part of both the administrative data and the survey data to be able to compare survey and administrative data findings. In our analyses, we assume that the head-of-household who filled out the household questionnaire is identical to the "head-of-household" according to the administrative records (see Sakshaug and Kreuter, 2012).

Administrative data are available as spell data. These contain multiple observations for each sample case which cover the beginning and the end of a span of time during which the case is in a certain state, like employed, unemployed or benefit recipient. The variables of interest are constructed from these data using reference dates. For respondents, the date of the household interview is used. For all nonresponding cases that were contacted in CATI, the date of the last contact was used. We do not have contact data for cases contacted in CAPI, so we use the end of the field period as the reference date for those cases.

3.2.3 Estimation Sample

In total, 5,179 of the 6,282 cases (82 %) who participated in waves one and two responded to the household interview in wave 3. For 5,063 of those, the head-of-household completed his personal questionnaire as well. For all units within the recipient sample, linkage to administrative data is straightforward as these data are part of the sampling frame. In total, for only five cases from the recipient sample administrative data could not be linked successfully. Most likely, these were wrong or temporary entries in the frame at the date of sampling and deleted from the records after the sample was drawn.

In contrast to the register-based recipient sample, cases from the population sample need to be searched for in the administrative records. However, due to data protection rules this is only allowed if linkage consent has been given beforehand. In the population sample, 79.07 % of the respondents gave their consent for data linkage, and 77.93 % of them could be found in the records using probabilistic record linkage procedures. Thus, 62 % of the population sample cases could be linked successfully (see Table 3.1).

	recipient sample	population sample
	n = 2,952	n = 3,330
available for linkage	100.00 %	79.07 %
part of the analyses	99.83 %	61.53 %

Table 3.1: Estimation sample for administrative data analyses.

Since the administrative data only contain information of individuals who received any kind of unemployment benefit, were registered as unemployed, or were employees subject to social insurance contributions, the cases which could not be linked are likely to be self-employed or civil servants. Given the analyses from Sakshaug and Kreuter (2012), we do not expect much consent bias in these data. In their analyses of the same survey, consent bias is only found for age and foreign citizenship, and it is very small compared to other bias sources, such as bias due to measurement error or nonresponse. Also, Beste (2011) finds that only respondents having a foreign citizenship and respondents who receive no income at all might be underrepresented in the linked data set. We do not expect that excluding respondents who could not be linked to the administrative data or did not agree to the linkage introduces significant biases into our analyses. For our analyses using the administrative data, we use the 2,947 cases of the recipient sample and the 2,049 cases of the population sample that were successfully linked to the administrative data — which makes 4,996 cases in total.

In order to apply the survey and the administrative data analysis methods to comparable estimation samples, cases that had to be excluded from the administrative data analysis because they could not be linked were excluded from the analysis using the survey data only. We checked this restriction and found that survey data findings do not change substantially when including the previously omitted cases. In addition, varying degrees of item nonresponse to the survey variables of interest resulted in differing numbers of cases available for the analyses (see Table 3.2).

3.3 Methods

For our bias analyses we use two different methods. First, we compute nonresponse bias directly using administrative data for respondents and nonrespondents of the survey as a gold standard. Since most surveys can not be linked to administrative

variable	survey data	administrative data
income	1669	2401
age	3995	4996
UB II	4078	4996
female	3995	4996
foreign	3993	4996
mini job	3995	4996
employed	3995	4996

Table 3.2: Cases used for the analyses

data, this method usually can not be used. We provide the nonresponse bias approximation proposed by Bethlehem (1988) as a complementary second method that will be compared to the gold standard.

In general, the bias of a statistic is given as the difference of the statistic's expectation and the true population value. Thus, the bias in the mean statistic of a random variable Y, using observations y, is given by:

$$\text{bias}(\bar{y}) = E(\bar{y}) - \bar{Y} \tag{3.1}$$

This bias can result from many sources (e.g., undercoverage or overcoverage, measurement error, nonresponse). The estimation of nonresponse bias in the administrative data is straightforward as information is available for all sample cases and other error sources can be neglected. Using administrative data, nonresponse bias is given as the difference of the estimated mean using the respondents only and the estimated mean of all sample cases:

$$\widehat{bias}(\bar{y}_{admin}) = \hat{\bar{y}}_{admin,respondents} - \hat{\bar{y}}_{admin,sample} \tag{3.2}$$

In order to be able to compare nonresponse bias statistics across variables, we use a measure of relative bias which equals the estimated bias standardized by the full sample mean of the respective variable.

$$\widehat{rel.bias}(\bar{y}_{admin}) = \frac{\widehat{bias}(\bar{y}_{admin})}{\hat{\bar{y}}_{admin,sample}} * 100 \tag{3.3}$$

When no administrative data are available, nonresponse bias can be approximated using survey data only. As Bethlehem (1988) shows, nonresponse bias in the mean statistic of y, based on survey respondents only, can be approximated by

$$bias(\bar{y}) \approx \frac{cov(y, \rho)}{\bar{\rho}} \tag{3.4}$$

with $\bar{\rho}$ being the mean response propensity of the (sub)population and $cov(y, \rho)$ being the covariance of the response propensity and survey variable y. The formula shows that for a given covariance an increasing mean response propensity always has a decreasing impact on nonresponse bias. Also, a decreasing variance in response propensity leads to decreasing bias even if the covariance of response propensity and survey variable stays the same (see Peytchev et al., 2010). The estimated nonresponse bias for wave 3 using the survey method is then calculated as

$$\widehat{bias}(\bar{y}_{survey}) \approx \frac{\widehat{cov}(y_{respondents\ wave3}, \hat{\rho}_{wave3})}{\bar{\hat{\rho}}_{wave3}} \tag{3.5}$$

where $\hat{\rho}_{wave3}$ is the estimated wave 3 response propensity and $\widehat{cov}(y_{respondents\ wave3}, \hat{\rho}_{wave3})$ is the estimated covariance of the wave 3 survey response and the estimated wave 3 response propensity. Since we are using panel data, survey data from previous waves can be used to estimate wave 3 response propensities ($\hat{\rho}_{wave3}$) for all sample units. Therefore, based on information from waves 1 and 2, wave 3 response (yes/no) is modeled dependent on treatment group and sociodemographic variables. We estimate the response propensities using logistic regression models. The covariates on the household level are all collected in wave two: household size, presence of children under four in the household, house ownership, UB II recipiency, whether the household lives in eastern or western Germany and adjusted household equivalence income. The household income is adjusted by the (old) OECD equivalence scale which weights the household income by the number of household members and their age. This specific scale assigns a value of 1 to the first household member, of 0.7 to each additional adult and of 0.5 to each child.

We use indicators for whether the head of household gave a personal interview and whether all household members gave personal interviews in the second wave. The model includes the number of missing values in the second wave for the set of variables everyone received. We additionally include personal characteristics of the respondent. These are age, gender, an indicator whether someone was born outside Germany, the weekly hours worked, marital status and education. For all variables, the most recent information available is used. We fill in missing age, gender, education and marital status from the first wave if available. Also, for people who did not fill out a personal questionnaire in the second wave at all we use the personal information collected in the first wave (180 cases). The model includes an indicator for recipient or population sample, whether the household in wave 3 was first approached by CATI or CAPI and whether the mode needed to be switched during the fieldwork. Finally, we include a dummy for the incentive treatment and interactions of the treatment with all other covariates. Of particular interest are interactions of incentives and variables that indicate social status, like unemployment benefit, household income, and the deprivation index, as these interactions reveal different impacts of incentives on different socio-economic groups.

There is a small amount of missing data in key variables that we impute for adjusted household income (28 cases missing), UB II recipient (13 cases missing), age (14 cases missing) and working hours in the last job (13 cases missing). The imputation model uses the same variables on household and personal level as the propensity models to make sure that it reflects the relations of these variables correctly. We perform multiple imputation using chained equations (m = 10). Chained equations run imputation models for each variable containing missing data conditional on the covariates of the imputation model and all other variables containing missing data. This approach is used because the imputed variables are of different type (continuous, categorical) and no joint density can be specified to perform imputation by joint modeling (Azur et al., 2011).

We predict the response propensity by first estimating a person's contact propensity and then estimating the probability of response, given contact. The response propensity is the product of these two propensities. We use a two stage estimation strategy to account for the sequential nature of the response process and to allow the covariates to have differential effects on contact and nonresponse (see Bethlehem et al. (2011)). The average pseudo R^2 over the 10 imputations is 0.1809 (min=0.1807; max=0.1813) modeling contact and 0.2494 (min= 0.2491; max= 0.2496) modeling response given contact.

For this second method, we also use the estimated relative bias to compare the conclusions drawn by both methods. As we do not have the wave 3 information for nonrespondents, we have to use the wave 2 survey mean of the full sample. The relative nonresponse bias using the survey method is then given by

$$\widehat{rel.bias}(\bar{y}_{survey}) \approx \frac{\widehat{bias}(\bar{y}_{survey})}{\hat{\bar{y}}_{respondents\ wave2}} * 100 \tag{3.6}$$

Thus, for the survey method, the standardization term has to be lagged by one year as compared to the standardization for the administrative data method. Depending on the time-varying nature of each considered variable, this is of course an approximation. However, the entire survey method is approximate and in most instances the only way to analyze nonresponse bias. Therefore we find it particularly interesting to compare the conclusions drawn by both methods, using the administrative data as the gold standard. To this we turn now. Firstly, based on estimates derived as stated in equation (3) we will compare the relative biases between the lottery and the cash group. Secondly, we will apply equation (6) to the survey data in order to see whether this method comes to similar conclusions as the administrative data method. Confidence intervals are computed using bootstrap, and 10,000 bootstrap replicates are computed for each relative bias estimation. The bootstrap 0.025 % and 0.975 % empirical quantiles serve as the 95 % confidence intervals.

3.4 Results

We find that the response rate is higher for the unconditional cash group (85.54 %) than for the conditional lottery group (79.39 %), and the difference is statistically significant ($p < 0.001$). Concerning response rates, we can see that in this experiment incentives work in the expected direction. However, as an increasing response rate does not ensure a decrease in nonresponse bias, we will next analyze the variable specific relative nonresponse bias for personal income and several socio-demographic variables.

Administrative Data Findings

Figure 3.1 shows the relative nonresponse bias (including 95 % confidence intervals) in income estimation for the two incentive groups using administrative data (for numbers see Table B.1 in Appendix B). Whereas there is no significant relative nonresponse bias for any income category in the cash group, there is significant relative nonresponse bias for the lowest and highest income category in the lottery group. For the lottery group, the proportion of people in the high income group is significantly overestimated and the proportion of people in the low income group is significantly underestimated. Even though the relative nonresponse bias is significantly different from zero for these two income categories in the lottery group, the lottery and the cash group do not significantly differ from one another for any of the income categories.

Figure 3.2 shows the relative nonresponse bias in the estimation of the proportions of people falling into the five age categories using administrative data. In trend, the

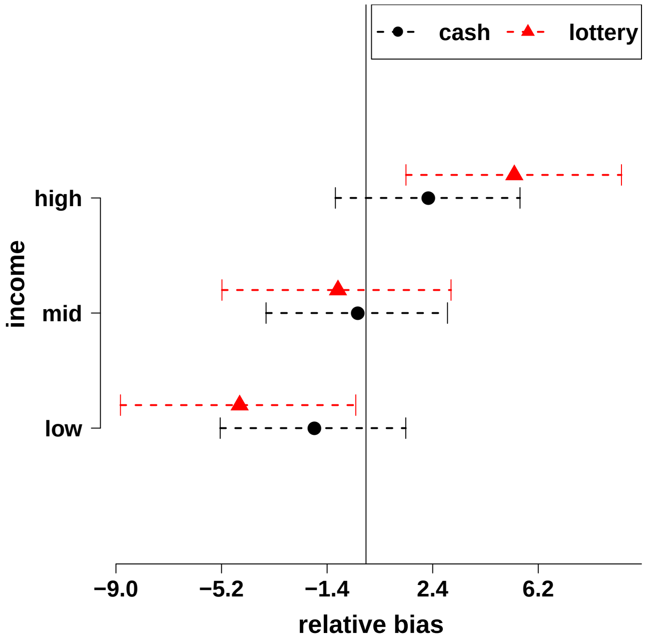


Figure 3.1: Relative nonresponse bias in income estimation using administrative data including 95% bootstrap intervals.

older age groups are overrepresented while the younger age groups are underrepresented. We find significant relative nonresponse bias for the lowest and second highest age group for both experimental groups. The relative nonresponse bias for the oldest age group is only significantly different from zero for the lottery group. Also, relative nonresponse bias in the extreme age groups is less for the cash group than for the lottery group. Summarizing, we can state that there is less relative nonresponse bias on age measures for the cash than for the lottery group, even though differences between the groups are not significant for any age group.

Figure 3.3 shows relative nonresponse bias for some socio-demographic variables. As for income and age, none of the variables show significant differences between the two incentive groups. Also, relative nonresponse bias is in the same direction for both experimental groups, except for the proportion of females. People who are born outside Germany are only significantly underrepresented in the cash group whereas relative nonresponse bias in the lottery group is not significantly different from zero. There is no significant relative nonresponse bias for the proportion of people having a "mini job" and the proportion of people being employed. Even though relative nonresponse bias for the proportion of females is in different directions for both experimental groups, it is not significantly different from zero for either group. The

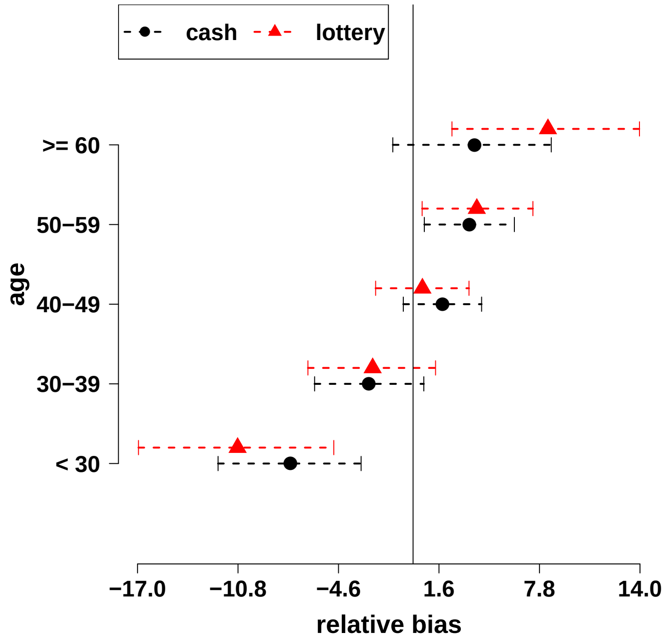


Figure 3.2: Relative nonresponse bias in age estimation using administrative data including 95% bootstrap intervals.

proportion of people receiving UB II does not show any significant relative nonresponse bias.

We combined the population and recipient samples for our analyses because we were interested in nonresponse bias for the whole study. However, when we looked at the two subsamples separately, we found that the relative bias for income and UB II was smaller in the more homogenous sample of recipients than in the population sample. Also, it was not significantly different from zero.

Survey Data Findings

The relative nonresponse bias in the income estimation shows very similar patterns across the income groups for the survey data approximation (see Figure 3.4) as for the administrative data method (see also Table B.2 in Appendix B). That is, we find no significant relative nonresponse bias in income estimation for the cash group and no significant differences between the experimental groups. For the lottery group however, the survey method fails to detect significant relative nonresponse bias for the highest income group.

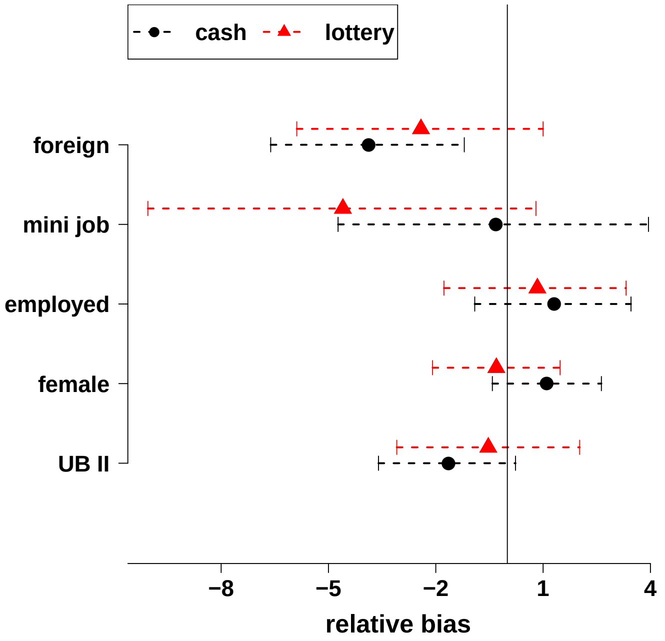


Figure 3.3: Relative nonresponse bias in estimation for socio-demographic variables using administrative data including 95% bootstrap intervals.

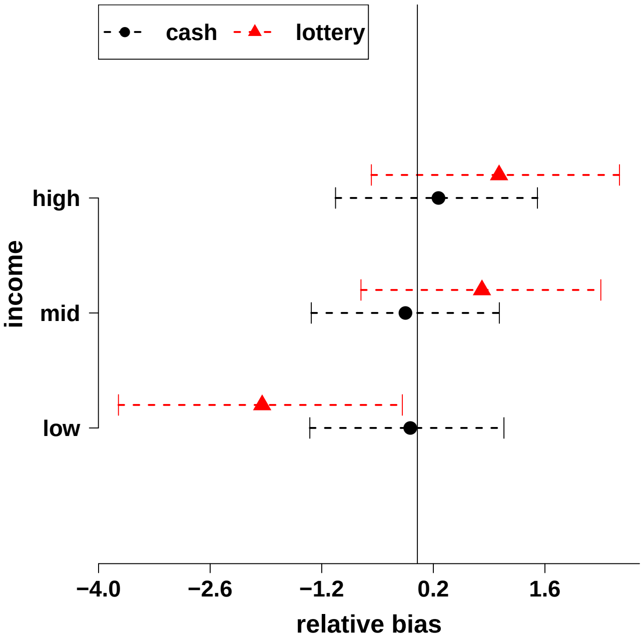


Figure 3.4: Relative nonresponse bias in income estimation using survey data including 95% bootstrap intervals.

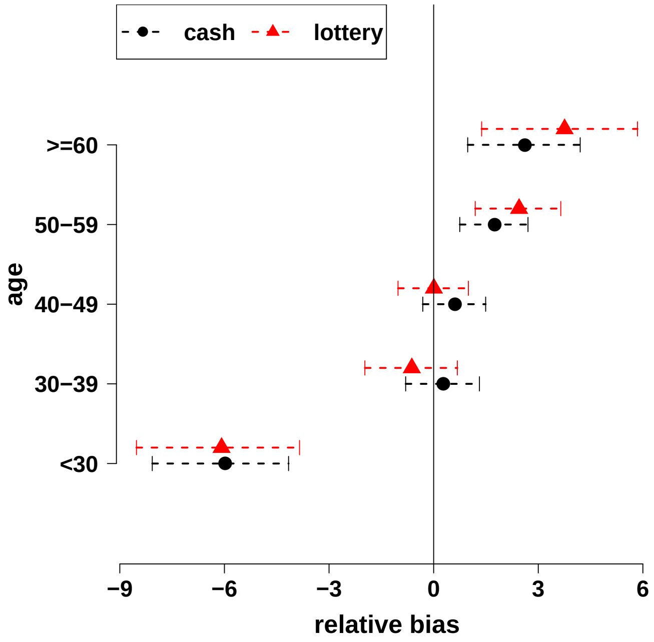


Figure 3.5: Relative nonresponse bias in age estimation using survey data including 95% bootstrap intervals.

Figure 3.5 shows the relative nonresponse bias in the estimation of people falling into the five age categories using survey data. Again, the survey data method shows very similar relationships as the administrative data method: the older age categories are overestimated and the younger categories are underestimated. The relative nonresponse bias is statistically significant for the youngest and second oldest age category. Also, the relative nonresponse bias in the oldest age category is found to be significantly different from zero for the lottery group. However, the survey method wrongly finds significant nonresponse bias for the oldest age category within the cash group. This difference to the administrative data analysis might be explained by the findings of Sakshaug and Kreuter (2012) who find a small consent bias for age. Again, we do not find any significant differences between the experimental groups.

The estimates of relative nonresponse bias in socio-demographic variables (Figure 3.6) show a similar pattern as the administrative data estimates. Again, none of the variables shows significant differences in nonresponse bias between the incentive groups. However, the survey method wrongly finds a significant overrepresentation of employed respondents and significant underrepresentation of UBII recipients in the cash group. Also, the survey method wrongly indicates a significant underrepresentation of females for the lottery group.

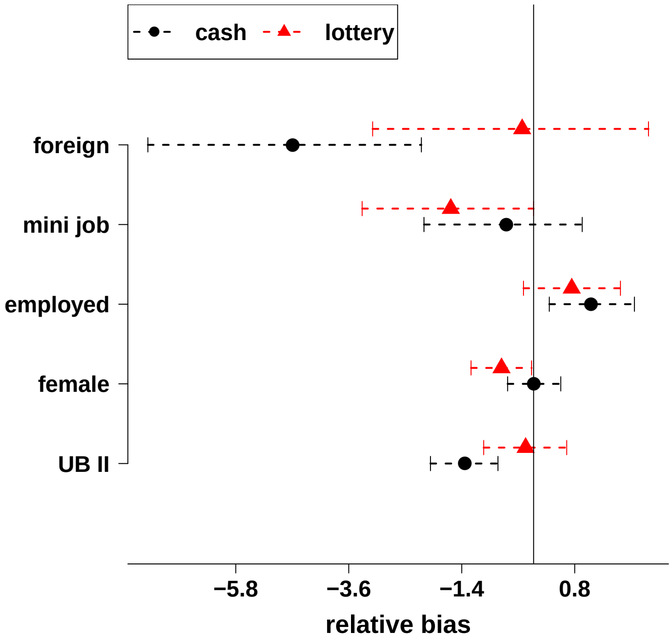


Figure 3.6: Relative nonresponse bias in estimation for socio-demographic variables using survey data including 95% bootstrap intervals.

3.5 Summary and Conclusions

We find that — compared to the conditional post-paid lottery group — unconditional prepaid cash incentives reduce the relative nonresponse bias for income. This reduction, however, is not significant as the confidence intervals for the relative bias estimates are overlapping. Also, there is no significant bias in income measures for the cash group. Finally, unconditional prepaid cash incentives reduce nonresponse bias in estimates of the age distribution and of the proportion of people having a "mini job", which are both related to income.

Although our analysis focuses on the aggregate survey outcome and we do not know how incentives work at the individual level, we think that these findings could be explained by leverage-saliency and economic theory. According to leverage-saliency theory the impact of cash incentives should be highest for people who are missing other motivation to participate in the survey and for whom money has the highest importance. Therefore, it might be that low income people are over-proportionately attracted by the cash incentive. Also, low income people have lower opportunity cost for survey participation and according to economic theory, a modest incentive of €10 will be more attractive to them as compared to people with a higher opportunity cost of time.

Our empirical approach has some limitations. The panel survey in which the incentive experiment was implemented overrepresents low income people.

Also, we were only able to link sample cases to the administrative data that gave consent. Since consent bias is usually found to be very small and cases were randomly assigned to the experimental groups, differences between the groups are not expected to be affected by consent bias. Furthermore, the attrition bias we analyzed in this Chapter might be different from nonresponse bias in the first wave of a panel or in a cross sectional survey. Also, the experiment alters two characteristics of the incentive that can not be separated.

Regarding our assessment of how well nonresponse bias can be characterized using survey data alone (the "survey method"), we find this to be a useful tool for analyzing nonresponse bias when no administrative data can serve as a gold standard. The bias estimates using the survey method broadly reflect the patterns of relative nonresponse bias using the administrative data but seem to fail to give correct confidence intervals for some variables. One possible explanation for the differences between the two methods could be measurement error in the survey variables. Since the survey approximation necessarily has to rely on the estimation of propensity models, differences could also be due to limitations of these models.

Our findings confirm that the unconditional cash incentive increases the response rate compared to the conditional lottery ticket. But what is more important, it decreases nonresponse bias in the key variables of the survey at the same time. Cash incentives have proved useful to decrease nonresponse bias in this low income and benefit-related survey. Future research should investigate whether this finding generalizes to other surveys and target populations.

Turning to the implications of our findings, we recognize that our finding that monetary incentives are especially effective for low income people might raise ethical concerns (Grant, 2002). According to the Helsinki Declaration, incentives would be called ethically problematic if they are coercive or if they give undue inducement to respond to surveys. While it is widely agreed that incentives can never be coercive (Wertheimer and Miller, 2008; Grady, 2001) since they are an offer to the respondent and not a threat, some scientists worry that incentives might be an undue inducement to respond. This would be so if they lead to respondents participating against their will and "against their better judgment" (Grant and Sugarman, 2004). Since incentives and other forms of payment are always used to induce people to do something they would not do otherwise, ethical problems might arise if incentives are used to induce people to do something they are verse to. One could argue that is the case if the offer is so attractive that respondents are not able to refuse and they decide to participate even though that means they have to take risks that they would not be willing to take otherwise. One might, however, argue that the incentive compensates an individual for her willingness to take certain risks, and that ethical concerns should be implemented using absolute standards that rule out a risk entirely if it is deemed unethical. In any event, studies in survey research do not provide evidence supporting the notion that higher incentives lead to acceptance of higher risks: In an experimental setting, Singer and Couper (2008) find that monetary incentives raise the willingness to participate to hypothetical surveys, and that willingness declines with risk of disclosure of sensitive data. But they do not detect a significant interaction between the amount of the incentive and the risk of disclosure. Similar relationships are found for two studies that explore incentives in the context of clinical studies. Halpern et al. (2004) find that people's willingness to participate in a hypothetical clinical trial increases as the payment level rises and as the risk of adverse effects or the risk of being assigned to the control groups decreases. They find no significant interaction between the risk people are willing to take and the incentive level. Bentley and Thacker (2004) also find that willingness to participate in studies is higher if incentives are higher and decreases with greater risk. They also do not find any interaction of incentive level and risk on the willingness to respond. They also find that a respondent's rating of the study risk does not depend on the amount of incentives. Although the harm of disclosure of confident information in social surveys may not be comparable to the risks in clinical studies, and the incentives in survey are small compared to those given in clinical studies, the rationale is the same for both kinds of studies, and incentives in social surveys are discussed along these lines by ethic boards.

4

The Influence of Within-Survey Incentives on Response Quality in a Web Survey

4.1 Introduction

Even though a sampled person may agree to participate in a survey, she may not provide answers to all of the questions asked or give wrong or imprecise answers, resulting in item nonresponse and measurement error. Particularly in web surveys, where no interviewer is present, the prevalence of satisficing responses like "no answer" and "don't know" can result in a significant proportion of missing data. Also, satisficing strategies (Krosnick, 1991) might lead to coarse, e.g., rounded or heaped data. Since all these forms of "bad" survey answers can bias survey outcomes strongly, researchers aim at reducing satisficing behavior.

Cannell et al. (1981) found that a request to think hard about the next question and to try to be as precise as possible can significantly increase data quality of personal interviews: people receiving such a request report more items in open-ended questions, give more precise answers (compared to medical records), and do more often check outside sources. These findings also hold for telephone interviews, even though the effects are in general smaller for the phone than for the personal mode (Miller and Cannell, 1982). Also, one recent study by Smyth et al. (2009) found that extra instructions can increase both the number of topics provided in answers to open-ended questions and response times but at the same time increase item nonresponse in a web survey. Overall, there is evidence that special verbal instructions informing respondents about the importance of correct answers and asking them to think hard can increase data quality in terms of response times, validity, and length of answers to open-ended questions. However, it is not clear whether other forms of "bad" answers can be decreased as well. Also, it is not clear whether the positive effect of the instruction can be increased even further by adding some monetary incentive to it.

There are basically two conflicting theories about the effect of incentives on response quality: Self-perception theory (Hansen, 1980) suggests a decrease in response quality. This is, respondents who are offered an external incentive will not be able to develop intrinsic motivation and therefore will be likely to choose satisficing strategies. According to this theory, even respondents who are generally willing to provide good answers can lose their intrinsic motivation, and the quality of their responses might decrease. The opposite effect is expected by leverage-saliency theory (Groves et al., 2000). Incentives can increase motivation or compensate for missing motivation and lead to increasing effort and thus to better data quality. It can also be argued that receiving an incentive creates an obligation to reciprocate by giving good answers (Gouldner, 1960).

There is only few literature about the effect of incentives on response quality, and it is only concerned with initial incentives that are given to increase survey participation (Singer and Ye, 2013). Also, it typically focuses on item nonresponse and length of answers to open-ended questions as indicators of data quality only. The findings are ambiguous. Curtin et al. (2007) for example find no effects of incentives on data quality, whereas Jäckle and Lynn (2008) find that prepaid incentives increase item nonresponse. The latter still find a net information gain by incentives due to decreasing unit nonresponse. Medway (2012) looks at the effects of several different incentives on many different indicators of survey quality, such as item nonresponse, straight-lining, rounding, interview length and response accuracy, and also finds basically no differences between incentive treatments. However, none of these studies analyzes incentives that are specifically given to increase response quality. That is, initial incentives might not have an effect in addition to increased participation (Medway, 2012). This might be the case because the respondent feels that she already fulfilled her duty by responding, or because she no longer thinks about the incentive after the interview has started.

Most studies of response quality are not able to distinguish between sample selection and survey quality: respondents who are brought into the survey by an incentive might be different from people participating without an incentive, and differences in response behavior and response distributions between incentive treatments might be due to different sample compositions instead of different reporting behavior. To the best of our knowledge, there has never been the attempt to combine the request to think hard about the next questions with additional incentives given during the interview.

In the web survey "Work and Consumption in Germany", we conduct an experiment to study the effects of instructions to think hard together with incremental incentives on response quality. We assess the response quality using different measures. In the first part we look at the time spent answering the question as an indicator of the effort being made to provide good answers. In the second part, we investigate item nonresponse (“don't know” and “no answers”), and in the third part, we analyze rounding and heaping in one recall question and two questions about subjective expectation (on a 0%-100% scale). In the fourth part, we are able to validate the answers to four questions: one knowledge question for which the right answer is known and the same for everybody, and three recall questions that we are able to validate by linking survey answers to administrative records. With our experiment we seek to answer the following research questions:

> Do respondents who receive the request to think hard give *better* answers?
>
> Does giving additional incentives increase data quality *even further*?
>
> Are effects *stronger* when the additional incentive is higher?

4.2 Incentive Experiment

As explained in Chapter 2.3, all sample cases of the web survey of the “Work and Consumption in Germany” study were promised an incentive of € 3 for completing the interview in form of an Amazon voucher. Even though for personal interviews unconditional cash incentives have been found to increase response rates more than other kinds of incentives (Singer and Ye, 2013), they are less common in web surveys (Göritz, 2006, 2008) as they are usually harder to realize. For instance, no postal addresses of sample cases are known where the cash money could be sent to. Although in our case it would have been possible to send unconditional cash incentives, we decided to use the more cost-efficient conditional incentives. Moreover, Göritz (2004) does not find differences in data quality for different kinds of initial incentives in online access panels. The user names and passwords allow to identify the respondents and link the survey data to the administrative records. The register data combine information from various sources on employees and basic income support recipients (“Unemployment Benefit II”, short UB II) in Germany (see Chapter 2.1). Due to German data protection laws, linking survey and administrative data is only possible for respondents who explicitly consent to data linkage (62.28 % in our survey).

The incentive experiment was placed towards the end of the interview, right before the socio-demographic section. For the experiment, respondents were randomly assigned to eleven treatment groups. Ten experimental groups were requested to take their time and think hard about the answers to the next questions, and to be as precise as possible. Of these ten groups, nine were further told that — as a thank-you in advance — an additional incentive was now added to the Amazon voucher that the were already promised to receive for completing the interview (see Chapter 2.3). This additional incentive was assigned randomly and ranged from € 0.50 to € 4.50. In this experiment, there are two control groups: one group received the request to think hard but no incentive, and the other did neither receive a request nor an incentive. This setting allows us to differentiate between the effect of the additional incentive and the effect of the request for extra effort.

Table 4.1 gives an overview of the treatment groups of the incentive experiment. Note that we exclude three respondents from our analyses because they obviously went back and forth in the interview and we were not able to keep track of that. We further exclude two cases who reported unreasonable years of birth (1900 and 9999) leaving us with 1063 cases for our analyses. The random assignment was performed right before the experiment started, and was independent of other experimental conditions. Also, we do not find any evidence that socio-demographic groups are not randomly represented in the experimental groups. Only the three strata are not equally distributed across the groups, and are therefore controlled for in the multivariate analyses. In total, 33 respondents broke off the interview after the experimental intervention started, and we do not find significant differences in break-off rates between treatment groups. Break-offs are in generally not part of our analyses.

treatment	request to think hard	amount of additional incentive	n
no incentive, no request	no	€ 0	116
request only	yes	€ 0	106
3	yes	€ 0.50	98
4	yes	€ 1.00	92
5	yes	€ 1.50	106
6	yes	€ 2.00	87
7	yes	€ 2.50	87
8	yes	€ 3.00	110
9	yes	€ 3.50	87
10	yes	€ 4.00	87
11	yes	€ 4.50	87

Table 4.1: Treatment groups included in our incentive experiment, the sample size is $n = 1063$.

The questions after the experimental intervention are taken from ongoing panel studies. That is, they have all been fielded and well tested, and we know that they are usually subject to item nonresponse and measurement error. We only chose questions for which we know which response patterns to expect, and for which we think that response quality can be increased by thinking hard enough about the answer.

The questions contain knowledge questions, recall questions referring to different time periods, sensitive questions, and questions about subjective expectations. Questions require either yes/no or numerical responses. The questions were divided into two blocks, "standard of living" and "health", of four questions each. The order of

question wording	answer type	abbreviation
"What do you think is the recommended daily number of calories for an average adult of your sex?"	numeric	calories
"How many times did you visit a doctor in 2011?"	numeric	doctor visits
"Have you ever been told by a doctor to have one of the following diseases?"	17 diseases listed from least to most common ; yes/no for each disease	diseases
"How likely do you think it is that you will live until age x?" (x depending on a persons age)	numeric (0%-100%)	life expectancy

Table 4.2: Questions within the health block of the web survey.

the blocks was randomized, but not the order of the questions within the block. The questions and answer types are shown in Tables 4.2 and 4.3.

In addition to the survey responses, we have time stamps for every question containing the date and time (to the precision of seconds) when the response was logged in. Our first question is whether there is a relationship between the additional incentives and the time the respondents spent on a question.

For all questions in the experimental block, we are first interested in the proportion of "don't know" and "no answer" responses. *Item nonresponse* is a big problem in all surveys as it can lead to biased estimates and a loss of precision (Rubin, 1976). In a second step, we are analyzing different forms of satisfising answers. Tables 4.4 and 4.5 give an overview of the satisfising behavior that is investigated for each of the questions. First, we are interested in the amount of *rounding* and *heaping* (rounding to prominent values) which also leads to biased estimates (Heitjan and Rubin, 1991). It is specifically problematic if the choice of the rounding interval is unknown and not random (Heitjan and Rubin, 1991). A special case which is very common in expectation questions, is rounding to 50%. There is some evidence that responding "50%" is very likely to be non-informative and can be viewed as an alternative way of saying "I don't know" (Bruine de Bruin et al., 2002, 2000; Manski and Molinari, 2010). Respondents who are very unsure about the true answer might prefer to say" 50%" just to fulfill the requirement of giving a numeric answer, but without having to commit themselves to a specific answer. In this sense, Gouret (2011) finds that 50% answers are mostly uninformative, and that this is the most often stated uninformative answer in expectation questions. For the recall questions that require a numeric response, heaping patterns are investigated according to patterns usually found for these same questions in different surveys. We study the extent of rounding monthly

question wording	answer type	abbreviation
“What was your last monthly income?”	numeric	income
“Did your household receive unemployment benefit (UB II) during the last 12 months?”	yes/no	UB II
“Please think of all employments in your life: how long was your longest period of employment you had without being unemployed in between?		
How many years	numeric	employment years
and months have you been employed in that period?”	numeric	employment month
“How likely do you think it is that your living standard will decrease in the next five years?”	numeric (0%-100%)	standard of living

Table 4.3: Questions within the finance block of the web survey.

income to multiple of €500. This pattern is frequently observed, for example in the PASS survey from which we took the income question. The doctor visits questions is similarly asked for in the German part of the SHARE survey, where strong heaps at 52, 24, 12, 6, 4, 2 are found. This corresponds to the answer heuristics “once a week”, “two times a month”, “once a month”, “every other month”, “every quarter year”, and “every half year”. Note that for the survey questions — with the exception of income and longest employment phase — we do not know a person's true answer. A rounded or heaped response is not necessarily wrong for a single person, and it only hints at a satisficing answer. Moreover, a value that is not rounded is not necessarily true. On aggregate, however, we assume that the decline in the proportion of rounded responses for a certain treatment group implies an increase in data quality. We also look at *measurement error* directly for some validation questions which can cause severe bias (Cochran, 1968) if it is non-random. For income and longest employment phase, we are able to control for the fact that seemingly rounded or heaped values can still represent a person's true answer. This is a quite unique situation, and the validation findings can be compared with findings using only indirect measures, which in most applications is the only way to assess response quality. Measurement error can be due to either reporting a wrong value on purpose (e.g. due to social desirability or data protection concerns), or to the respondents' incapability to provide the true

question type	question	research question
sensitive recall question referring to last months	income	rounding to multiples of 500, difference of survey and register data
sensitive recall question referring to the last 12 month	UB II	difference of survey and register data
recall question	employment year/employment month	rounding to multiples of 5 for employment year, difference of survey and register data
subjective expectation question	living standard	heaping to 50%

Table 4.4: Questions within the health block and research questions connected to them.

question type	question	research question
general knowledge question	calories	right answers (2000 kcal for everyone)
recall question referring to 2011	doctor visits	heaping to 52, 24, 12, 6, 4, 2
recall question	diseases	amount of yes-answers to most common diseases
subjective expectation question	life expectancy	heaping to 50%

Table 4.5: Questions within the finance block and research questions connected to them.

value (e.g. due to recall problems).

In the following, we subsume all kinds of undesirable answers — item nonresponse, rounding, heaping, and direct measures of measurement error — under the umbrella term “unsatisfactory response”.

4.3 Results

We assess our research questions by comparing the incentive groups for several indicators of survey quality: First, the experimental groups are compared with respect to the time to answer each question. Second, the amount of item nonresponse (“don't know” and “no answer”) in each question is compared. The third part contains a comparison of rounded answers between the groups. To this end, indicators for rounding

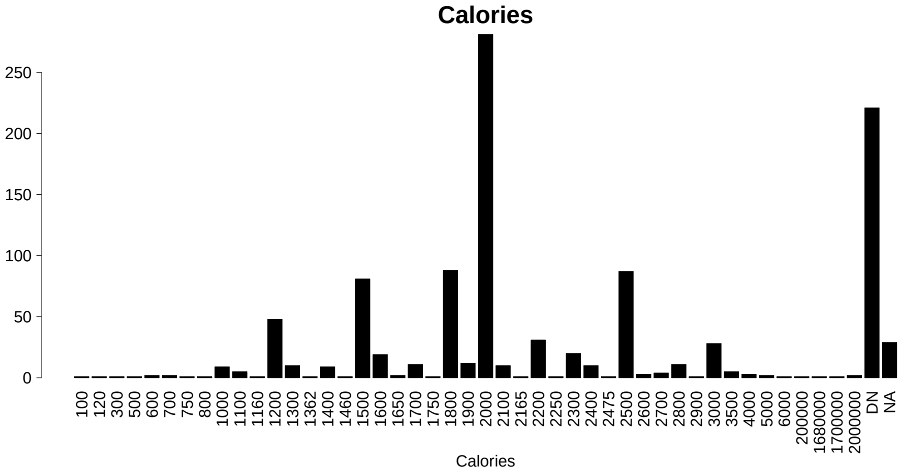


Figure 4.1: Frequency of answers to calories question.

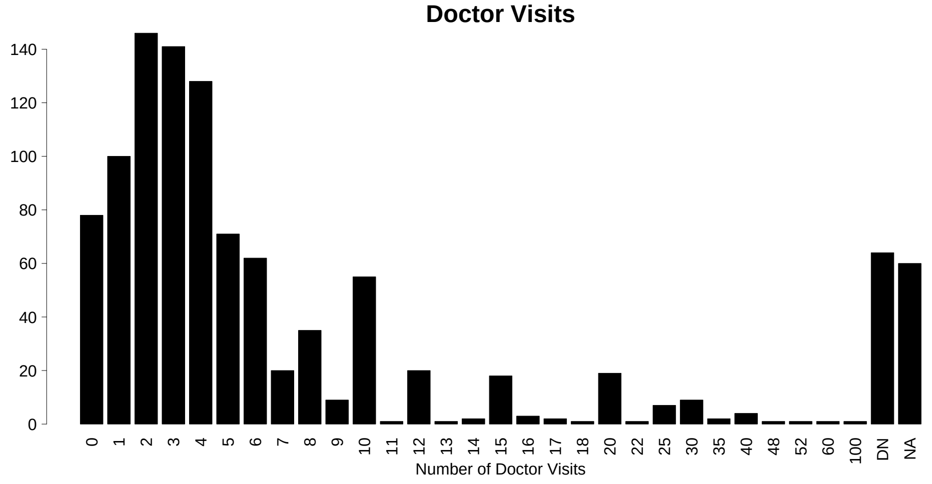


Figure 4.2: Frequency of answers to doctor visits question.

and heaping (yes/no) according to Tables 4.4 and 4.5 are built, and the answers for the calories, UB II, and employment questions are validated. Finally, we assess all questions of the experiment jointly.

4.3.1 General Findings

Before analyzing the incentive experiment, we provide some descriptives of the survey response distributions for the experimental questions. Overall, the expected response patterns can be observed for our survey. Figures 4.1 to 4.9 show the responses to the eight questions of the incentive experiment. For the calories question, we find a strong peak at the right answer (2000 calories, Figure 4.1). The distribution for the doctor visits (Figure 4.2) is not quite as expected. We see heaps at 2, 8, 10, 12, 15 and 20 but were expecting heaps at 52, 24, 12, 6, 4 and 2. In our survey, 52 and 24 are not chosen at all. There is a heap for the life expectancy question (Figure 4.4) at 50%, but

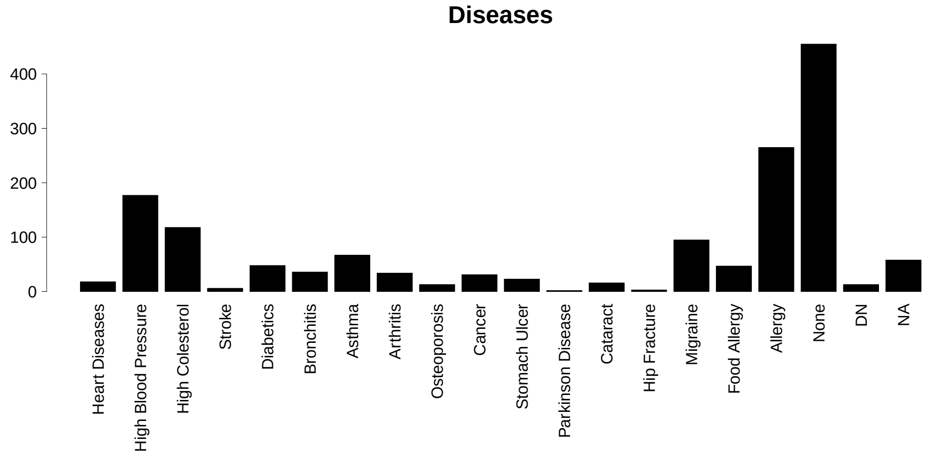


Figure 4.3: Frequency of answers to diseases question.

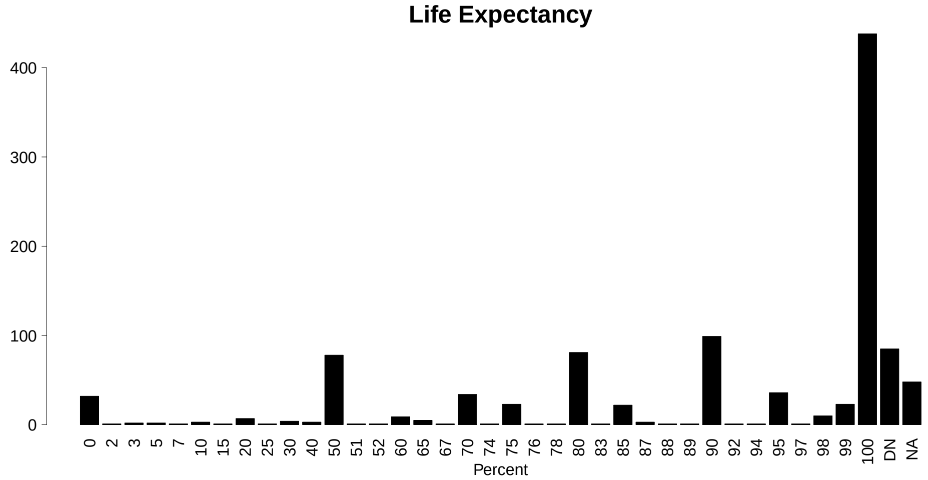


Figure 4.4: Frequency of answers to live expectancy question.

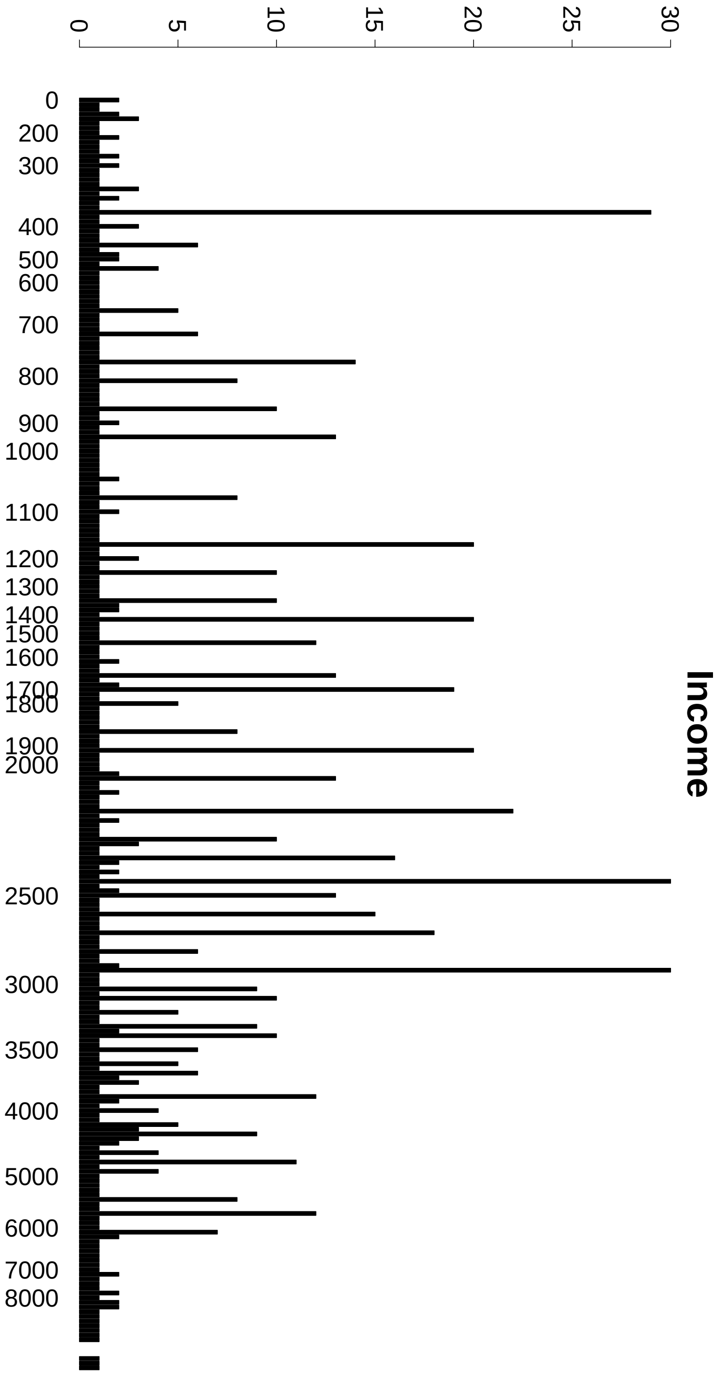


Figure 4.5: Frequency of answers to income question.

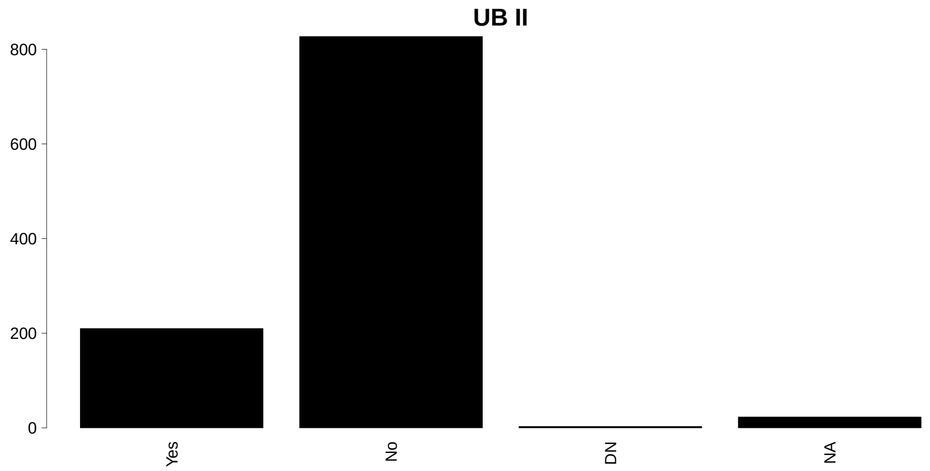


Figure 4.6: Frequency of answers to UB II question.

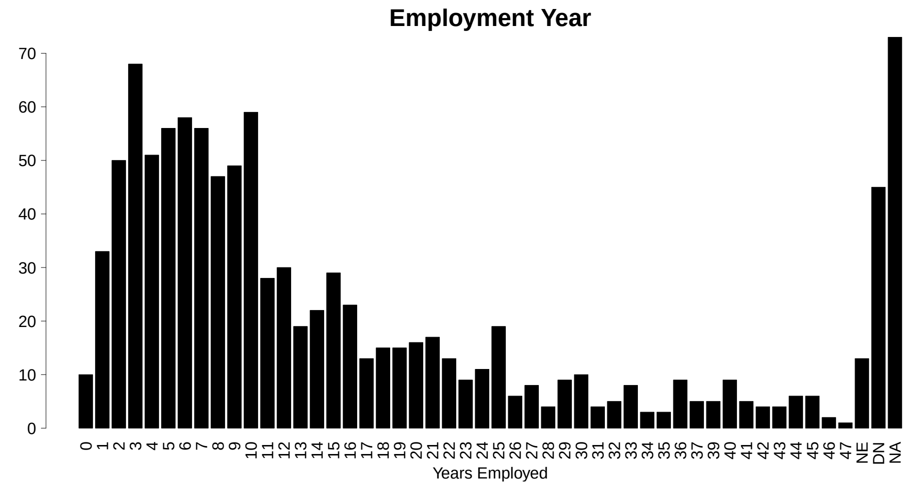


Figure 4.7: Frequency of answers to employment years question.

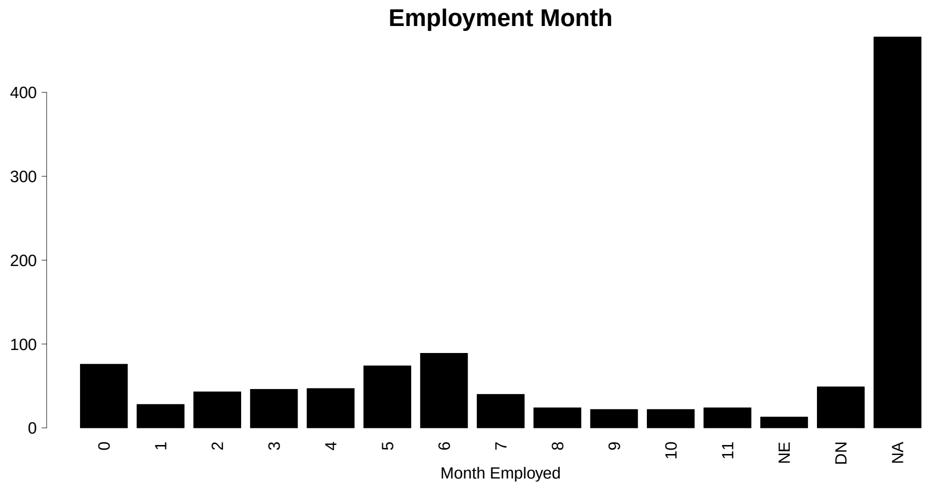


Figure 4.8: Frequency of answers to employment months question.

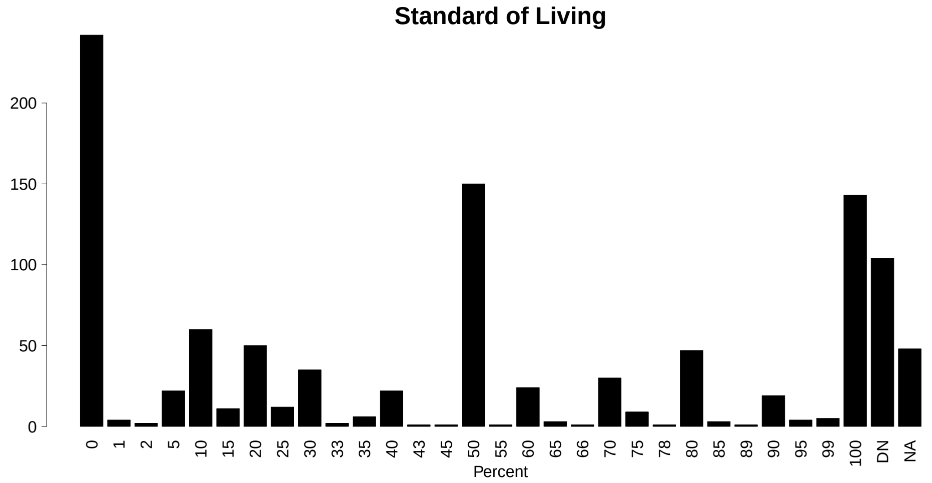


Figure 4.9: Frequency of answers to standard of living question.

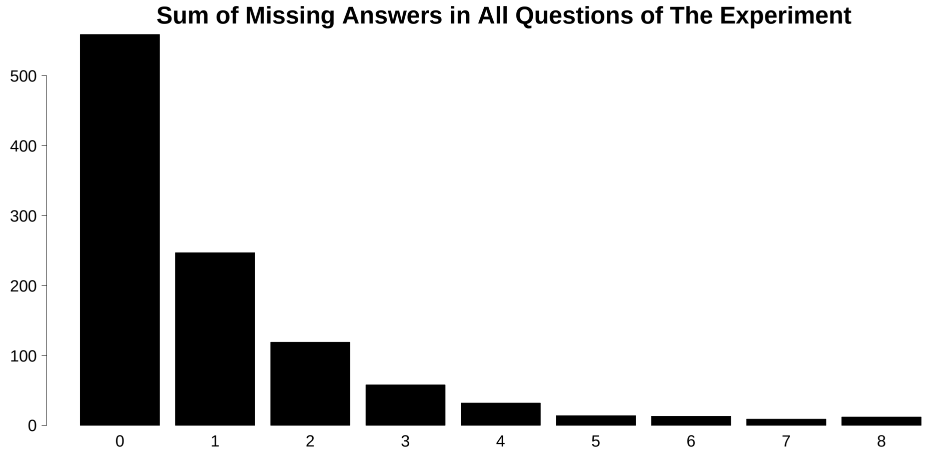


Figure 4.10: Distribution of item nonresponse in all questions.

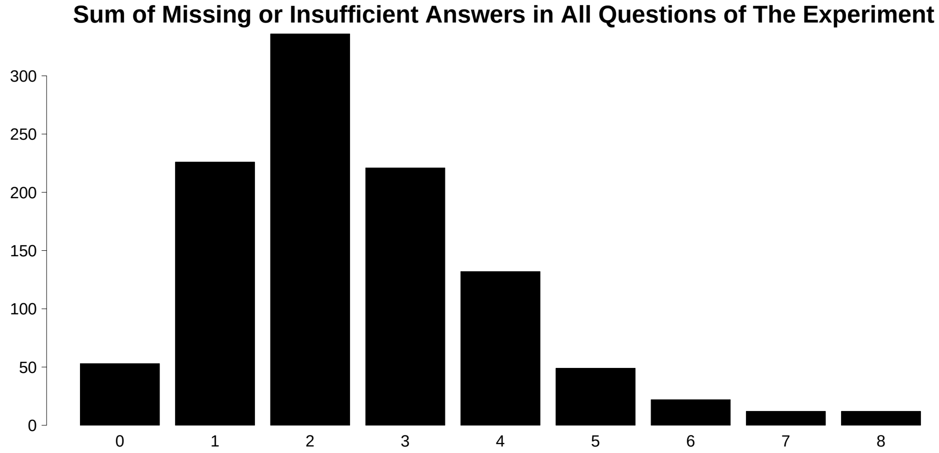


Figure 4.11: Distribution of item nonresponse and insufficient answers in all questions.

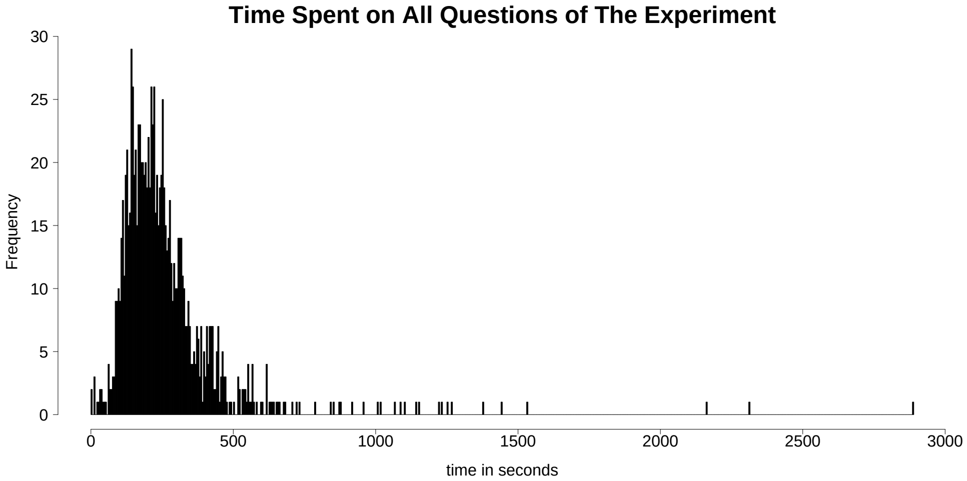


Figure 4.12: Time spent on all questions of the experiment.

there are stronger heaps at 80%, 90% and 100%. We find a much stronger heap at 50% for the living standard question than for the live expectancy question (Figure 4.9), but living standard shows a stronger heap at 0% than at 50%. For income (Figure 4.5), the expected heaps at multiples of €500 are found, but there is also a strong heap at €400, which corresponds to the standard income from a mini job. At the same time, €400 could be a rounded version of income from UB II, which actually was explicitely not asked for. We find heaping to multiples of 5 years for the employment question (Figure 4.7).

From Figure 4.10, we see that item nonresponse in the experimental blocks is quite low: most respondents show no or only one missing value in this part of the questionnaire. Combining missing and rounded/heaped/incorrect answers (Figure 4.11), we find that most people give two of these unsatisfactory answers. The median time spent on the eight questions (Figure 4.12) is 223 seconds but there are strong outliers. Probably these respondents took a break during the interview resulting in a block length of up to about 48 minutes. Because of these strong outliers, we assess at the median instead of means times spent on each question.

The comparison of the questions, irrespective of incentive groups, shows that item nonresponse and time to answer varies considerably across the questions (see Table 4.6).

Item nonresponse is found to highest for the variable *employment month*, while the amount of item nonresponse for *employment year* is much lower. These two questions were on the same page, and it was possible to go on in the questionnaire without receiving an error message if only the year of employment was filled in. The answer to *employment month* was then automatically set to "no answer". Because of the immense number of missing information in *employment month*, we exclude it from further analyses. As known from the literature, there is a lot of missing information for *income*,

question	% item nr	% don't know	% no answer	time to answer
knowledge				
calories	24	21	3	14
recall				
doctor visits	12	6	6	19
employment year	11	4	7	49
employment month	49	5	44	49
diseases	7	1	6	25
sensitive				
income	22	4	18	28
UB II	3	1	2	9
expectation				
standard of living	14	10	5	27
life expectancy	13	8	5	20

Table 4.6: Proportions of item nonresponse, "don't knows" and "no answers" and median time to answer for the variables of the incentive experiment.

mostly due to "no answer". In general, we find more "no answer" than "don't know" responses for sensitive questions, i.e., income, UB II, doctor visits, and more "don't know" responses for knowledge and expectation questions. For our analyses, it is not important why the information is missing, and we do not differentiate between "don't know" and "no answer". Instead, we only look at item nonresponse as a whole.

For all respondents who gave substantive answers, i.e., excluding "don't know" and "no answer", we find the following for our indicators of unsatisfactory answers:

- 65% of the respondents give a wrong answer to the calories question.
- 38% of the respondents round to prominent values in the doctor visits question.
- 66% of the respondents do not report at least one of the last three diseases.
- 8% of the respondents choose "50%" in the life expectancy question.
- 24% of the respondents round their income to multiples of €500, 69% to multiples of €100.
- 6% of the consenting respondents misreport their UBII status.
- 22% of the respondents round the years of employment to multiples of 5.
- 16% of the respondents choose "50%" in the living standard question.

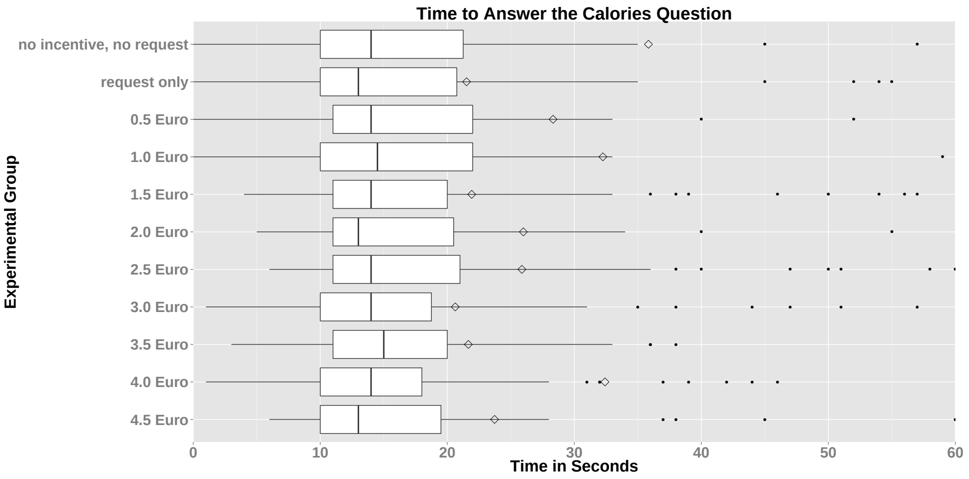


Figure 4.13: Time spent on the calorie question.

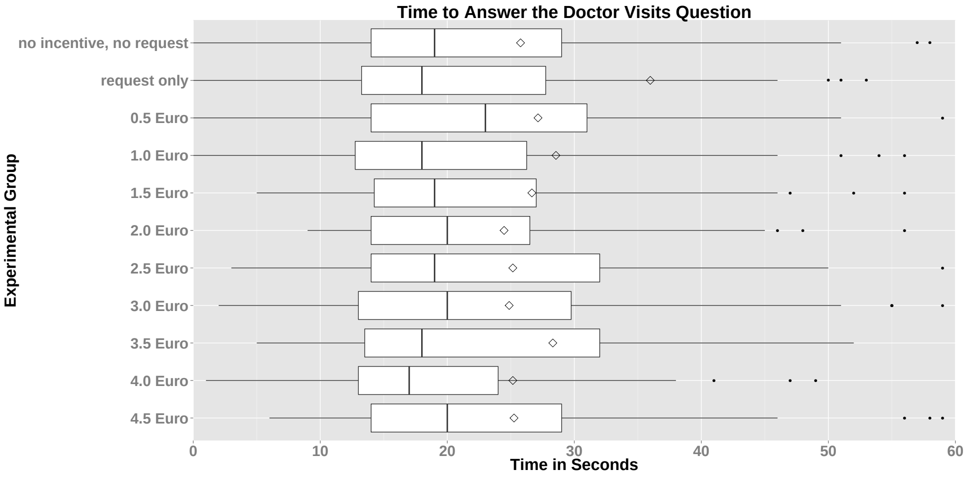


Figure 4.14: Time spent on the doctor visits question.

4.3.2 Bivariate Findings

Next we compare time spent on each question, the amount of item nonresponse, and unsatisfactory answers across the 11 incentive groups. In our experimental setting, we randomly assigned respondents to the treatments groups. Thus, comparing group means will give us a first indication on the effect of incentives on measurement error. In a next step, we use multivariate models to analyze the incentive experiment. By doing this, we can include both indicators of the respondents' behavior prior to the experiment and interactions with the incentive treatment. Finally, we will test different functional forms of the effect of additional incentives on unsatisfactory responses.

Time Spent on a Question

Figures 4.13 to 4.20 show the boxplots of response times for each of the eight questions of the incentive experiment, depending on the incentive treatment. Distributions of

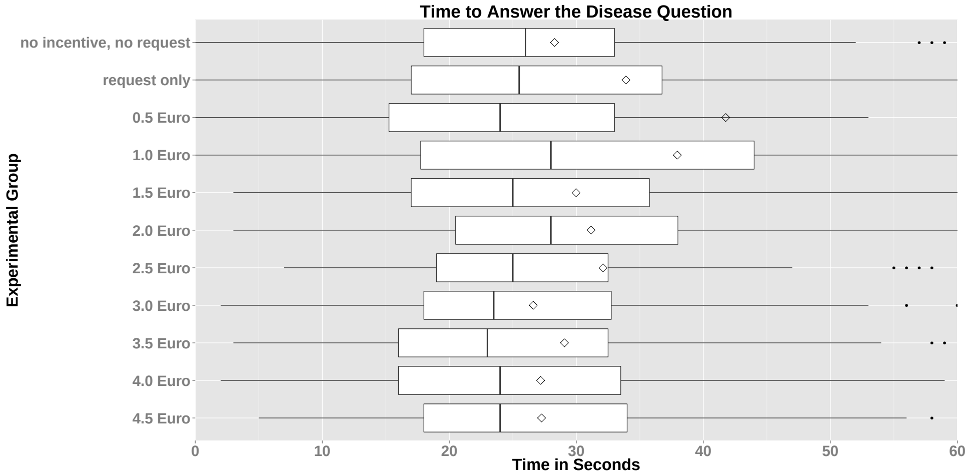


Figure 4.15: Time spent on the disease question.

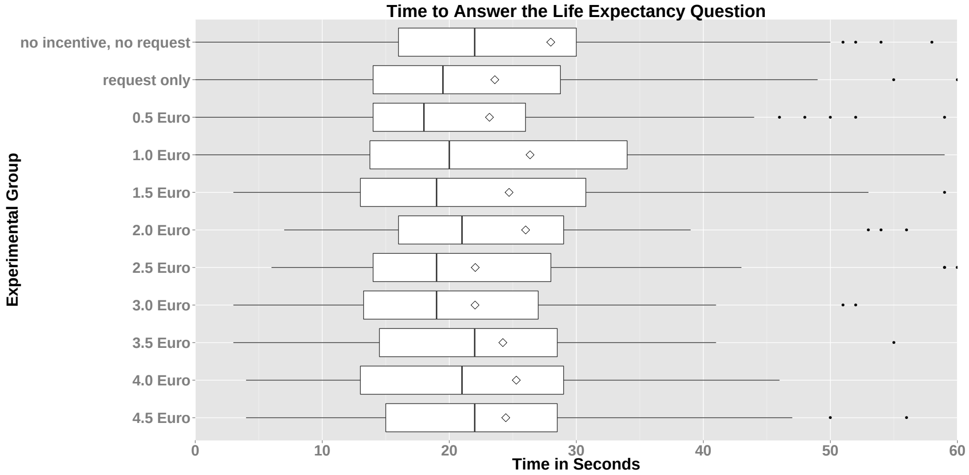


Figure 4.16: Time spent on the life expectancy question.

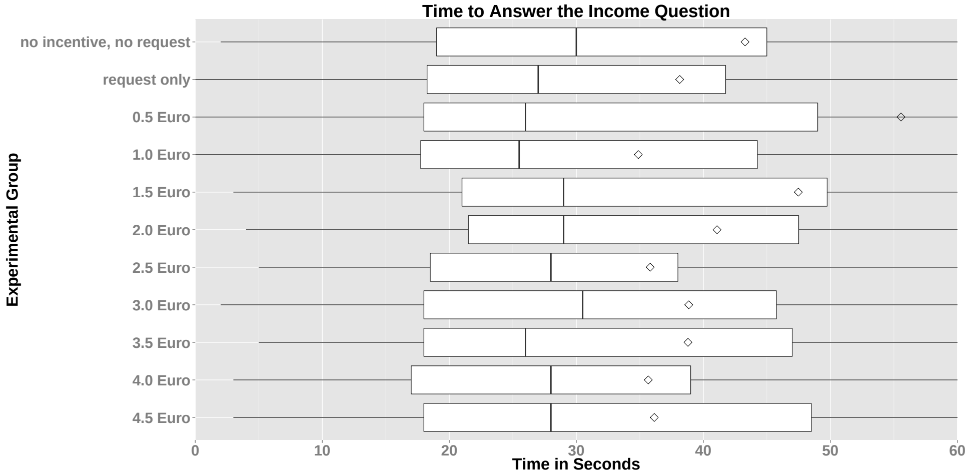


Figure 4.17: Time spent on the income question.

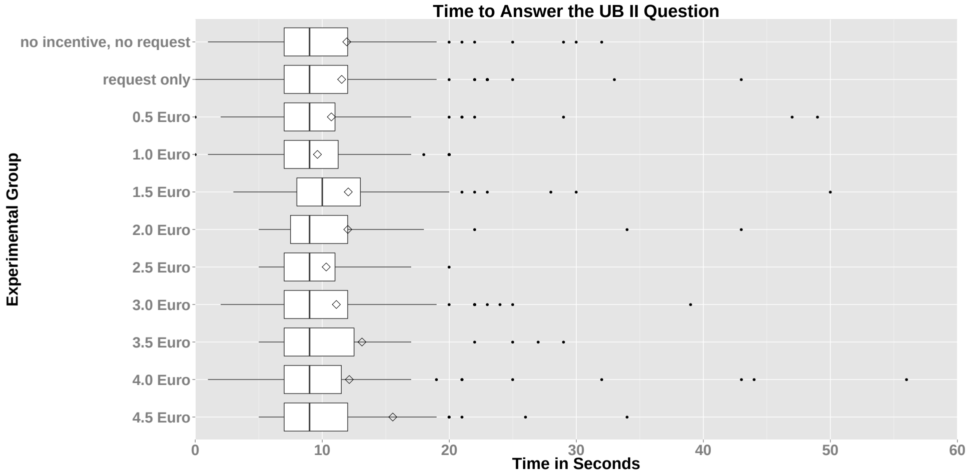


Figure 4.18: Time spent on the UB II question.

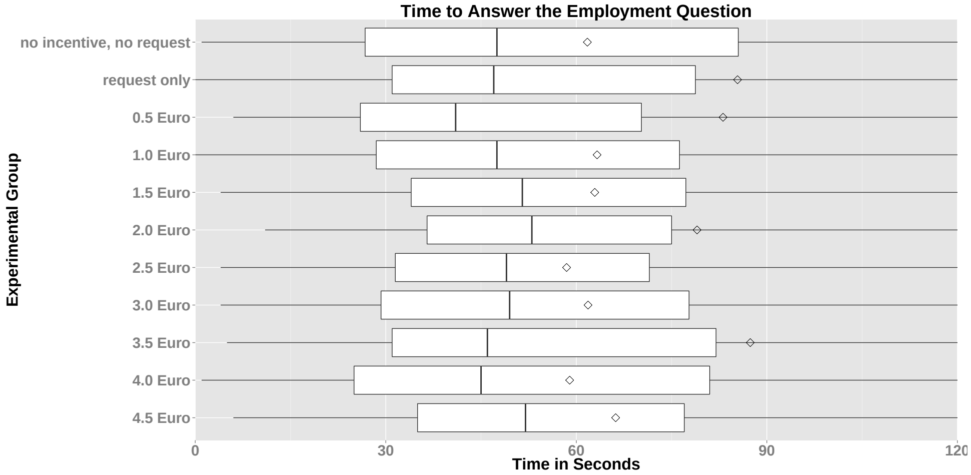


Figure 4.19: Time spent on the employment question.

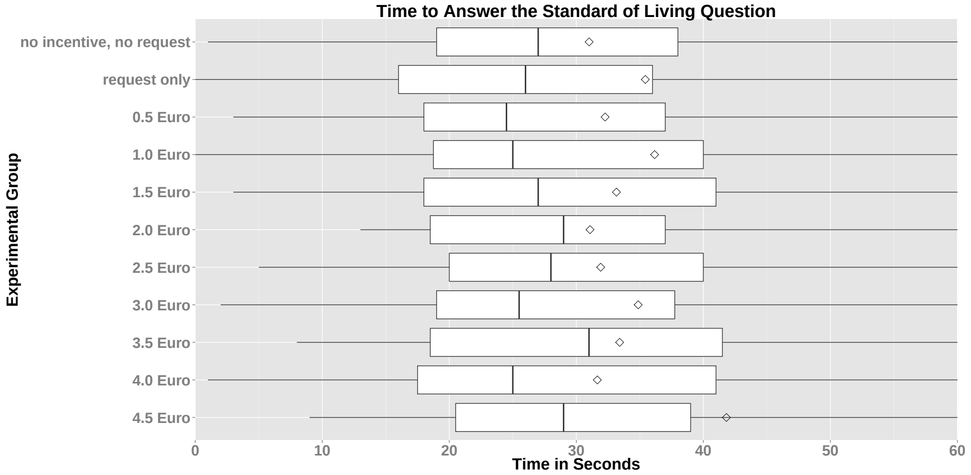


Figure 4.20: Time spent on the standard of living question.

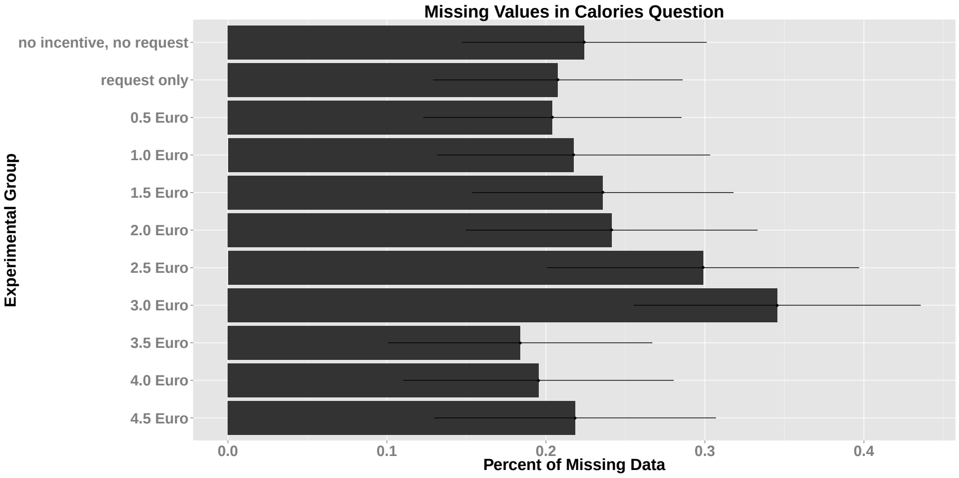


Figure 4.21: Item nonresponse in the calories question.

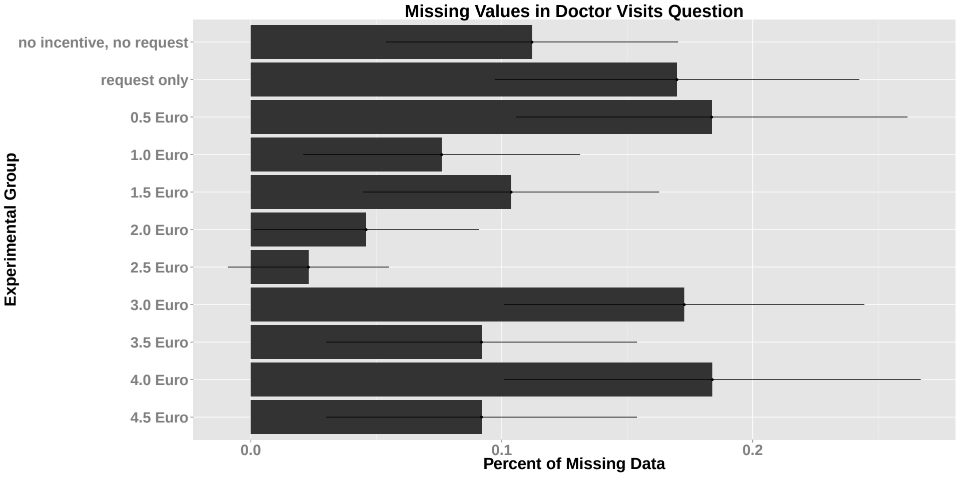


Figure 4.22: Item nonresponse in the doctor visits question.

response times are skewed for all questions. Many of them show strong outliers. For all questions, we perform F-tests to assess whether the mean response times differ between the experimental groups. Because normality is doubtful for the distributions of the response times, we perform the same analysis nonparametrically, i.e., by means of Kruskal-Wallis tests. Both tests do not indicate significant differences between the incentive groups for any of the questions[1]. Note that the Figures do not show the strong outliers but only zoom in to the range of 0 to 80/120 seconds.

Item Nonresponse

Figures 4.21 to 4.28 show the amount of item nonresponse for each of the questions depending on the incentive treatment. Even though there is some variation in the amount of item nonresponse, we do not find a general pattern. However, a test of

[1] Tables showing p-values for the tests of group differences for response times, item nonresponse and insufficient answers can be found in Appendix B.

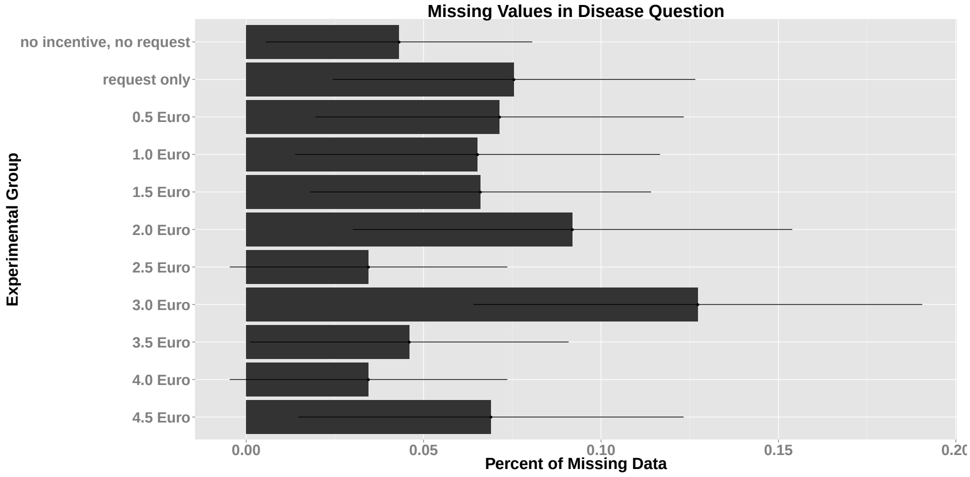


Figure 4.23: Item nonresponse in the disease question.

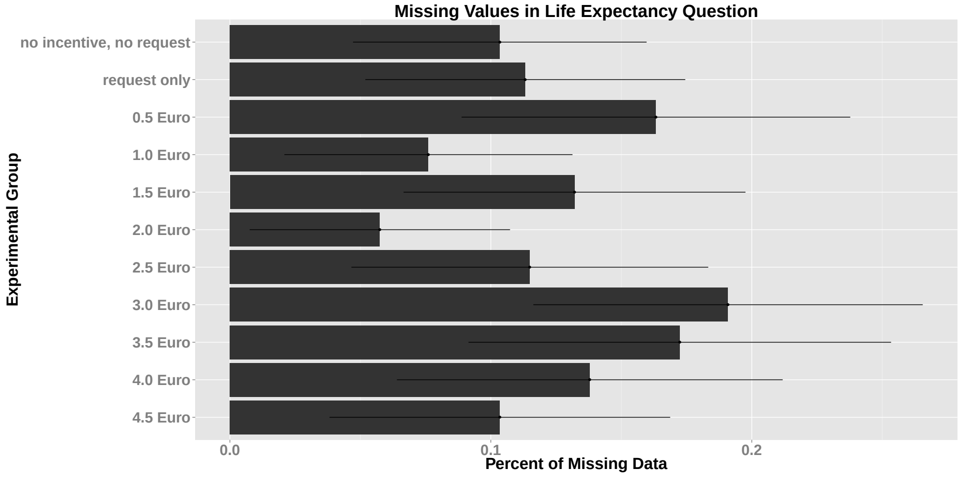


Figure 4.24: Item nonresponse in the life expectancy question.

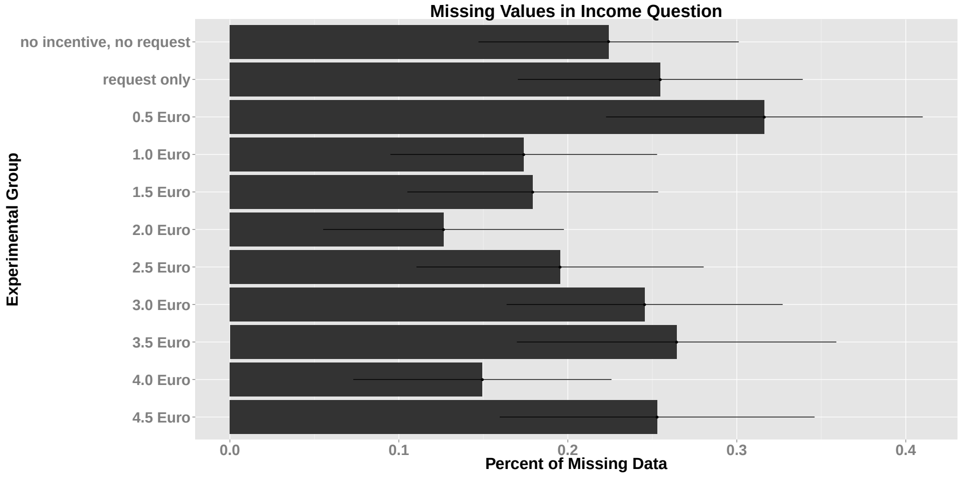


Figure 4.25: Item nonresponse in the income question.

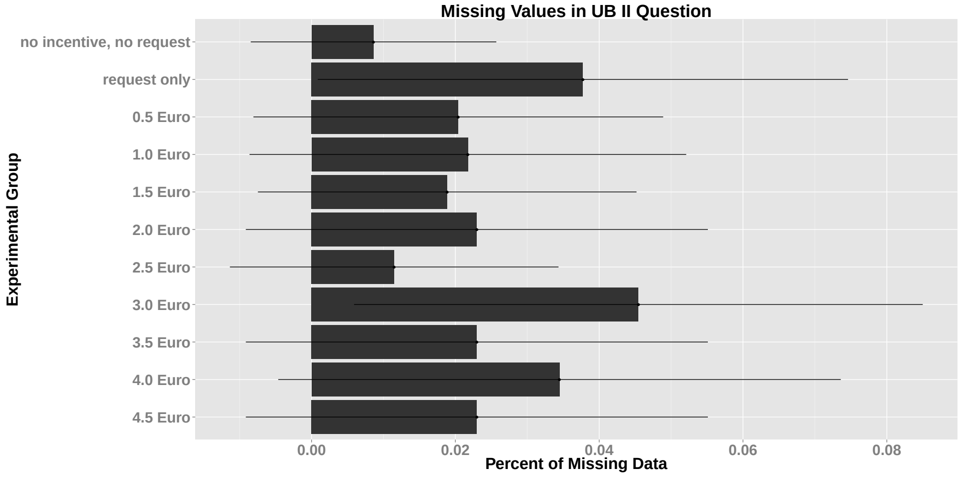


Figure 4.26: Item nonresponse in the UB II question.

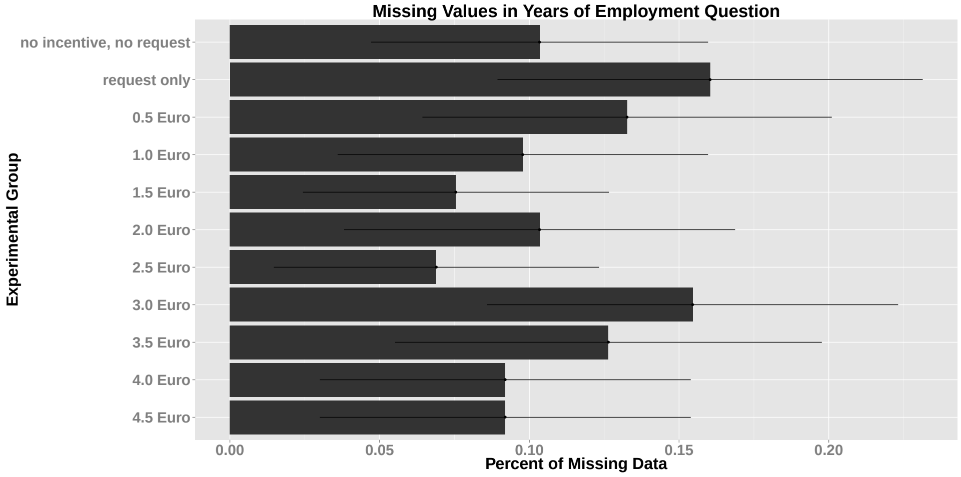


Figure 4.27: Item nonresponse in the employment question.

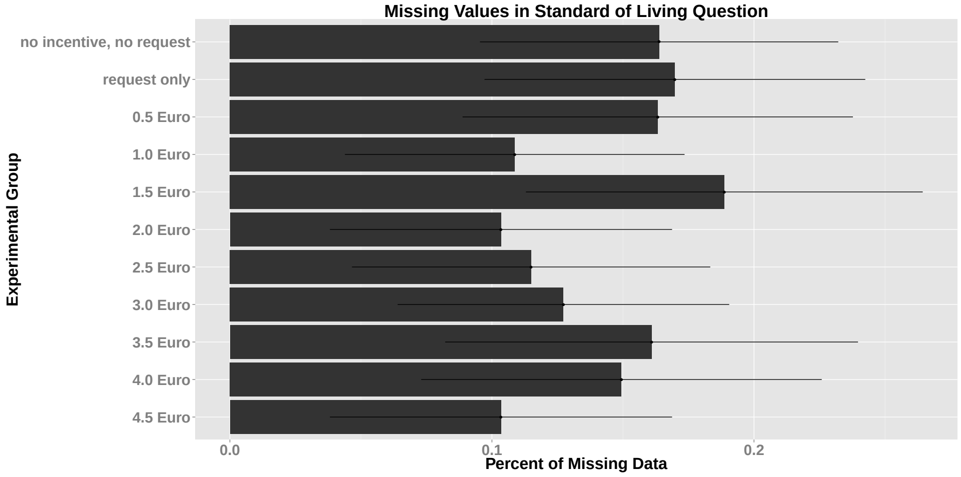


Figure 4.28: Item nonresponse in the standard of living question.

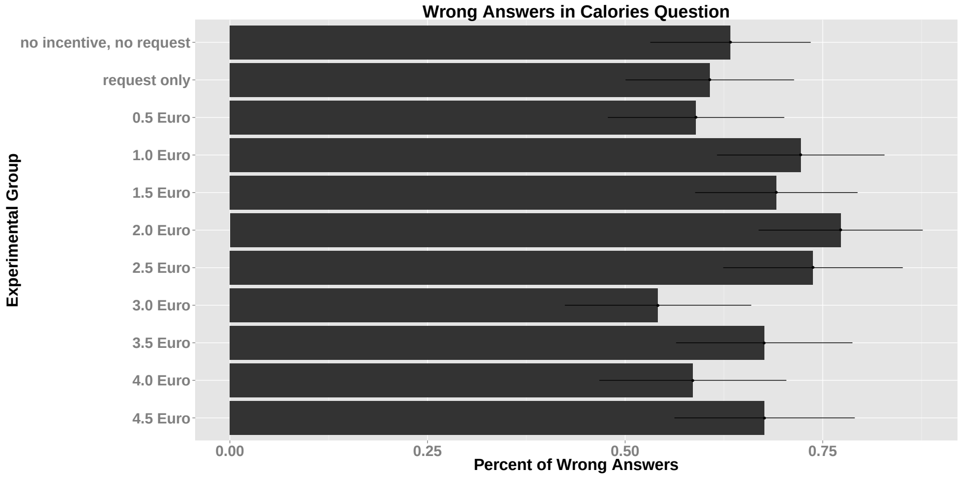


Figure 4.29: Wrong answers in the calories question.

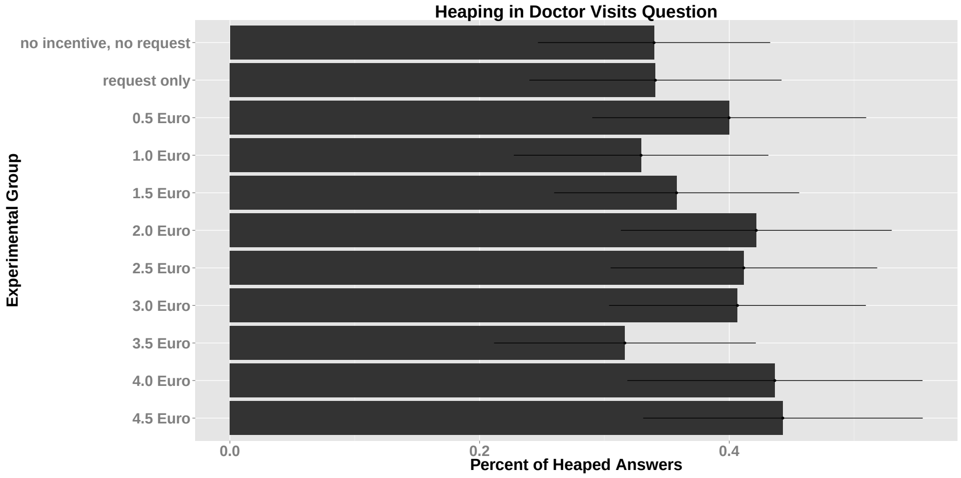


Figure 4.30: Heaping in the doctor visits question.

equal proportions between all groups finds significant differences between the incentive groups for *doctor visits* and *income* (10%-level). Multivariate analysis of the effects of the experimental treatment on the amount of item nonresponse for all questions will be studied further in Section 4.3.3.

Unsatisfactory Answers

Figures 4.29 to 4.34 show the amount of rounding and heaping and incorrect answers depending on the incentive treatment. We do not find differences between the incentive treatments for most of the questions, but we do find differences for the amount of wrong answers in the calories question and for heaping in the life expectancy question on the 10%-level. Multivariate analysis will be performed in Section 4.3.3.

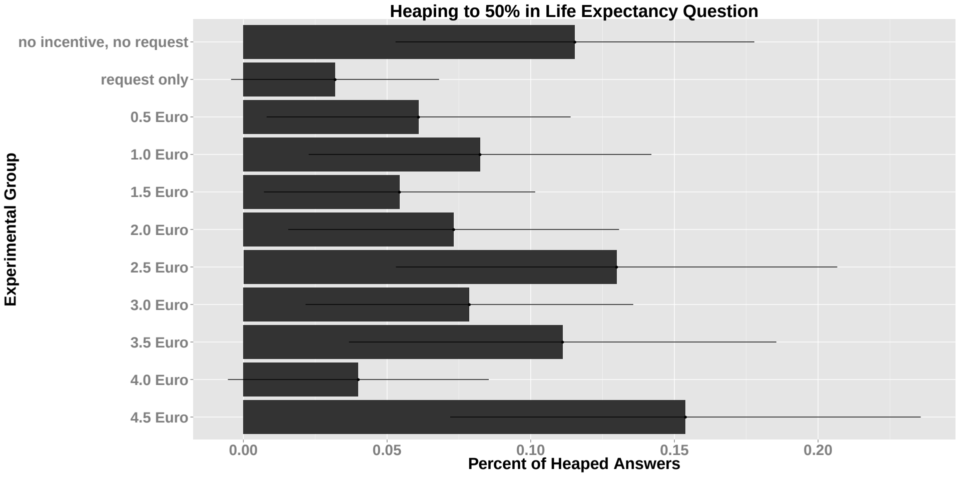


Figure 4.31: Heaping in the life expectancy question.

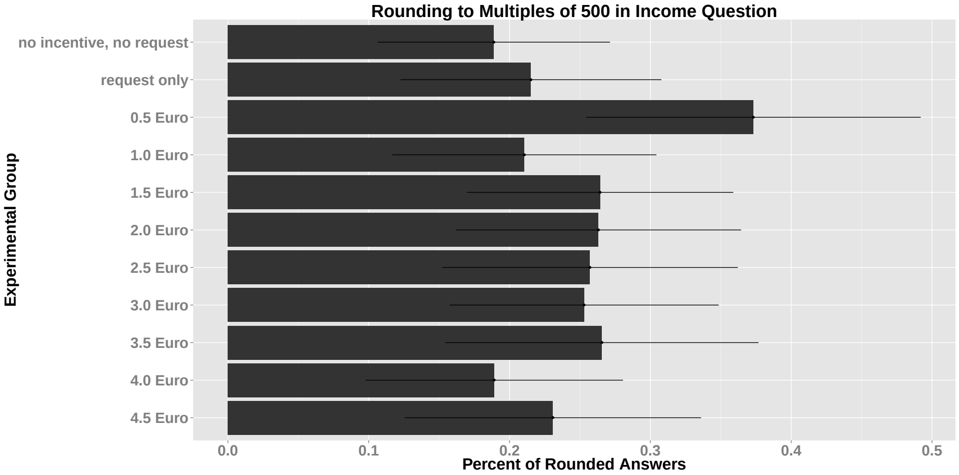


Figure 4.32: Rounding to multiples of 500 in the income question.

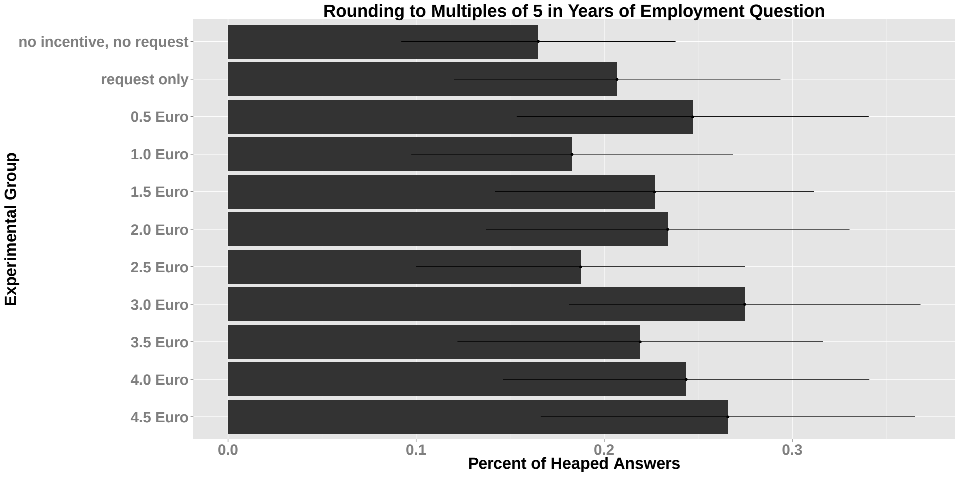


Figure 4.33: Heaping in the employment question.

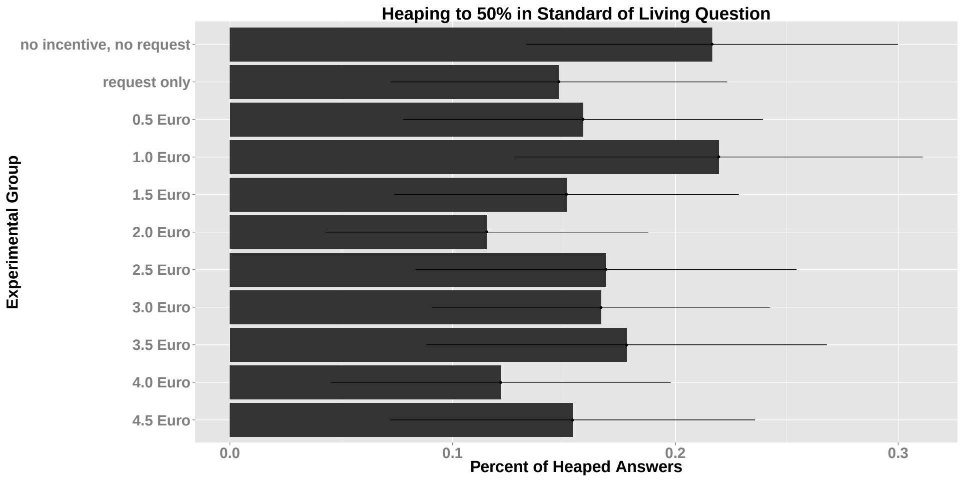


Figure 4.34: Heaping in the standard of living question.

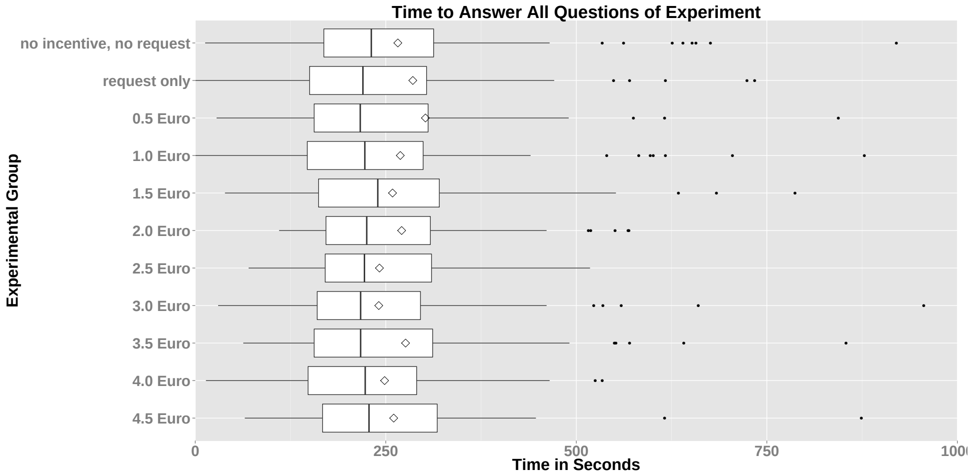


Figure 4.35: Time spent on all questions of the experiment

Time Spent on All Variables and Missing or Heaped and Incorrect Answers on all Variables

Figures 4.35, 4.36 and 4.37 show the amount of time spent on all questions of the experiment and the number of missing and unsatisfactory answers in all eight questions of the experiment. As expected from the previous Sections we do not find any significant differences between the experimental groups.

Validation of Unemployment Benefit II, Income and Employment History

In this Section, we validate the responses to the UB II, income, and employment history questions using the administrative data of the FEA. By doing this, we are able to directly assess the measurement error in income and the years of the longest employment phase as well as misreporting in the UB II question. Although the questions were designed to match the administrative data information, there are some chal-

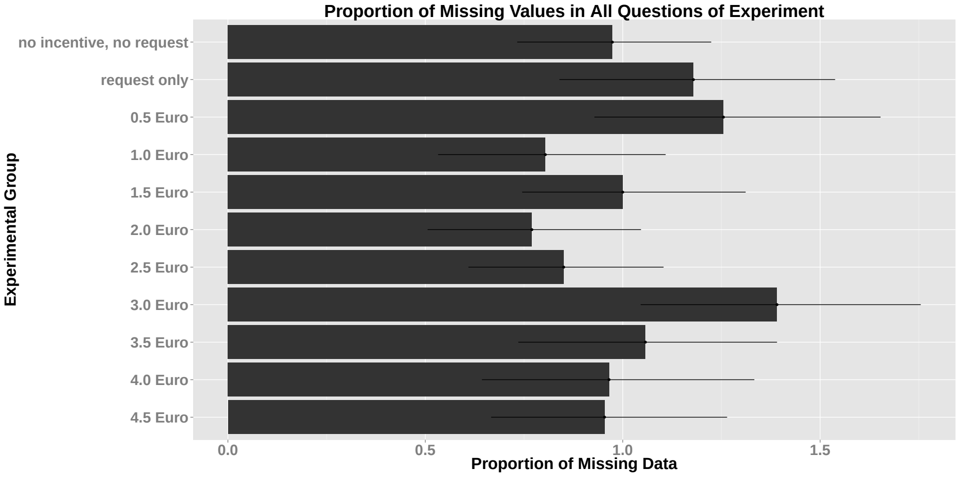


Figure 4.36: Proportion of missing items in all questions of the experiment

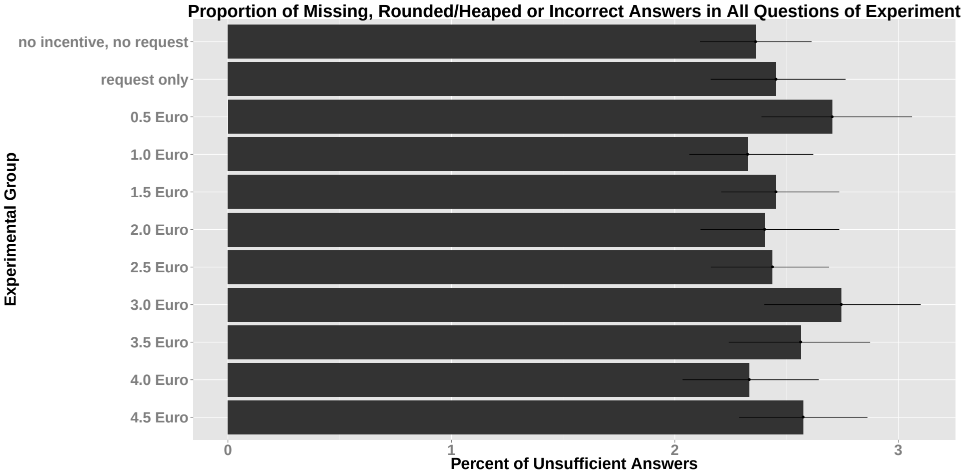


Figure 4.37: Proportion of unsatisfactory answers in all questions of the experiment

lenges that are described in the following paragraphs.

For the validation of UB II, income, and employment history, we can only use the 662 respondents who consented to linking their survey response to the administrative data of the employment agency. As the sample cases were drawn from these administrative data, information on all respondents is easily available [2]. The administrative data do neither contain information on self-employment nor information on civil servants (see Chapter 2.1). We find four respondents whose reported age differs from the age information in the register data. As we can not be sure that we have interviewed the right person, we drop these cases, leaving 658 cases for the analyses. The placement of the request for consent to data linkage has been experimentally varied: 75% of the respondents received the request at the beginning of the interview, 25% at the end of the interview, i.e., after the incentive intervention. The placement of the consent question might affect the responses to the experimental questions: respondents receiving the linkage consent question prior to the experiment might be aware of the fact that we know the true answers to the questions, and therefore behave differently than respondents who are not informed about the administrative data yet. All multivariate analyses of the validation questions control for the consent placement.

To find possible sources of discrepancy of survey answers and administrative records, we first compare the reported employment status with the administrative status (Table 4.7). According to administrative data, there are only few respondents who have more than one job at the same time (4.4%, 29 respondents). All of these have a mini job and an additional job, i.e., one respondent is doing a vocational training, 13 respondents have a part-time job, and 15 have a full-time employment. For these respondents, we define the job with the higher income as their main job to be compared with the survey response.

The match between the reported and administrative employment status is quite good. As expected, most of the people who report being self-employed do not have a job according to the administrative data. Only four of them have a part-time or a mini job, and it is likely that those respondents define self-employment as their main job. As we do not have information on their income from self-employment, we drop all respondents who report to be self-employed from the analyses for income and longest employment phase, leaving 621 cases.

The analyses for *unemployment benefit* are based on all consenting respondents. Only 11 consenting respondents did not respond to the UB II question, and are therefore not part of the analyses. Respondents were asked for UB II on household level, which can differ from the benefit community level as it is used in the administrative data

[2] The data sets used for these analyses are IAB Beschäftigtenhistorik (BeH), Version 08.07, Nürnberg 2012 and IAB Leistungshistorik Grundsicherungsempfänger SGB II (LHG), Version 06.04, Nürnberg 2011 as well as IAB Leistungshistorik Grundsicherungsempfänger in zugelassenen kommunalen Trägern (XLHG), Version 01.10, Nürnberg 2011.

	full time	part time	mini job	training	not employed
no answer	4	2	0	0	7
don't know	0	0	1	0	1
self-employed	0	2	2	0	33
full time	251	10	0	1	22
part time	14	75	4	0	8
mini job	1	2	31	0	7
vocational training	0	0	0	13	4
maternity leave	2	3	1	0	20
unemployed	1	0	5	0	62
not working	2	0	9	0	58

Table 4.7: Match of reported and administrative employment status

(see Section 2.1). In principle, respondents can live in a household that consists of more than one benefit community, and thus asking for UB II on the household level is not equal to asking for UB II on the community level. Some "misreport" might be due to the difference between household and benefit community. As the respondents were randomly assigned to the experimental groups, respondents who are wrongly found to have misreported their UB II status should equally likely be found in the treatment groups as well. Comparisons between groups should not be affected by these "false-positives", but the discrepancy between administrative records and the survey question can add some additional noise. We do not find any difference in UB II recipiency between the treatment groups according to the administrative data.

The 60 respondents who did not report the number of years of their *longest employment phase* are dropped, leaving 561 cases for the analyses. When asked for the longest employment phase, respondents were told not to count a job change as a break, but were not given further instructions how such a break is defined. Therefore, we use three different ways of defining a break in the employment phase:

1. We always count it as a break when a respondent did not have a job for at least one day.
2. We only count it as a break if the next job starts later than the first day of the following month after one job ended.
3. We only count it as a break if the next job starts later than the next month.

Respondents were asked to report their longest employment phase as years and months of this phase. For validation, we construct the longest employment phase according to administrative data in the same way. Due to a high amount of missing information in the *month* variable, we only validate the years. The three definitions lead to the same longest employment phase for 81.5% of the respondents (Figure

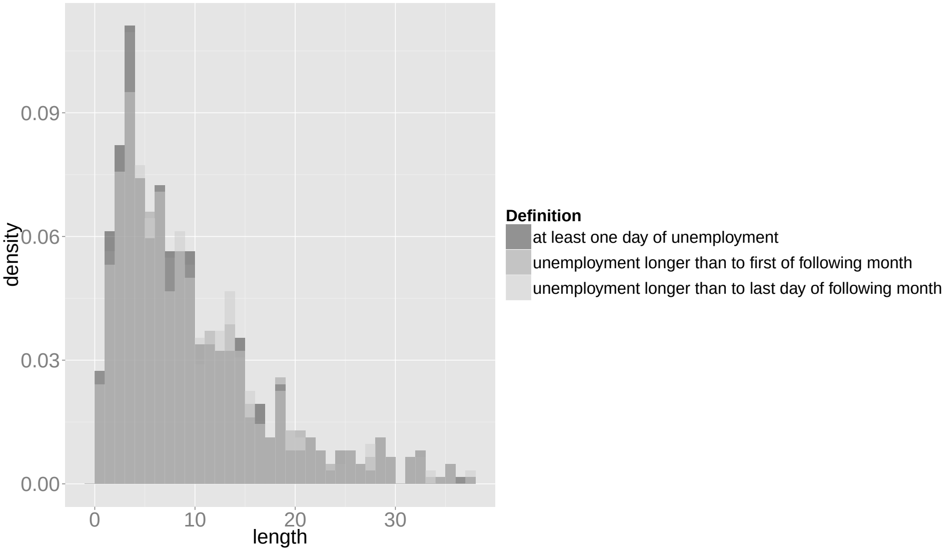


Figure 4.38: Histogram of longest employment phase in register data according to the three definitions.

4.38), the correlation of the reported and administrative longest employment phase is about 0.8 for all definitions. Although all respondents held a social security contributing job (sozialversicherungspflichtige Beschäftigung) within the last 10 years, the administrative data do not necessarily cover their whole employment history as phases of self-employment between jobs are not captured. As a consequence, times of self-employment only appear to be a break in the employment phase in the administrative data. Moreover, register data is only available from 1975 in West Germany and 1993 in East Germany. This implies that the longest employment phase from the administrative data is only the lower bound for the true longest employment phase. Reported durations that are longer than those computed from the administrative data might but need not be correct. However, the comparison between the groups is still valid as we do not find differences in the administrative data longest employment phase between the incentive groups for any of the three definitions.

For the validation of the *personal income*, we drop 108 respondents who did not report their income. The register income of 44 respondents equals the monthly social security contribution assessment ceiling (Beitragsbemessungsgrenze) of the year 2012, or, in case of unemployment, equals the income ceiling of the year of their last employment (see Section 2.1). We drop these respondents from the measurement error analysis because it is clear that their reported income does not equal the register income due to censoring in the register data. This leaves 467 respondents, 329 who are currently employed and 138 who are currently unemployed. The experimental groups do not differ significantly with respect to their administrative data income (overall and separated by employment status).

The income question was asked to match the administrative information, i.e., respondents were asked for their last income from work before taxes. In the case of more than one job, the income was summed up. Respondents who were currently unemployed were asked to report the last monthly income from their last job. As it can be assumed to be far more complex to recall an income one received years ago, we separate our analyses according to the employment status as reported by the administrative data.

In the univariate analyses, we have seen a strong peak at €400 that we do not expect from theory. Taking a closer look at these respondents, we can see that they basically belong to two groups: respondents who hold a mini job, and respondents who are currently unemployed according to the administrative data (Table 4.8).

register income in €	reported income in €	register employment prev. month
188.17	400	mini job
405.48	400	mini job
270.86	400	mini job
406.41	400	mini job
403.62	400	mini job
357.43	400	mini job
341.33	400	mini job
396.18	400	mini job
372.62	400	mini job
206	400	not employed
78	400	not employed
834	400	not employed
44.7	400	not employed
379.68	400	not employed
407.22	400	not employed

Table 4.8: Administrative and register income for people reporting an income of exactly €400.

In 2012, the legal income limit for a mini job was €400. None of the respondents reporting exactly €400 actually received exactly €400 in the particular month. Rather, they seem to have heaped their real income to the legal limits. Especially for the respondents earning less than €300, it is obvious that they do not follow a simple rounding pattern like rounding to the next 100, but to the income they usually earn but did not earn in the previous month. We note that, as most of these interviews were taken in February, the income question refers to January, and it is likely that some of the respondents were on holiday at the beginning of the year, and therefor received a lower income than usual. In total, 17.65% of the respondents having a mini job round to exactly €400. The standard amount of unemployment benefit is €382. The unemployed respondents might round to €400 because they had a mini job before, or they might receive UB II while being unemployed, and wrongly report a rounded value as their income. Note again that the unemployed respondents are excluded from the income measurement error analyses.

Table 4.9 shows how many reported income values correspond to a true administrative income that is rounded to the next 5, 10, 100, 500 or 1000. Employed and non-employed respondents exhibit about the same amount of values that are rounded to some extent. Rounding to the next 50 or 100 is the most common rounding scheme for the employed, and rounding to the next 500 for the non-employed. In general, non-employed respondents tend to round in broader intervals which might be due to larger uncertainty (Ruud et al., 2014). However, most of the reported income values do not correspond to a rounded version of the administrative income.

In the following analyses, we define measurement error as the difference between the log reported and the log administrative income as it is common in the literature (Bound and Krueger, 1991; Kim and Tamborini, 2012)). Figures 4.39 and 4.40 show the empirical distributions of the reported and administrative incomes for employed and non-employed respondents. We assume that most respondents had their income information available when they filled in the online questionnaire, and could in principle have looked up the correct numbers. Since we specifically asked to be as precise as possible in the experimental request, we expect to see less measurement error for the groups who received the request. We also expect measurement error to decrease with the amount of the additional incentive.

reporting pattern	currently employed (in %)	currently not employed (in %)
rounding to 5	0.72	0.6
rounding to 10	0.72	0.91
rounding to 50	6.52	4.83
rounding to 100	6.52	6.34
rounding to 500	2.9	6.95
rounding to 1000	2.9	5.44
no rounded form	72.51	76.09

Table 4.9: Extend of income rounding for currently employed and non-employed respondents.

Table 4.10 shows true and reported UB II recipiency during the last 12 months. Overall, only about 7% of the respondents misreport UB II recipiency. Surprisingly, we find more than 8 times more overreporting than underreporting. As receiving unemployment benefit is socially undesirable, we usually expect respondents to underreport recipiency. Social desirability might not have been a problem in this survey because the respondents knew that the survey was conducted by the employment agency. Moreover, these respondents consented to data linkage and might show less misreporting than the non-consenters, as they can expect that their answer can be

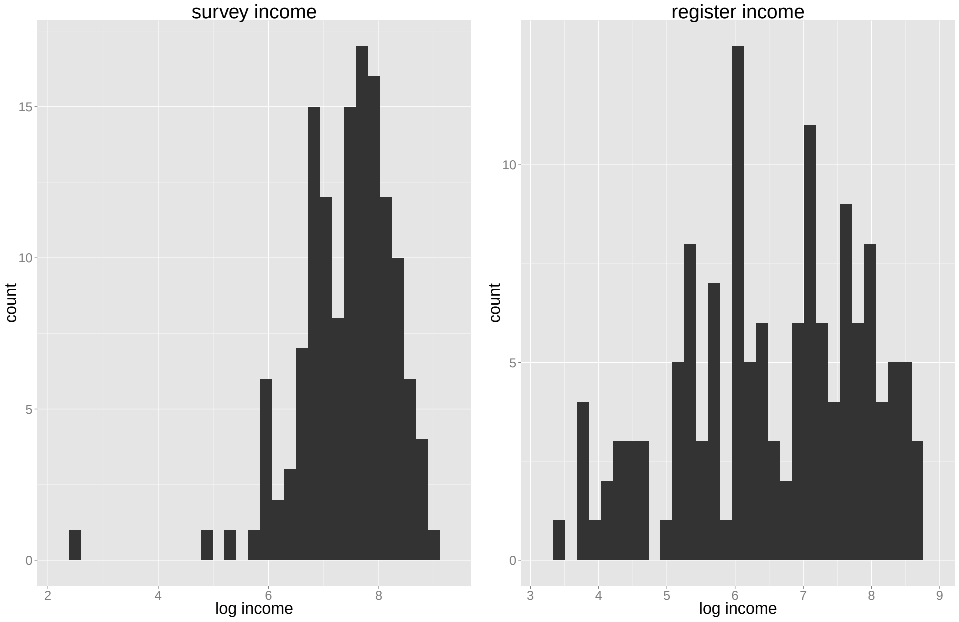


Figure 4.39: Histogram of last survey and register income for non-employed respondents.

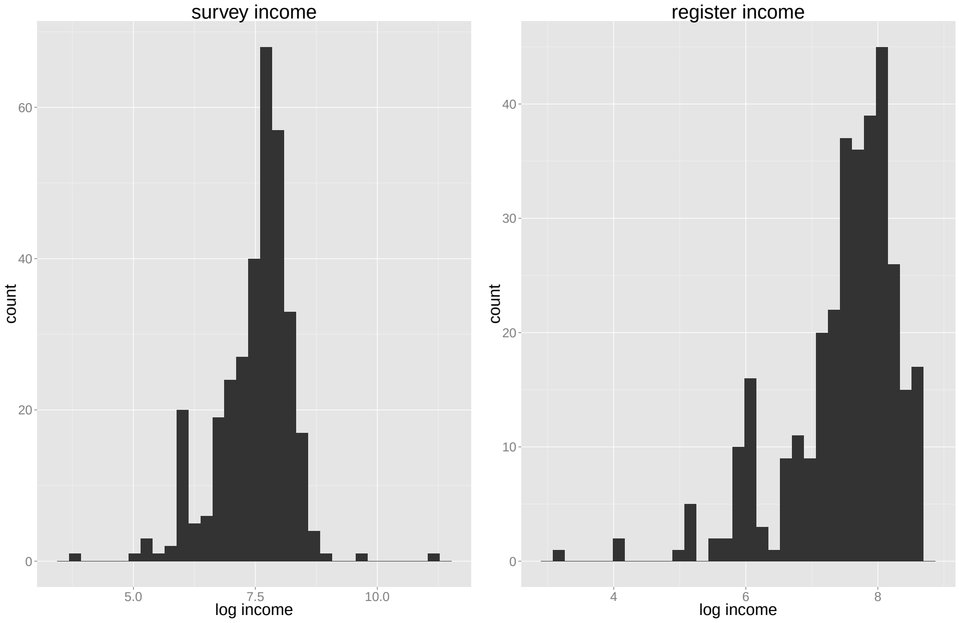


Figure 4.40: Histogram of last month's survey and register income for employed respondents.

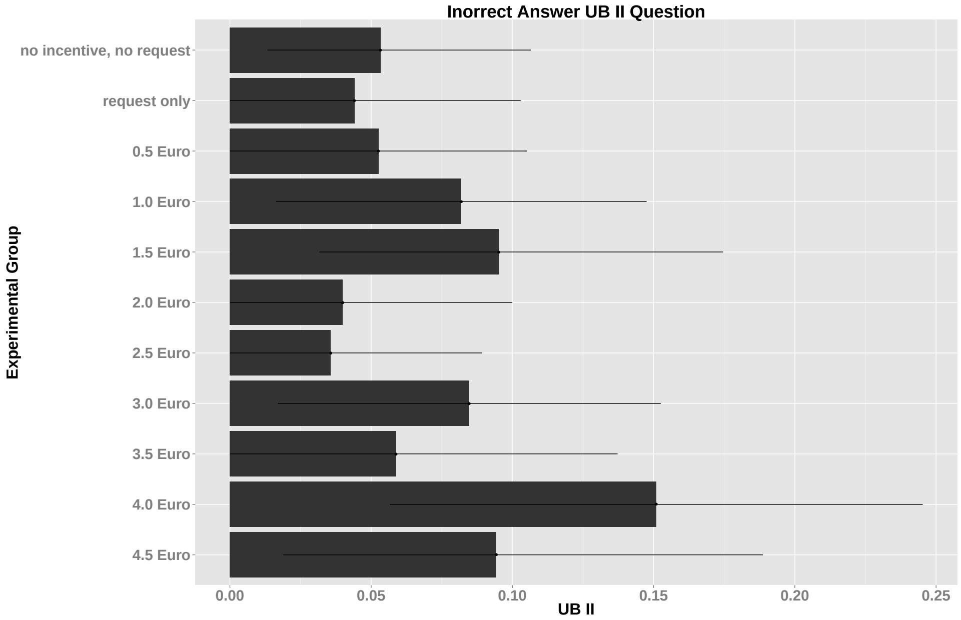


Figure 4.41: Wrong answer in UB II question.

validated. On the other hand, the amount of overreporting indicates that this recall question is hard to answer for some respondents: 8 respondents who overreport UB II recipiency in the last 12 months received UB II earlier in 2011, whereas 26 received UB II at some point. This can be interpreted as a recall problem rather than intentional misreporting. Being reassured that misreport is mostly due to question complexity, we expect that thinking hard about the answer can help to improve accuracy. Thus we expect to find an effect of the additional incentive. Since there are only 5 respondents underreporting UB II recipiency, we do not distinguish between underreporting and overreporting, and only analyze both errors combined as misreport.

	administrative: no UB II	administrative: UB II
survey: no UB II	513	5
survey: UB II	41	87

Table 4.10: True and reported UB II recipiency.

Figure 4.41 shows the proportion of respondents giving a correct answer to the UB II question across all experimental groups. There are neither significant differences between the groups, nor can clear patterns be observed.

Figures 4.42, 4.43 and 4.44 show the difference between the reported and administrative years of the longest employment phase according to the three definitions. As expected, there is much more overreporting than underreporting, and overreporting is highest for the strictest definition. The median difference is about 1 year for all definitions and for all incentives groups. A possible reason for overreporting the longest

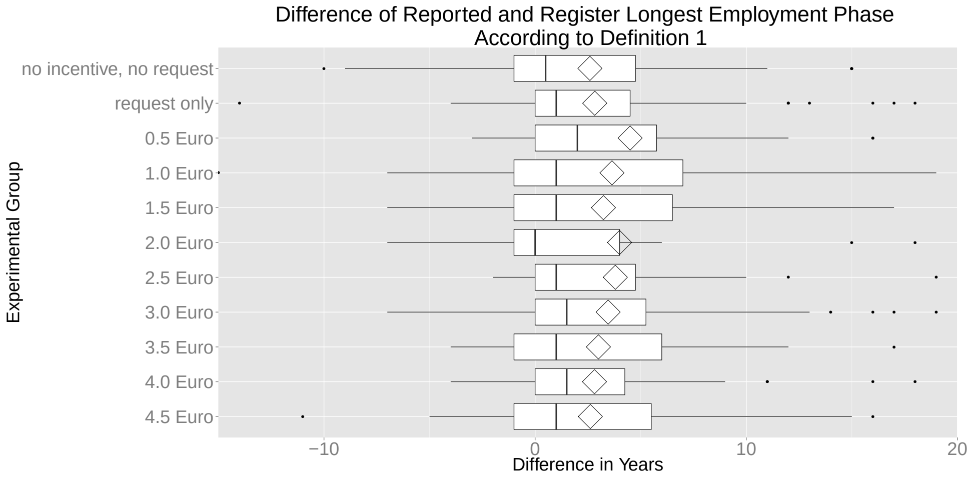


Figure 4.42: Difference of reported and register longest employment phase according to definition 1.

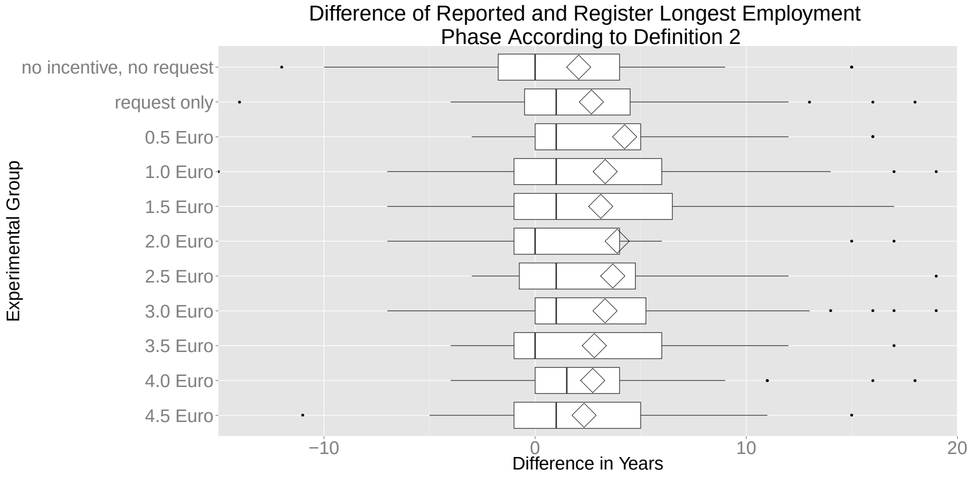


Figure 4.43: Difference of reported and register longest employment phase according to definition 2.

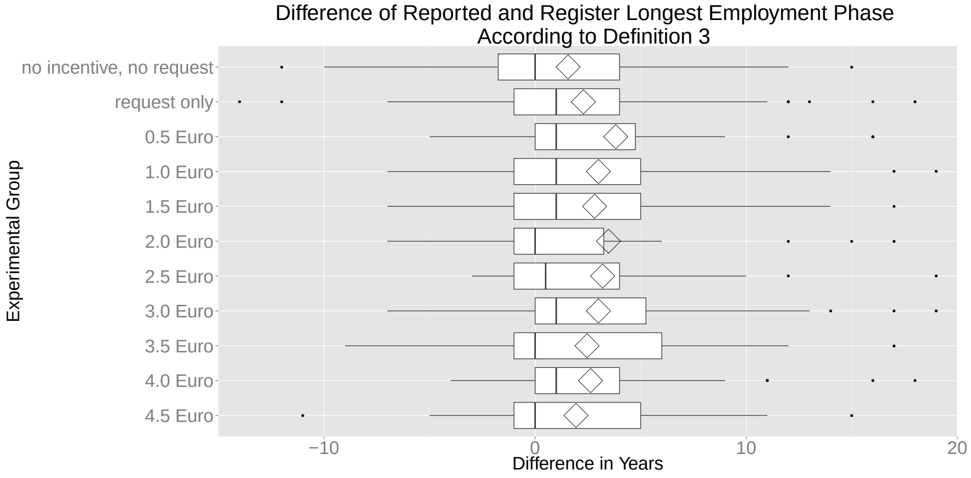


Figure 4.44: Difference of reported and register longest employment phase according to definition 3.

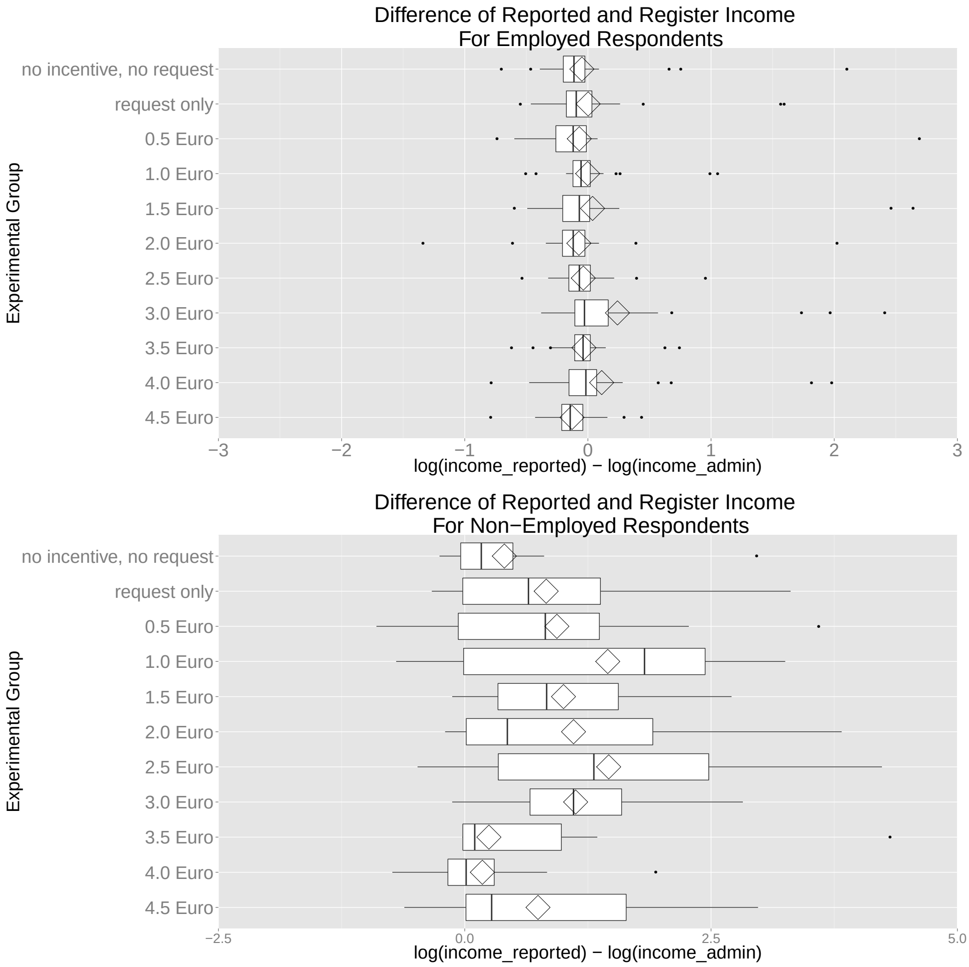


Figure 4.45: Difference in log Survey and log register income for employed and not employed respondents.

phase might be that respondents do not count episodes as breaks during which they have neither been employed nor been registered as unemployed. Underreporting can be due to counting gapless job changes as breaks. As there is no big difference between the three definitions, the first, the strictest one, is used for further analyses.

Figure 4.45 shows the difference between the reported and the administrative income for currently employed and unemployed respondents. The employed respondents exhibit an average measurement error that is close to zero for all incentive groups. In contrast, the non-employed respondents overreport their true income systematically. Again, no effect of the experimental conditions can be found.

4.3.3 Multivariate Findings

Modelling Time to Answer a Question, Item Nonresponse and Unsatisfactory Answers

In the following, we run multiple regression models using time to answer the question, item nonresponse and unsatisfactory answers as response variables. As the time to answer a question exhibits many strong outliers, we use median regression to model the common trend. In contrast to mean regression median regression is robust to outliers. Logistic regression is used for item nonresponse and unsatisfactory answers. It is not necessary to control for socio-demographic variables as these variables are distributed randomly across the experimental groups. However, the sample stratum however is controlled for as the stratum variable was not fully random for the experimental conditions. Further, all models include an indicator of the order of the health and finance block in order to control for the fact that the treatment might only influence the questions directly fielded after the experimental request. Further variables of interest are the incentive treatment, indicators for the response behavior prior to the experiment (time spent on the questionnaire before the request ("time before"), measured in quartiles, and the proportion of missing answers prior to the experiment (prop. NA), measured in tertiles), as well as interactions of the behavior and the treatment groups. For unsatisfactory answers, we also test whether the time spent on the question (in quintiles) improves the model fit, as the time spent on the question might improve the quality of the answer.

For all models, we test a linear against a quadratic term of the incentive, and only report the results for the better model in the following. We start with models only containing main effects and then add interactions if they improve the goodness of fit. The improvement is investigated in terms of likelihood-ratio tests for the logistic models, and in terms of Wald tests (Koenker and Bassett, 1982) for the median regressions.

Time to Answer a Question The request to think hard and the additional incentives are found to have no significant effect on the time to answer the questions, see Table 4.11 and Table 4.12 ($\alpha = 5\%$). In general, we find that the time a respondent spent on the questionnaire before the request is positively associated with the time spent on each of the questions of the experiment. Interactions with the incentive treatment are only part of the model for the employment question, though not significant. For the same question, the proportion of missing items prior to the experiment is also found to interact with the additional incentives: the negative effect of the proportion of item nonresponse is reduced by the additional incentives.

	Calories		Doctors		Diseases		Life Expectancy	
	Value	Std. Error	Value	Std. Error	Value	Std. Error	Value	Std. Error
Intercept	11.0000	0.8549	13.9930	1.4012	18.0010	1.0195	12.3750	0.8268
request	0.0003	0.7695	0.5064	1.3024	1.4259	1.0782	-0.7490	0.6532
incentives	0.0000	0.1477	-0.0010	0.2636	-0.5711	0.3313	0.2500	0.2444
time before cat. 2	2.0007	0.4727	2.0008	0.7860	5.2866	1.0419	3.1250	0.4466
time before cat. 3	3.0017	0.5119	5.9993	1.0587	8.0010	0.9073	6.3750	0.9983
time before cat. 4	6.0013	0.8239	12.4986	1.2218	14.1433	1.6829	11.001	1.3556
prop. NA cat. 2	-0.0013	0.5099	-0.4989	0.8683	0.2866	0.9171	0.6240	0.7143
prop. NA cat. 3	-1.0007	0.5095	-1.0002	0.9362	-1.5721	1.0652	-0.3760	0.6023
UB II stratum	0.9997	0.6267	0.5059	0.9672	3.1433	1.4139	1.6250	0.9593
employed stratum	-0.0013	0.4460	-0.4969	0.8107	0.8567	0.8332	0.3740	0.6250
question in first block	0.9997	0.4133	2.0013	0.7194	-0.5701	0.8237	5.2500	0.5419

Table 4.11: Results of median regression for time to answer the questions of the health block.

	Income		UB II		Employment		Living Standard	
	Value	Std. Error	Value	Std. Error	Value	Std. Error	Value	Std. Error
Intercept	24.0215	1.9963	7.9980	0.3725	46.5553	5.6244	22.3105	1.3816
request	0.0595	1.9120	0.0010	0.3319	-3.7078	3.8404	-0.3115	1.4625
incentives	-0.2400	0.4268	0.0001	0.0681	-0.5068	1.7793	0.3750	0.3727
time before cat. 2	4.3405	1.2662	1.0000	0.2303	9.5113	4.0010	5.8145	1.2884
time before cat. 3	11.4195	1.5759	1.9990	0.2299	17.4447	4.5914	9.4375	1.3050
time before cat. 4	20.6390	2.0678	3.0010	0.3612	36.1315	5.8417	15.5010	1.4888
prop. NA cat. 2	-1.9795	1.3636	0.0010	0.2269	-3.6858	5.4948	1.2500	1.1725
prop. NA cat. 3	-5.8410	1.4616	0.0000	0.2311	-22.4054	4.7266	-1.3730	1.1470
UB II stratum	0.4995	1.4550	1.0010	0.2471	-8.7928	2.2085	-1.3125	1.3335
employed stratum	0.3395	1.3342	-0.9990	0.2130	4.3014	2.2230	0.4365	1.0620
question in first block	-2.1815	1.1407	0.0000	0.1853	-0.8685	1.8243	-5.56050	0.9544
incentives*time before cat. 2	-	-	-	-	-0.3233	1.5681	-	-
incentives*time before cat. 3	-	-	-	-	-0.7532	1.5841	-	-
incentives*time before cat. 4	-	-	-	-	1.3670	2.7742	-	-
incentives*prop. NA cat. 2	-	-	-	-	0.8066	1.8920	-	-
incentives*prop. NA cat. 3	-	-	-	-	3.7293	1.6146	-	-

Table 4.12: Results of median regression for time to answer the questions of the finance block.

Item Nonresponse The results for the regression of item nonresponse on incentives can be found in Table 4.13 and Table 4.14. We do not include an analysis of item nonresponse in the UB II question, as the level of item nonresponse is very low (3%). We find significant quadratic effects of the incentives on the amount of item nonresponse for the calories question that interact with the proportion of item nonresponse prior to the request. The effect of the additional incentives on item nonresponse is reverse u-shaped. An illustration of the effect depending on nonresponse prior to the intervention can be found in Figure 4.46, with all other variables kept at their means or modes. Item nonresponse decreases with the amount of the incentive only for the respondents who already show low item nonresponse. In contrast, item nonresponse is even increased for respondents who show high amounts of item nonresponse in general. Item nonresponse in the calories question is mostly due to choosing the "don't know" option. For this question, "don't know" might in fact be an honest and correct answer. Thus, an increase in item nonresponse might be interpreted as an increase in response quality. It might be that highly motivated respondents, i.e., respondents who show little item nonresponse prior to the experiment, actually looked up the true answer and therefore decreased item nonresponse, whereas less motivated respondents decided to be honest and choose "don't know" when appropriate. Incentives also affect item nonresponse in the doctor visits question (10% level). The relation between the additional incentives and item nonresponse is found to be u-shaped with a minimum at about €2 to €3 on average.

	Calories		Doctors		Diseases		Life Expectancy	
	β	Std. Error	β	Std. Error	β	Std. Error	β	Std. Error
Intercept	-1.1515	0.3628	-1.1518	0.4099	-1.8108	0.5679	-1.0572	0.4148
request	-0.2640	0.3041	0.4631	0.3843	0.4577	0.5233	0.0156	0.3768
incentives	0.5590	0.2895	-0.7156	0.2596	-0.0331	0.0915	0.0664	0.0722
incentives^2	-0.0998	0.0638	0.1541	0.0570	-	-	-	-
time before cat. 2	-0.0598	0.2175	-0.2496	0.2652	-0.5438	0.3484	-0.4184	0.2713
time before cat. 3	0.0909	0.2124	-0.7860	0.2923	-0.3118	0.3279	-0.3235	0.2629
time before cat. 4	0.1347	0.2118	-0.5341	0.2754	-0.8480	0.3726	-0.8059	0.2858
prop. NA cat. 2	-0.5681	0.3719	-1.9789	0.3176	-1.6981	0.3633	-2.5795	0.3824
prop. NA cat. 3	-0.4297	0.3569	-0.9597	0.2289	-1.5005	0.3358	-0.9563	0.2178
UB II stratum	0.4157	0.1879	0.5800	0.2453	0.1677	0.3103	0.6681	0.2341
employed stratum	0.0673	0.1794	0.1035	0.2503	-0.0208	0.3101	-0.0789	0.2515
question in first block	0.1139	0.1496	0.0925	0.2009	-0.0493	0.2539	-0.0995	0.1975
incentives*prop. NA cat. 2	-0.3107	0.4474	-	-	-	-	-	-
incentives*prop. NA cat. 3	-0.1099	0.4049	-	-	-	-	-	-
incentives^2*prop. NA cat. 2	0.0019	0.1075	-	-	-	-	-	-
incentives^2*prop. NA cat. 3	0.0106	0.0919	-	-	-	-	-	-

Table 4.13: Results of logistic regression for item nonresponse in the questions of the health block.

Quadratic effects of the incentive on the probability of item nonresponse are also found for the income (10% level) and the employment questions. As for the doctor visits question, item nonresponse is lowest for an additional incentive of about €2 to €3 on average. We do not find effects of incentives on the two expectation questions: Neither item nonresponse in the standard of living, nor in the life expectancy question is affected by the incentive treatment. This might be explained by the fact that these two questions allow for a "50%"-response, which is often treated as an additional way of saying "don't know" (Bruine de Bruin et al., 2000, 2002). Thus, respondents choosing "don't know" might not lack motivation to choose a number, but rather be very unsure, a problem that can not be solved by the additional incentive. We do not find effects of the incentive treatment for the disease question, either. For this question, however, nonresponse is quite low and mostly driven by "no answer" (see 4.6). We find u-shaped effects for the doctor visits, income and employment questions. Separate logistic regressions for "don't know" and "no answer" show that incentives only significantly affect the "don't know" response for the calories, income and doctor visits questions. For these questions, the propensity of "no answer", which can be seen as reluctance to respond to the questions, is not affected by the incentive. This and the fact that there is no effect in the disease question can be interpreted as incentives being able to influence motivation to choose some substantive answer rather than "don't

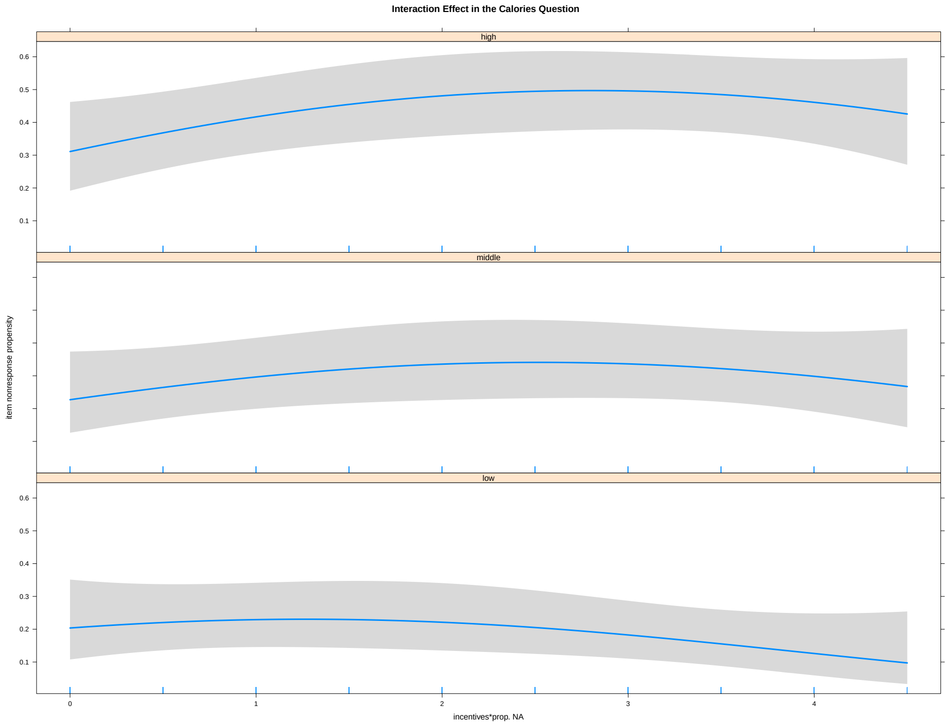


Figure 4.46: Effect of the additional incentive, interacted with the proportion of missing answers prior to the intervention, on item nonresponse in the calories question.

Effect of the additional incentive, interacted with the proportion of missing answers prior to the intervention, on item nonresponse in the calories question. The propensity for item nonresponse is found to decrease with the amount of incentives for respondents who already show a low item nonresponse propensity. For respondents with a high item nonresponse propensity, however, item nonresponse propensity in the calories question is further increased.

know", but not being able to affect reluctance to give an answer. We generally find that the amount of missing information before the request is positively, the time spent on the interview before the request is negatively associated with item nonresponse in the questions of the experiment. Only for the calories question, incentives interact with the behavior prior to the intervention.

	Income		Employment		Living Standard	
	β	Std. Error	β	Std. Error	β	Std. Error
Intercept	-0.1523	0.3158	-0.8505	0.4630	-0.8899	0.3684
request	0.1160	0.2990	0.4253	0.4458	-0.1045	0.3186
incentives	-0.3722	0.2045	-0.6154	0.3023	-0.0626	0.0690
incentives^2	0.0798	0.0446	0.1268	0.0668	-	-
time before cat. 2	-0.4974	0.2147	-0.9154	0.3088	-0.2273	0.2481
time before cat. 3	-0.5477	0.2149	-1.1000	0.3172	-0.5398	0.2616
time before cat. 4	-0.5409	0.2151	-1.2263	0.3212	-0.4674	0.2568
prop. NA cat. 2	-1.4248	0.2023	-3.2310	0.5253	-1.4109	0.2512
prop. NA cat. 3	-1.0190	0.1832	-1.7532	0.2876	-0.8843	0.2151
UB II stratum	0.0804	0.1994	1.2000	0.2718	0.9539	0.2208
employed stratum	0.0702	0.1830	-0.0587	0.3128	0.0341	0.2371
question in first block	0.1324	0.1553	0.0478	0.2299	-0.1299	0.1844

Table 4.14: Results of logistic regression for item nonresponse in the questions of the finance block.

Modelling Unsatisfactory Answers Tables 4.15 and 4.16 show the results of the regressions for rounding, heaping, an incorrect answers, for all respondents who gave a substantive answer to the questions of the incentive experiment. For the life expectancy question, we find significant effects of the request to think hard and the additional incentives. The request to think hard decreases the probability of heaping to 50%. At the same time, the additional incentives increase the heaping propensity. Still, the groups receiving an additional incentive show less heaping than the "no incentive, no request" group up to an incentive of €4. We do not find interactions between the incentive and behavior prior to the request for any of the experimental questions. Time spent on the question is found to have an effect on the doctor visits and calories questions: respondents taking less time than the third quintile show less, respondents taking more time show more wrong answers in the calories question. In contrast, respondents in the fifth time quintile show less heaping in the doctor visits question than the respondents in the third quintile.

Modeling the Time Spent on the Experimental Block and Total Amounts of Item Nonresponse and Unsatisfactory Answers We do not find significant effects of the request to think hard or the additional incentives on the time spent on all eight questions of the incentive experiment. Moreover, the total amount of item nonresponse and unsatisfactory answers does not significantly differ by the experimental conditions. We find that the time spent on the interview before the experiment has a positive effect on the time spent on the questions of the experiment, and a negative effect

	Calories		Doctors		Life Expectancy	
	β	Std. Error	β	Std. Error	β	Std. Error
Intercept	0.4325	0.3423	-0.8861	0.3286	-2.5952	0.5012
request	0.1009	0.2670	0.1155	0.2521	-0.8802	0.4134
incentives	-0.0122	0.0553	0.0557	0.0515	0.1998	0.0928
time before cat. 2	0.3963	0.2152	0.3021	0.2032	0.2984	0.3905
time before cat. 3	0.1214	0.2164	0.6505	0.2039	0.5911	0.3728
time before cat. 4	-0.1191	0.2264	0.6337	0.2157	0.5636	0.3735
prop. NA cat. 2	0.0598	0.1793	0.0168	0.1613	0.6929	0.2974
prop. NA cat. 3	0.1635	0.1930	-0.5379	0.1786	0.3874	0.3255
UB II stratum	0.0334	0.2083	0.1140	0.1855	0.5052	0.2984
employed stratum	0.0380	0.1720	0.0715	0.1574	-0.0541	0.2913
question in first block	-0.2144	0.1523	0.1859	0.1390	-0.7317	0.2491
response time 1	-0.2956	0.2386	-0.0544	0.2268	-	-
response time 2	-0.4369	0.2267	-0.0881	0.2099	-	-
response time 4	0.4185	0.2407	-0.2240	0.2142	-	-
response time 5	0.6336	0.2387	-0.8312	0.2240	-	-

Table 4.15: Results of logistic regression for rounding and heaping in the questions of the health block.

	Income		Employment		Living Standard	
	β	Std. Error	β	Std. Error	β	Std. Error
Intercept	-1.2742	0.3536	-2.2407	0.3726	-1.4955	0.3618
request	0.4290	0.3160	0.2372	0.3111	-0.3664	0.3051
incentives	-0.0392	0.0608	0.0579	0.0586	-0.0204	0.0678
time before 2	0.2051	0.2372	0.8325	0.2529	0.2613	0.2612
time before cat. 3	0.1025	0.2395	0.7082	0.2536	-0.0511	0.2704
time before cat. 4	0.0296	0.2450	0.5592	0.2583	0.2732	0.2589
prop. NA cat. 2	-0.4405	0.1975	0.0540	0.1942	0.2257	0.2200
prop. NA cat. 3	0.0625	0.2052	0.4541	0.2008	0.3352	0.2283
UB II stratum	-0.9308	0.2588	-0.5360	0.2364	0.1220	0.2379
employed stratum	0.2047	0.1802	-0.0325	0.1773	-0.0563	0.2065
question in first block	-0.0841	0.1652	0.0788	0.1606	-0.1603	0.1800

Table 4.16: Results of logistic regression for rounding and heaping in the questions of the finance block.

on the sum of item nonresponse. Item nonresponse prior to the request is positively associated with item nonresponse in the experiment, and it is negatively associated with the time spent on the experiment. None of these indicators interact with the incentive experiment.

Modelling The Correct Answers to the Validation Questions

For the validation questions, we use the same covariates as before. We also use logistic models for the misreport in UB II, and median regression for measurement error in the years of the longest employment and measurement error in income. As we have seen in Section 4.3.2, the measurement error in the longest employment spell is not symmetric but skewed to the right. The measurement error distribution for income is not symmetric either, especially for currently non-employed respondents. As for the models for the unsatisfactory answers, we include the time spent on the questions, measured in quintiles. For the employment history and income question, we also include an indicator for current employment according to the administrative data

	Time Spent on The Experiment		Sum of Missing Items		Sum of Missing or Unsatisfactory Items	
	β	Std. Error	β	Std. Error	β	Std. Error
Intercept	177.0826	13.7667	-0.6958	0.1422	0.6785	0.0843
request	-1.3884	13.1672	0.0111	0.1111	0.0222	0.0714
incentives	-0.6116	2.3678	-0.0072	0.0224	0.0059	0.0143
time before cat. 2	36.2498	7.5944	-0.3169	0.0830	-0.0015	0.0549
time before cat. 3	76.1366	7.7990	-0.3501	0.0832	-0.0225	0.0550
time before cat. 4	158.0541	13.8539	-0.4036	0.0841	-0.0616	0.0557
prop. NA cat. 2	-9.0826	7.7425	0.6008	0.1011	0.1226	0.0515
prop. NA cat. 3	-31.9429	8.3888	1.4067	0.0914	0.4321	0.049
UB II stratum	-0.2498	8.1739	0.4248	0.0733	0.1182	0.0496
employed stratum	-5.6922	7.6561	0.0268	0.0773	0.0155	0.0462
question in first block	5.5005	6.6853	0.0220	0.0607	-0.0247	0.0390

Table 4.17: Results of Poisson regression for item nonresponse and unsatisfactory answers and median regression for time spent on the experiment.

(yes/no). We expect these questions to be harder to answer for respondents who are currently non-employed. Thinking longer about the answer could be interpreted as an indicator of uncertainty.

We do not find an effect of the request or the additional incentive on misreporting UB II (see Table 4.18). The respondents sampled from the employed stratum show significantly less misreporting than respondents from the other two strata, whereas respondents from the UB II stratum exhibit most misreporting. As respondents from the employed stratum have never receive any social benefit, the question is not hard to answer for them, and only respondents from the other two strata might suffer from recall error or have an incentive to give the socially desired answer. Surprisingly, there is more misreporting for respondents taking more time to answer the question than for the quicker ones.

There is no effect of the request or additional incentive on measurement error for the employment phase (Table 4.18). Respondents who spent the most time on this question show the highest measurement error. As expected, measurement error is lower for currently employed than currently non-employed respondents.

The best model for measurement error in income includes interactions between the additional incentives and behavior prior to the experiment, though not significant (Table 4.18). Measurement error is smaller for the last month's income of currently employed respondents than for the last income of currently non-employed respondents.

4.4 Summary and Conclusion

In this Chapter, we have presented the results of an incentive experiment within a web survey. In the experiment, we tried to increase the quality of survey response by asking some of the respondents to take their time, think hard about the next questions, and give as good answers as possible. We have followed the research by Cannell et al.

	UB II misreport		Employment phase measurement error		Income measurement error	
	β	Std. Error	β	Std. Error	β	Std. Error
(Intercept)	-2.8687	0.8103	1.7665	1.2925	0.3274	0.1764
request	0.0540	0.6395	0.0060	0.9237	0.1198	0.0797
incentives	0.2078	0.1219	0.0958	0.1822	0.0212	0.0964
incentives^2	–	–	–	–	-0.0037	0.0204
consent late	0.2921	0.6546	0.7725	1.0784	0.0131	0.0679
time before cat. 2	-0.0233	0.5174	0.5030	0.6757	0.0721	0.0958
time before cat. 3	-0.3467	0.5244	0.8443	0.7261	-0.0003	0.1053
time before cat. 4	-0.3063	0.5018	0.7725	0.7662	0.0452	0.1164
prop. NA cat. 2	0.0507	0.3929	-0.4491	0.5800	0.1033	0.0854
prop. NA cat. 3	-0.1262	0.4195	0.1198	0.7015	0.0509	0.1038
UB II stratum	0.8498	0.3581	-0.5150	0.6939	0.0499	0.0624
employed stratum	-1.0374	0.5311	-0.5868	0.5911	-0.0176	0.0430
question in first block	-0.3537	0.3291	-0.5030	0.5160	0.0097	0.0427
employed (admin)	–	–	-2.1677	0.6199	-0.6118	0.1353
response time 1	-1.2191	0.7006	0.5269	0.7188	–	–
response time 2	-0.4871	0.6954	0.7904	0.7727	–	–
response time 4	0.1324	0.4891	0.7305	0.7774	–	–
response time 5	1.1204	0.4398	3.4251	0.9698	–	–
incentives*time before cat. 2	–	–	–	–	0.0705	0.1226
incentives*time before cat. 3	–	–	–	–	0.0615	0.1310
incentives*time before cat. 4	–	–	–	–	0.0424	0.1302
incentives*prop. NA cat. 2	–	–	–	–	-0.1167	0.1035
incentives*prop. NA cat. 3	–	–	–	–	0.0467	0.1289
incentives^2*time before cat. 2	–	–	–	–	-0.0213	0.0294
incentives^2*time before cat. 3	–	–	–	–	-0.0196	0.0300
incentives^2*time before cat. 4	–	–	–	–	-0.0155	0.0297
incentives^2*prop. NA cat. 2	–	–	–	–	0.0300	0.0246
incentives^2*prop. NA cat. 3	–	–	–	–	-0.0096	0.0293

Table 4.18: Results of logistic and median regression for validation questions.

(1981) who found for personal interviews that such requests can increase response quality as measured by the numbers of topics mentioned in answers to open-ended questions and precision of answers. In our experiment, for most of the respondents who received the request to think hard, the request was combined with an additional incentive from €0.50 to €4.50. We have hypothesized that the request to think hard increases the quality of the answers and the time spent on the question compared to the respondents who do not receive this request. Further, we have assumed the quality to increase as the amount of the additional incentive increases. Our hypotheses are only confirmed in part. We find the experimental intervention not to affect response times of any of the questions and only to affect the quality of the answers to some of the questions.

Moreover, incentives lead to reduction of item nonresponse for some questions. In the multivariate analyses, we find that the additional incentive affects item nonresponse in the calories, doctor visits, income and employment questions. Incentives are found to have no effect in the expectancy questions. This might be explained by the fact that these questions allow less motivated respondents to choose the non-informative "50%"-response, and item nonresponse might be differently motivated than for questions that do not have such a response option. In general, there is some evidence that incentives are able to motivate respondents not to respond "don't know", but they are not able to prevent "no answer". For the calories question, the additional incentives interact with the proportion of item nonresponse prior to the experiment: only respondents who already show little item nonresponse show a further decrease in item nonresponse as the incentive increases. For respondents showing middle or high item nonresponse prior to the intervention, the item nonresponse propensity is not affected or even increased. We argue that for this question, "don't know" might be the best answer for some respondents, and the increase in item nonresponse for the less motivated respondents, i.e., respondents showing more than average item nonresponse prior to the intervention, might be a sign for increased response quality. The highly motivated respondents show a decrease in item nonresponse that might be due to looking up the right answer. Moreover, we find that the request to think hard decreases heaping to 50% in the life expectancy question, but not in the standard of living question for which the heap at 50% is much more pronounced. The propensities for unsatisfactory answers is not affected by the incentives for any other question. This might be due to respondents just not knowing how to provide good answers. To keep the instruction short, we did not include explanations how to improve response quality.

In order to overcome some of the shortcomings of this experiment, based on this Chapter, we conducted an additional experiment in the later IAB web survey "Herausforderungen am deutschen Arbeitsmarkt 2014". In this experiment, respondents did not only receive the request to think hard, but also were asked to agree to giving good answers, as was done by Cannell et al. (1981). Furthermore, respondents were told to check outside sources or look up answers if possible. In this experiment however, no additional incentives were given. At the end of the interview, respondents were asked to judge the question complexity and report about their use of outside information. These data will help to understand respondents reaction on the request to think hard. Validation data for this survey will be available as of the beginning of 2016.

5 Item Nonresponse and Measurement Error in Income Questions

5.1 Introduction

It is a well known fact that surveys asking for income suffer from a high amount of item nonresponse in income questions. For personal and telephone surveys, one often finds income to have more missing information than any other question (Krumpal, 2013). Even if an income is reported, the quality of this report is usually unknown. Any difference between the reported and true income value can lead to biased estimates. Especially if item nonresponse and measurement error are related to the true unknown income value, this will lead to bias that can not in general be solved by imputation or correction procedures (Rubin, 1976).

Personal income is often found to affect measurement error in personal interviews (see for example Moore et al. (2000); Bound et al. (2001, 1994). In contrast, it is not clear whether this relationship exists in web surveys as well. There is some evidence that web surveys are able to decrease measurement error in sensitive questions (Kreuter et al., 2008) but it is not known if the same findings hold for income questions.

In the following Chapter, we analyze the effect of personal income and other personal characteristics on the propensity of not reporting an income value and the quality of the reported income. We do this by linking survey data from the telephone and web survey "Work and Consumption in Germany" (see Chapter 2.3) to administrative data from the German Federal Employment Agency, assuming that the administrative data represent the true income (see Chapter 2.1). Our main interest is to analyze whether the effect of personal income is the same for both modes, or if we can find mode differences.

We will first study item nonresponse and measurement error separately for both "sub-surveys" to see whether the usual findings can be replicated for our survey. As the web survey is found to be highly selective for some socio-demographic variables (e.g., age, gender, employment type), differences in income effects between the modes can not necessarily be attributed to the survey mode. To be able to conduct causal inference on the effect of survey mode on item nonresponse and measurement error, we perform propensity score matching of respondents to both modes in a second step.

5.2 Background

Errors can occur in all steps of the question answering process which in general consists of three steps (Tourangeau, 1984): respondents need to (i) comprehend the question, e.g., the income concept, (ii) recall their past income, and (iii) report an answer that matches their income. In the first two steps, errors can for example occur because a respondent is confused about the income concept and definitions (e.g. the definition of gross and net income), is not able to recall her exact income, or just does not know her income (Moore et al., 2000). Moreover, respondents might forget small income amounts, or include black labor (Nordberg et al., 2004). In the third step, measurement error might occur because the respondent might not want to share (true) income information because she thinks it is too private and personal, or sensitive (Bradburn and Sudman, 1979).

In general, three aspects of *sensitive* questions can be distinguished (Tourangeau et al., 2000): social (un)desirablity of the answer, invasion of privacy, and risk of disclosure of the answers to third parties. Regarding the *income* question, we expect the following interrelations to be present to some extent: As *social desirability* is found to be a bigger problem in personal interviews than in web surveys (Heerwegh, 2009; Kreuter et al., 2008), the survey mode could affect income measurement error. While there is no reason why the *invasion of privacy* should be different between survey modes, the *risk of disclosure* of the answers could be perceived to depend on the mode.

As Heerwegh (2009) summarizes, building trust is harder in the impersonal web survey than in a personal interview. Also, the fear of database hacks can inhibit trust-building. However, there is no empirical support for this assumption. In our case, we do not expect the perceived risk of disclosure to differ between the modes: both surveys were conducted by the same well-known field agency, were stated to be conducted by the IAB, and anonymity was ensured in the invitation letters of both surveys. In contrast, social desirability concerns are likely to differ between the modes. As Heerwegh (2009) argues, the social distance is larger in a web than in a face-to-face survey. As in the telephone survey there is also an interviewer present, one could argue that the social distance is lower in the web than in the telephone survey, as well. We therefore expect that the web survey will lead to less socially desirable responses than the telephone interview.

In general, there is some empirical evidence that self-administered interviews show less social desirability bias than face-to-face interviews (Tourangeau and Yan, 2007). Most studies however can not make use of validation data and are limited to comparing distributions and evaluating their findings under the assumption that higher reports of social undesired behavior are a sign for less social desirability bias.

For example, Heerwegh (2009) concludes that the web mode shows less socially desirable answers than the face-to-face mode when comparing response distributions of questions on happiness, subjective health, and integration. At the same time, he finds a decrease in data quality in the web in terms of higher item nonresponse and "don't know"-answers. Tourangeau and Smith (1996) find that the differences in the reported number of sex partners between women and men decreases for self-administered surveys compared to personal interviews as self-administration reduces the fear of embarrassment and increases privacy. Presser and Stinson (1998) find that respondents in self-administered surveys report less socially desirable weekly religious attendance as compared to conventional interviews.

Kreuter et al. (2008) conclude from their validation study that the web mode yields more accurate reports of sensitive information than the IVR and CATI mode, and that differences are larger for socially undesired than for socially desired items. Moreover, the perceived item sensitivity depends on survey mode and a person's true status. Respondents who have an undesired status find questions more sensitive than others. Respondents, asked about their expectations, reported to expect more misreport in the CATI than the web or IVR mode.

The effect of mode on reporting one's income is specifically interesting as income is a particularly sensitive question for most people as social class can be expected to be determined by income to a large extent. Moreover, income is a focal quantity in economic research, and any item nonresponse or measurement error can seriously bias corresponding analyses and inflate standard errors.

The imputation of missing income values on socio-demographic and other survey variables will not lead to unbiased results if the missing mechanism is related to the true income (Rubin, 1976), i.e., the income information is missing not at random. Hence, imputation on observables will in general lead to biases. A relationship between missing income information and true income is often conjectured (Rässler, 2000; Riphahn and Serfling, 2005). It has been studied by analyzing the relation of the propensity for item nonresponse and survey indicators of income (see for example Schräpler (2004, 2006)) but has to our knowledge never been analyzed in a validation study. Knowing a person's true income, we can directly model the dependence of item nonresponse and income rather than being forced to make assumptions about the relationship of true income and some survey indicators. We are the first to compare income item nonresponse between a telephone and a web survey.

Lacking access to validation data, Smith (1991) finds for the General Social Survey

(GSS) that item nonresponse in the income variable is higher among respondents who are less educated and work on lower employment levels than on average. In the context of mode effect studies for financial asset questions, Essig and Winter (2009) conclude that a self-administered drop-off questionnaire yields better data quality (in terms of item nonresponse and heaping) than a personal interview.

Other than for item nonresponse in income questions, validation studies on income measurement error have been conducted frequently, and measurement error is usually found to be non-classical. In the context of linear regression, even classical measurement error will always lead to attenuation bias, i.e., bias towards zero, whereas the direction of bias in the case of non-classical measurement error can not be predicted (Carroll et al., 2006; Angrist and Pischke, 2008; Biemer and Lyberg, 2003). The present study builds on prior studies that analyze measurement error in reported income by comparing the survey report to some kind of external income source that they presume to be true (see for example Bound et al. (2001) and Moore et al. (2000) for an overview).

Most of the validation studies were conducted in the US, and many of them use administrative data from the Social Security Administration (SSA) containing employers' tax reports. SSA data are used to validate survey reports in the Current Population Survey (CPS) (Bound and Krueger, 1991; Bollinger, 1998), the Survey of Income and Program Participation (SIPP)(Pedace and Bates, 2000; Gottschalk and Huynh, 2010; Kim and Tamborini, 2014; Abowd and Stinson, 2013), and the Health and Retirement Survey (HRS) (Bricker and Engelhardt, 2008). Other studies use employers' payroll information from one single firm to validate the reported income in the Panel Study of Income Dynamics (PSID) (Duncan and Hill, 1985; Bound et al., 1994).

There are also some studies from Scandinavian countries which link responses to administrative data. Examples are Nordberg et al. (2004) for the first wave of the Finnish part of the European Community Household Panel (ECHP), (Kapteyn and Ypma, 2007) for the Swedish part of the Survey of Health, Ageing and Retirement in Europe (SHARE), and (Kristensen and Westergaard-Nielsen, 2007) for the Danish part of the ECHP.

All studies, while being different in focus, find mean-reverting measurement error in reported income. This is, the correlation between the measurement error and true income is usually found to be negative, indicating that respondents with a high income under-report, respondents with a low income over-report their incomes. Also, some studies come to the conclusion that the negative correlation between a person's true and reported income is largely driven by strong over-reporting among low-income respondents and moderate under-reporting of income in the higher deciles (e.g Bollinger (1998); Nordberg et al. (2004)). For the CPS and SIPP, Bound and Krueger (1991) and Gottschalk and Huynh (2010) also show that the measurement error is correlated over panel waves.

Some studies weaken the assumption of the administrative data being free of errors by allowing for measurement error in both data sources (Kapteyn and Ypma, 2007; Abowd and Stinson, 2013). For the Swedish part of the SHARE, Kapteyn and Ypma (2007) discuss the possibility of a mismatch between survey and administrative data, finding that the mean-reverting effect gets smaller once one allows for error in the matching procedure. They also find the administrative data to be biased to some degree. Abowd and Stinson (2013) allow for measurement error in both data sources and give three possible reasons for not considering the administrative data as "truth": definitional differences between the data sources, error in the administrative data, and error in the matching process. The administrative income records we are using are known to be very accurate (see Chapter 2), and as we draw our sample from administrative data and can link survey responses back to the administrative data, errors in the matching process are virtually impossible. However, as explained in Chapter 2, there are small definitional differences between the survey and administrative data. These can be expected to occur in the same way in both modes, not corrupting the mode comparison.

The literature on effects of socio-demographics and job characteristics on measurement error in reporting of income shows mixed results: There is some evidence for income under-reporting by younger and older single persons (Nordberg et al., 2004). Also, females are found to misreport their income less often than men (Bollinger, 1998; Bound and Krueger, 1991; Kristensen and Westergaard-Nielsen, 2007). Part-time employed respondents report with more error than full-time employed, and reporting quality increases with age (Kristensen and Westergaard-Nielsen, 2007). The effect of age differs by gender (Gottschalk and Huynh, 2010). Some studies include education in their analysis, but only find weak or no association with measurement error (Bound and Krueger, 1991; Bricker and Engelhardt, 2008; Pedace and Bates, 2000). Also, race is not found to affect income measurement error Pedace and Bates (2000).

In a recent validation study, comparing SIPP responses to SSA records, Kim and Tamborini (2014) allow the effects of socio-demographics to differ by true income. Running separate regressions for respondents belonging to different income quintiles based on the SSA data, they find that some effects differ between the income subsamples: in the low-income groups, black respondents show higher over-reporting than whites; in the high-income groups, blacks show higher under-reporting, suggesting that mean-reversion is stronger among black than white respondents. A similar result is found in Kim and Tamborini (2012). Moreover, women are less likely than men to over-report their income in the lowest income quintile, but more likely to under-report their income in the highest quintile. Both effects cancel each other out. (Not) having included the interaction of true income and other socio-demographic variables within regression analyses might explain the mixed results in the studies cited above.

All these studies need to link survey and administrative data, a task which is to some degree based on probabilistic procedures. The strength of our study is that we drew the sample from the administrative records and can *directly* link the survey information back to the administrative data using a unique person identifier — provided that consent to data linkage is given. Most of the validation data from the studies mentioned above stem from tax information. As Moore et al. (2000) point out, tax reports have some drawbacks: they only cover earnings above a certain threshold, some people are not captured in the tax system, and special regulations such as joint return for married couples make it hard to identify a single person's income. The validation data set we are using captures the respondents' individual income before taxes. However, not all respondents are captured in the social security system and hence part of the administrative data (see 2.1).

Last, while all previous studies focus on face-to-face surveys, we are the first to study item nonresponse and measurement error in a web survey, and to compare the results to a telephone survey based on the same population.

Our research questions are:

- Can previous findings be replicated using German administrative records?
- Is the relation between income and item nonresponse in the income question moderated by the survey mode?
- Is the relation between income and measurement error in the income question moderated by survey mode?

5.3 Data

For our analyses, we use the survey data from the telephone and web survey "Work and Consumption in Germany" (for a description of the study see Chapter 2). Like we did in Chapter 4, we are linking the consenting respondents from both surveys (2,281 consenters for the CATI, 662 consenters for the web survey) to the administrative records of the FEA. We again exclude all respondents whose reported age differs from the administrative age by more than one year (see Chapter 4.3.2), as we can not be sure that we have interviewed the right person. We also need to drop all respondents who did not receive the question in CATI[1], respondents who report to be self-employed, and respondents whose administrative income is censored at the social security contribution assessment ceiling (for more information on the social

[1] Other than in the web survey, the income question was filtered by employment status in the CATI survey.

security contribution assessment ceiling see Chapter 2.1). Finally, we can only focus on respondents who are employed according to the administrative data and receive some income from labor. Seven strong outliers, five in the CATI and two in the web survey mode, with a reported monthly income of at least €10,000 were dropped, leaving us with 1,234 and 381 respondents (see Table 5.1) for the item nonresponse, and with 1,069 and 327 of item respondents for the measurement error analyses, respectively.

	CATI	web
consenters	2281	662
after age validation	2260	658
received question	1509	658
without self-employed	1429	621
under assessment ceiling	1367	568
employed with income (administrative)	1239	383
response to income question	**1074** (RR = 86.7 %)	**329** (RR= 85.9 %)
employed with reported income $<$ €10.000	**1069**	**327**

Table 5.1: Case numbers for CATI and web.

In both surveys, the income question was asked in exactly the same way.[2] The only difference is that the web survey respondents received the additional request to report the last income ever if they are unemployed at the time of the interview. We used the standard question that is used in all surveys of the IAB because this question is well tested and yields at the income that researchers are interested in. The question wording is:

> *What was your employment income last month? In the case of multiple jobs please calculate the total sum. Please indicate your gross income, i.e. your income before the deduction of tax and social security contributions. If you had special payments last month, such as a Christmas bonus or back payments, do not include these. However, do include any pay for overtime.*
>
> *For self-employed work please indicate your monthly profit before tax.*

This question differs slightly from the information in the administrative data (see Chapter 2) as the administrative data contains the income on a spell level, not on a monthly level, and includes bonuses. Just like we did in Chapter 4, we make the assumption that the bonus payments are negligible and that the income is constant over the whole spell. By doing this, we are able to compute a respondent's last month's income, and to validate the income question.

[2] For the German wording see Appendix A

	1st Qu.	Median	Mean	3rd Qu.
CATI	1,014	1,732	1,972	2,797
Web	1,272	2,121	2,230	3,039

Table 5.2: Summary of administrative income for CATI and web respondents.

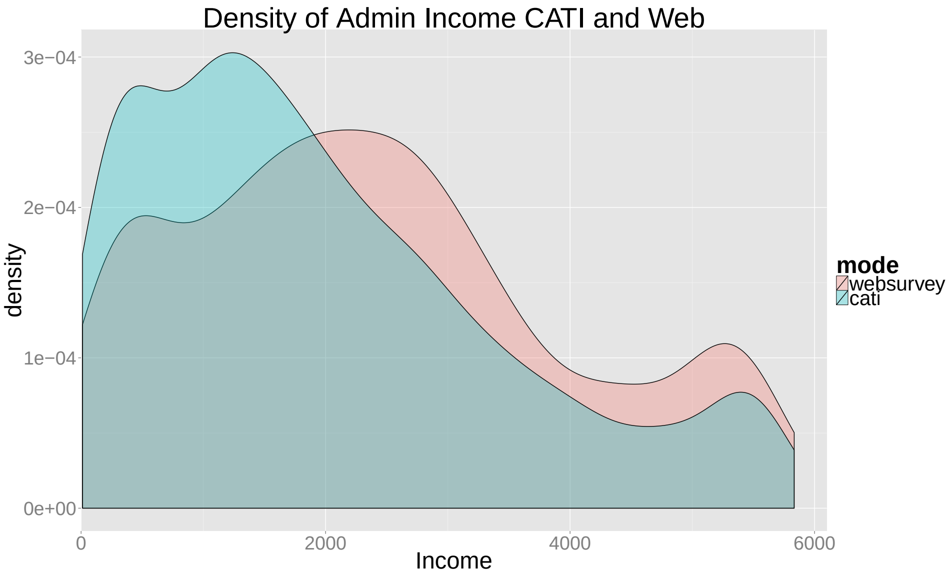


Figure 5.1: Kernel density estimates of administrative income for all CATI and web survey respondents. Note that these also include respondents who did not consent to data linkage and are not part of the further analyses.

As the cases were assigned to the modes randomly, the income distribution of the gross sample does not differ between the samples. However, the income distributions obtained from the administrative data differ substantively between the respondents to the CATI and web survey (see Figure 5.1 and Table 5.2; the densities for the web and telephone survey samples can be found in Appendix B).

Unit nonresponse analyses for these data show that the CATI respondents represent the general population quite well, and on aggregate do not show nonresponse bias due to income or other socio-demographic characteristics (Felderer and Kirchner, 2013). In the web survey, however, individuals with a high income are overrepresented while low-income people are underrepresented. The average income is higher in the web survey than in the CATI survey, and there is more variation in the web survey income (see Table 5.2). Note that the distribution is censored at the income limit (see Chapter 2).

Figures 5.2 and 5.3 show histograms of the reported income in the respondent sample for each study, including nonconsenters to the data linkage question. For readability, the graphs are focused on the range of €0 to €10,000. The reported income shows much higher outliers in the web than in the CATI mode which is

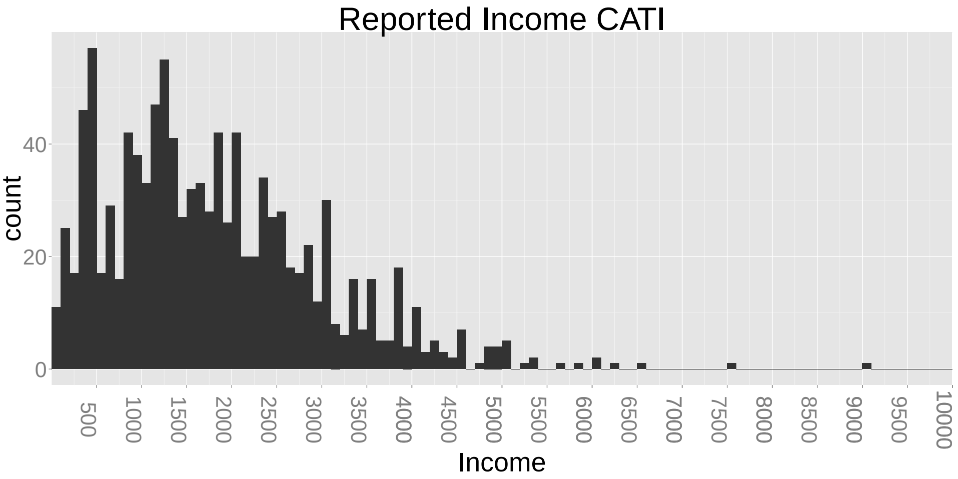


Figure 5.2: Reported income CATI.

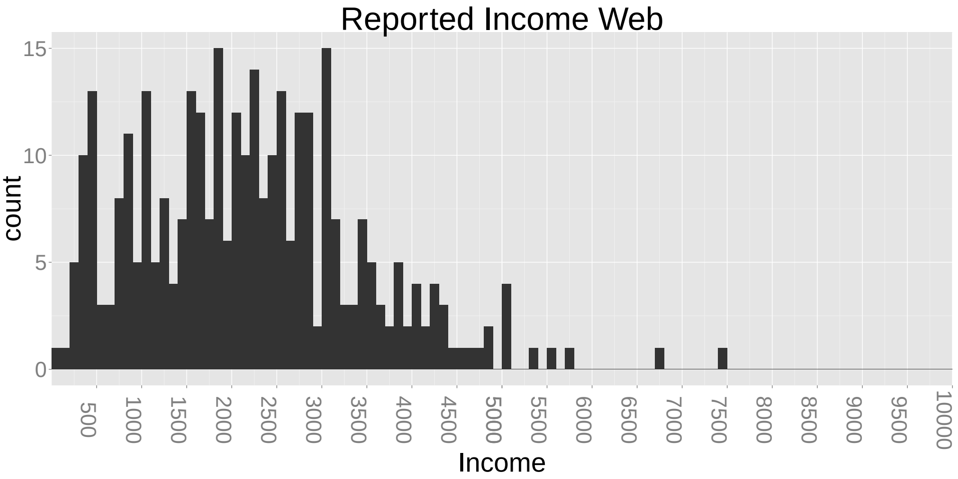


Figure 5.3: Reported income web.

probably due to the absence of control of an interviewer. In both modes, the usual heaping to multiples of 500 and 1,000 can be observed.

We analyze item nonresponse and measurement error in the income question in two steps: In the first step, we run logistic regressions for item nonresponse and linear regressions for measurement error. In both models, we employ the same regressors, i.e., administrative income and controls for both modes. Measurement error is thereby defined as the difference between the reported and the administrative income,

$$\text{Measurement error}_i = \text{Income}_i^{reported} - \text{Income}_i^{admin}.$$

As Kim and Tamborini (2012) show, the income effect is likely to differ with respect to socio-demographic characteristics. Rather than running separate regressions for the different income groups as in Kim and Tamborini (2012), we include interaction terms of administrative income and socio-demographic variables in every regression model. For all models, polynomial terms are used for the administrative income. We allow for polynomials up to the power of three and interactions of administrative income with all covariates. The best models after AIC selection are presented in the following sections. If the best model after AIC selection included interactions of a control variable *and* one or more income polynomial terms, *all* remaining interactions of the particular control variable and the income terms were added to the final model. Selecting the models by AIC can lead to models containing different variables for the CATI and web mode which can not be compared directly. Our aim is to find the best prediction model for each mode by avoiding over-fitting which occurs if variables that do not contribute to predict item nonresponse and measurement error are included.

Even though sample cases were randomly assigned to the survey modes, the actual respondents differ between the web and CATI mode. Due to this self-selection, the allocation of the respondents to the two modes is not random. If the relationship between the true administrative income and income misreporting is found to differ by mode, this can have two possible causes: It may either be due to different *reporting*, caused by different aspects of the survey mode, or by the fact that different *individuals* respond to the survey modes. In order to attribute differences in the income relationships between the surveys to the survey modes, we perform propensity score matching to disentangle response and mode effects: For each respondent of the web survey, we search for a telephone survey respondent who shares the same characteristics as the web survey respondent, represented by her propensity score. Comparing the web survey respondents to the respondents of the telephone survey – who could just as well have answered the web survey if they would have been asked to – solves the problem of self-selection, and it allows to interpret differences in income misreporting as being induced by the survey mode. A similar approach has been used by Atkeson et al. (2014).

All models contain covariates available from the survey *or* administrative data. Gender, age, employment status (full time employed, part time employed, vocational training and mini job) are taken from administrative records, as well as unemployment benefit within the last 12 month (yes/no). The same holds for the length of the current employment spell (with a maximum of 365 days, see chapter 2.1), and an indicator whether the respondent has more than one job. Education (low, middle, high) is taken from the survey. For all respondents who did not report the level of education (less than 1%), we took the last available information on education from the administrative records. Indicators for the spell duration and having more than one job are derived from the administrative records and included in the models. This can be justified by the fact that a relationship between ongoing job episodes and measurement error has been found before (Kim and Tamborini, 2014) and as it might be hard to sum up income from different sources. The employment status is lagged by one month as the income question refers to the previous month. There are a few respondents who were not employed in the month of the interview but have been employed in the month before. We therefore include an indicator for current employment in the interview month (yes/no). Models also control for the positioning of the linkage consent question (consent late: yes/no) where appropriate. The sample stratum is not included as it is captured by "UB II in the last 12 month".

5.4 Results

First, results for item nonresponse and measurement error are presented separately for both modes. In the second step, propensity score matching is performed, and results are presented for the combined sample of comparable CATI and web respondents.

5.4.1 Income Nonresponse

From the set of possible covariates described above, we do not include the positioning of the consent question in the item nonresponse model as all respondents who received the question after the experiment did reply to the income question. Also, employment type is not included as "vocational training" predicts item response perfectly. For the same reason, current employment in not included in the model. For all item nonresponse models, the effects of administrative income are shown per € 100.

In the telephone mode, administrative income is found to have a cubic association with item nonresponse which is interacted with gender, age and education (see Table 5.3). The relation between income and item nonresponse heavily depends on the covariates, but none of the effects is significant.

The final item nonresponse model for the web survey also contains a cubic effect for administrative income and interactions with all control variables except for "multiple jobs" (see Table 5.4). Even though all controls and interactions are selected into the final model, only the effect of education is significant.

While we do not find a significant effect of income on item nonresponse in the CATI survey, we find income to affect item nonresponse in the web survey. The effect is moderated by education.

Our analysis does not confirm the often stated hypothesis of a general u-shaped relation between income and item nonresponse. Rather, depending on the socio-demographics, the effect of income on the item nonresponse propensity can have any functional form from u-shaped to monotonic increase or monotonic decrease.

Different socio-demographic variables are found to influence item nonresponse in both modes although none of the effects — except for education in the web mode

	Estimate	Std. Error
(Intercept)	2.9275	3.1304
admin income	-0.4635	0.6211
admin income^2	0.0122	0.0307
admin income^3	-0.0001	0.0004
age	-0.2273	0.1592
age^2	0.0023	0.0018
low education	0.2509	0.9142
middle education	0.4795	0.9416
female	-0.8106	0.7253
multiple jobs	-0.4704	0.3167
admin income*female	0.1051	0.1122
admin income*age	0.0192	0.0304
admin income*age^2	-0.0002	0.0004
admin income*low education	0.0370	0.1368
admin income*middle education	-0.0409	0.1414
admin income^2*female	-0.0036	0.0050
admin income^2*age	-0.0005	0.0015
admin income^2*age^2	0.0001	0.0001
admin income^2*low education	-0.0029	0.0057
admin income^2*middle education	0.0017	0.0061
admin income^3*female	0.0001	0.0001
admin income^3*age	0.0001	0.0001
admin income^3*age^2	-0.0001	0.0001
admin income^3*low education	0.0001	0.0001
admin income^3*middle education	-0.0001	0.0001

Table 5.3: Best model for item nonresponse in the CATI survey after AIC selection.

— is significant. The indicator for multiple jobs is only influential in the CATI mode while UB II recipiency in the past 12 months and German nationality only affect item nonresponse in the web mode. Whether these differences are due to the survey mode or to differences between respondents of the modes will be studied in the second part of the analysis.

5.4.2 Income Measurement Error

Figures 5.4 and 5.5 show the densities of the reported and administrative income for item respondents for both modes. If all income was correctly reported, there should be no difference between the reported and administrative income densities for each mode. Also, in the case of classical measurement error with zero mean, these densities should not differ. In order to facilitate interpretation, the graphs are focused on the range from €0 to €10,000. For both modes, we can see that the general form of

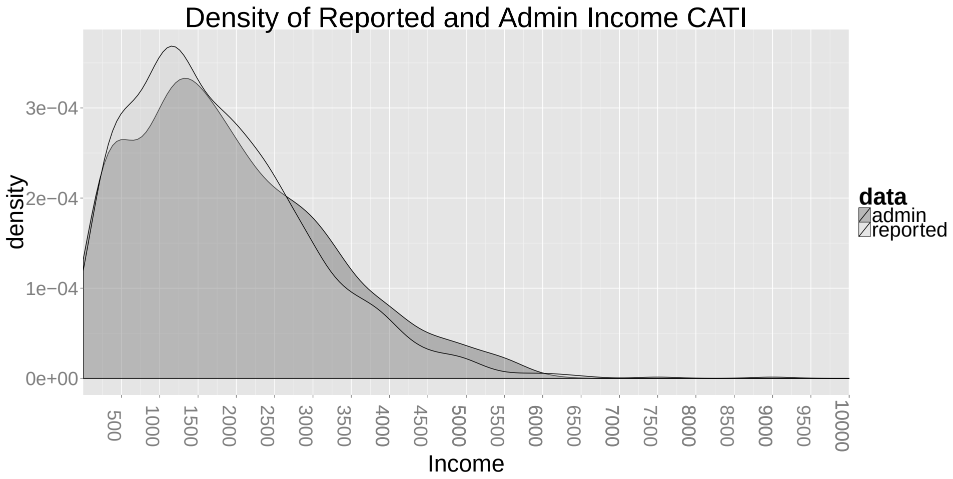


Figure 5.4: Kernel density estimates of reported and administrative income CATI.

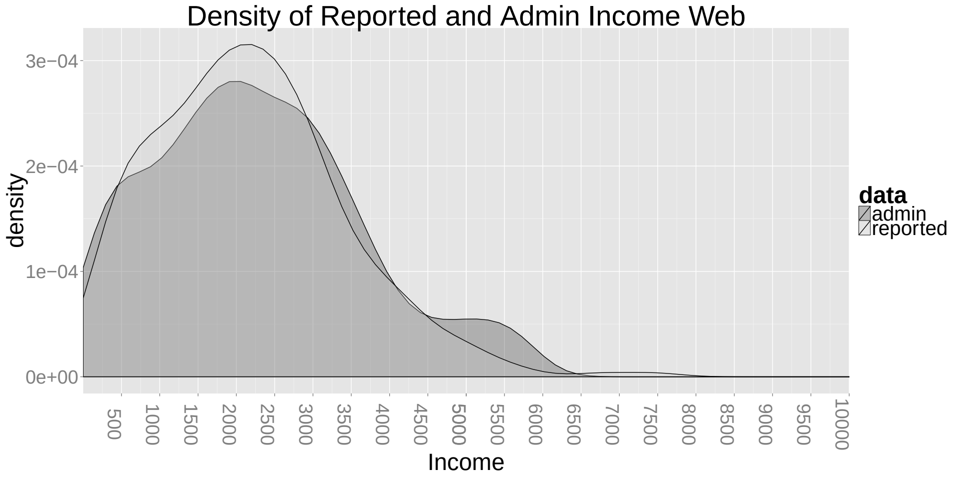
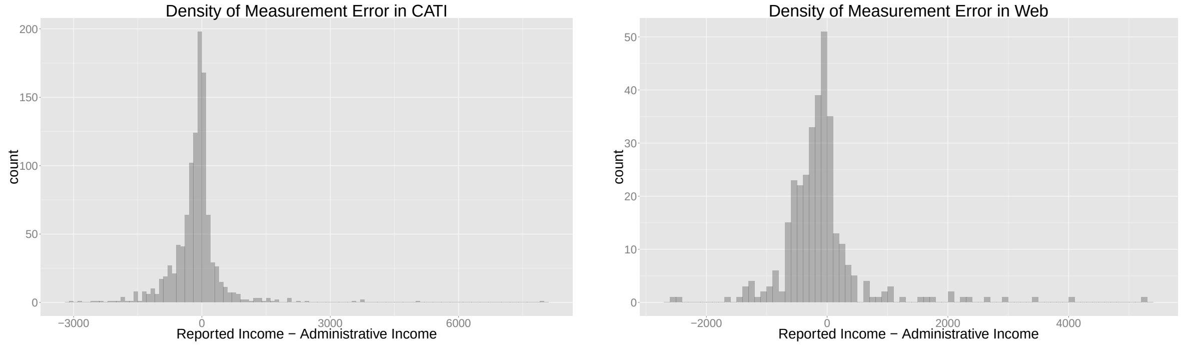


Figure 5.5: Kernel density estimates of reported and administrative income web.

the densities does not differ much between survey and administrative data, but the density of administrative income shows heavier tales: there is more probability mass in the extreme administrative incomes than for the reported income. This can be especially well seen for the web survey. At first glance, the usual finding of mean reversion is confirmed for our surveys.

In a bivariate analysis of the relationship between income measurement error and administrative income, the mean reversion of administrative income is confirmed (see Figures 5.7a and 5.7b). Both modes show a negative relation between measurement error and administrative income, indicating overreporting of low and underreporting of high income.

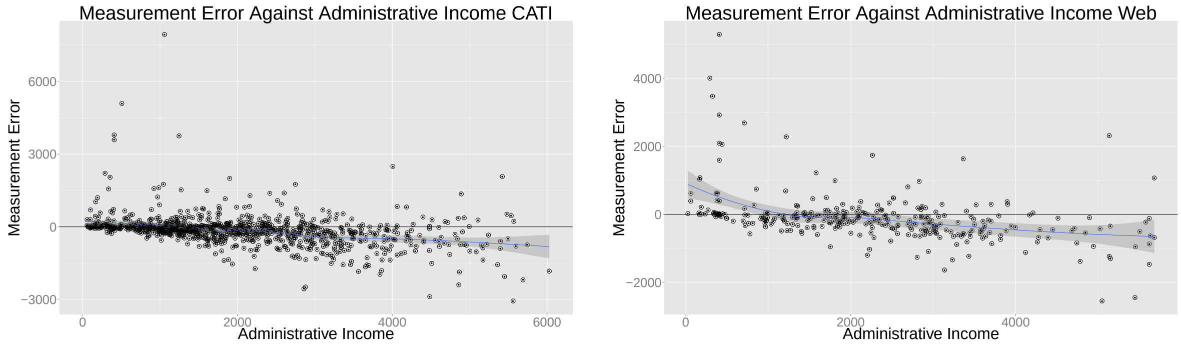
As can be seen from Figures 5.6b and 5.6a, income measurement error for both modes is centered around zero, but is not symmetric. Both densities are left-skewed, indicating more underreporting than overreporting of personal income. Moreover, both densities show heavier tails compared to a normal distribution (see Figures 5.8b



(a) CATI

(b) web

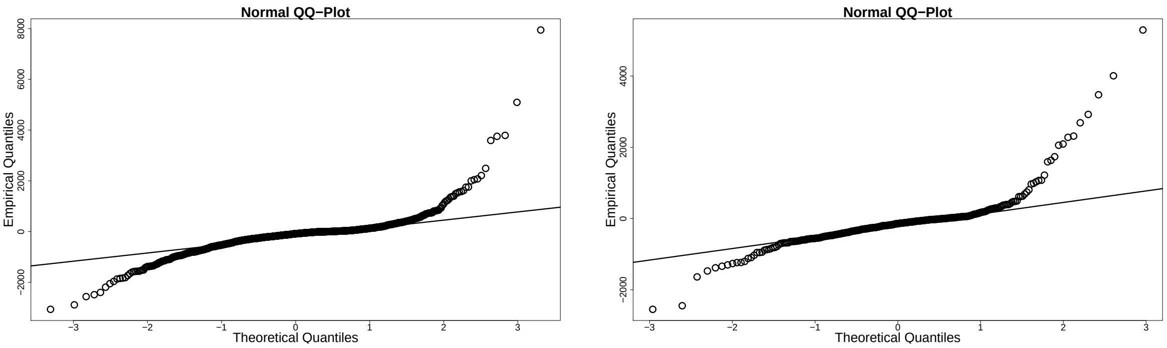
Figure 5.6: Kernel density estimate of measurement error.



(a) CATI

(b) web

Figure 5.7: Scatter plots of measurement error against administrative income, including fits of the nonparametric LOWESS smoother.



(a) CATI

(b) Web

Figure 5.8: Normal QQ-plots for income measurement error

and 5.8a). To conclude, income measurement appears not to be classical for our surveys.

Applying regression models, we find our hypothesis for a mean-reverting effect well confirmed for both modes (see Tables 5.5 and 5.6). As for the item nonresponse analyses, the effects of income are shown per €100. The final models for both modes contain cubic effects of administrative income that indicate a nonlinear negative relationship between income and measurement error. In line with the findings by Kim and Tamborini (2014), we find income to interact with socio-demographic variables which are different for both modes. German nationality, education and the consent placement are only part of the web model.

To summarize, we find a mean-reverting effect of personal income on income measurement error for both modes which interacts with socio-demographics. Different variables are selected for the two modes which can either be due to the modes themselves or a consequence of the different sample composition.

	Estimate	Std. Error
(Intercept)	7.9962	8.9087
admin income	-2.3367	2.9417
admin income^2	0.1747	0.2373
admin income^3	-0.0046	0.0058
age	-0.0813	0.3508
age^2	0.0007	0.0039
female	-0.1759	1.6573
low education	-2.6524	1.5271
middle education	-1.5173	1.4658
german	-3.1650	5.5755
duration	-0.0126	0.0378
duration^2	0.0001	0.0001
UB II	-0.6537	1.7332
admin income*female	-0.0732	0.3225
admin income*age	0.0348	0.0824
admin income*I(age^2)	-0.0004	0.0009
admin income*low education	0.6366	0.3312
admin income*middle education	0.2631	0.3347
admin income*duration	0.0032	0.0099
admin income*duration^2	-0.0001	0.0001
admin income*UB II	-0.3155	0.8152
admin income*german	0.2453	2.2267
admin income^2*female	0.0107	0.0181
admin income^2*age	-0.0017	0.0047
admin income^2*age^2	0.0001	0.0001
admin income^2*low education	-0.0288	0.0169
admin income^2*middle education	-0.0054	0.0188
admin income^2*duration	-0.0007	0.0007
admin income^2*duration^2	0.0001	0.0001
radmin income^2*UB II	0.0445	0.0841
admin income^2*german	-0.0182	0.1991
admin income^3*female	-0.0002	0.0003
admin income^3*age	0.0001	0.0001
admin income^3*age^2	-0.0001	0.0001
admin income^3*low education	0.0004	0.0002
admin income^3*middle education	-0.0001	0.0003
admin income^3*duration	0.0001	0.0001
admin income^3*duration^2	-0.0001	0.0001
admin income^3*UB II	-0.0011	0.0025
admin income^3*german	0.0007	0.0051

Table 5.4: Best model for item nonresponse in the web survey after AIC selection

	Estimate	Std. Error
(Intercept)	-1285.2632	1172.9828
admin income	745.2523	372.0818
admin income^2	-79.4458	38.3482
admin income^3	2.5451	1.2172
female	-45.4743	139.8164
age	20.7060	37.1323
age^2	-0.3019	0.4147
low education	-181.8462	47.6601
middle education	-102.2887	42.1020
duration	9.4685	5.2073
duration^2	-0.0177	0.0101
multiple jobs	-202.4477	224.8641
mini job	144.4528	569.3743
part time employed	97.9364	280.8372
vocational training	-556.7984	534.9677
currently employed	272.3660	713.0267
UB II	-152.5408	164.1647
admin income*female	-37.0095	22.5446
admin income*duration	-3.9383	0.9535
admin income*duration^2	0.0069	0.0018
admin income*mini job	-332.2910	744.4997
admin income*part time employed	-0.6770	44.0060
admin income*vocational training	176.6693	143.0741
admin income*currently employed	-260.2319	322.5938
admin income*age	1.5987	6.9402
admin income*age^2	-0.0131	0.0789
admin income*multiple jobs	-2.8681	34.4497
admin income*UB II	26.5320	41.8437
admin income^2*female	2.4839	1.0136
admin income^2*duration	0.2528	0.0480
admin income^2*duration^2	-0.0004	0.0001
admin income^2*mini job	173.3464	309.1067
admin income^2*part time employed	-1.1904	2.0592
admin income^2*vocational training	-22.0096	12.1033
admin income^2*currently employed	47.0146	37.2301
admin income^2*age	-0.1259	0.3336
admin income^2*age^2	0.0012	0.0038
admin income^2*multiple jobs	0.3256	1.5016
admin income^2*UB II	-1.5998	2.9715
admin income^3*female	-0.0388	0.0130
admin income^3*duration	-0.0040	0.0007
admin income^3*duration^2	0.0001	0.0001
admin income^3*mini job	-23.7891	39.4074
admin income^3*part time employed	0.0268	0.0289
admin income^3*vocational training	0.6610	0.3008
admin income^3*currently employed	-2.0593	1.2109
admin income^3*age	0.0031	0.0044
admin income^3*age^2	-0.0001	0.0001
admin income^3*multiple jobs	-0.0117	0.0189
admin income^3*UB II	0.0158	0.0575

Table 5.5: Best model for measurement error in the CATI survey after AIC selection

	Estimate	Std. Error
(Intercept)	14844.4380	3397.9817
admin income	-1925.5700	603.1857
admin income^2	72.8616	29.2811
admin income^3	-0.8377	0.4093
female	-1403.8850	360.2275
age	-147.9232	88.4484
age^2	1.9442	0.9917
low education	-573.0559	365.3022
middle education	-1536.8686	353.7833
duration	24.0452	10.1689
duration^2	-0.0550	0.0217
multiple jobs	1350.7542	543.0496
consent late	-1124.1311	614.2310
mini job	-1286.1191	1406.5819
part time employed	-1225.3777	861.6029
vocational training	-427.5727	7787.1301
currently employed	-10144.3982	2687.0556
UB II	154.1615	704.4289
german	-1200.5483	840.2121
admin income*female	112.5775	55.5503
admin income*age	24.0827	15.0946
admin income*age^2	-0.3169	0.1751
admin income*low education	85.4285	61.7132
admin income*middle education	193.5686	53.1681
admin income*duration	-3.8499	1.7161
admin income*duration^2	0.0084	0.0036
admin income*multiple jobs	-243.9202	85.8021
admin income*consent late	134.9359	102.6054
admin income*currently employed	1390.3504	488.3450
admin income*german	185.1068	116.3589
admin income*mini job	-630.0925	1218.0657
admin income*part time employed	113.7696	108.2127
admin income*vocational training	-566.2163	2710.1754
admin income*UB II	57.9335	226.2762
admin income^2*female	-3.1725	2.4785
admin income^2*age	-0.9830	0.6981
admin income^2*age^2	0.0129	0.0081
admin income^2*low education	-4.0621	2.7890
admin income^2*middle education	-7.1532	2.3031
admin income^2*duration	0.1578	0.0854
admin income^2*duration^2	-0.0003	0.0002
admin income^2*multiple jobs	11.7697	3.9149
admin income^2*consent late	-5.3460	4.2609
admin income^2*currently employed	-55.5534	24.7062
admin income^2*german	-7.7321	4.8168
admin income^2*mini job	605.8969	469.8818
admin income^2*part time employed	-2.8178	4.2105
admin income^2*vocational training	89.5642	303.9385
admin income^2*UB II	-10.7707	20.4371
admin income^3*female	0.0285	0.0320
admin income^3*age	0.0121	0.0091
admin income^3*age^2	-0.0002	0.0001
admin income^3*low education	0.0480	0.0353
admin income^3*middle education	0.0745	0.0288
admin income^3*duration	-0.0020	0.0013
admin income^3*duration^2	0.0001	0.0001
admin income^3*multiple jobs	-0.1673	0.0511
admin income^3*consent late	0.0624	0.0478
admin income^3*currently employed	0.6832	0.3652
admin income^3*german	0.0889	0.0551
admin income^3*mini job	-90.2666	58.3169
admin income^3*part time employed	0.0134	0.0507
admin income^3*vocational training	-3.3144	10.9492
admin income^3*UB II	0.3828	0.5335

Table 5.6: Best model for measurement error in the web survey after AIC selection

5.4.3 Causal Analyses of Mode Effects Using Propensity Score Matching

Although the sample was randomly assigned to either web or telephone mode, participation in both modes is not random. While the telephone sample is found to show little nonresponse bias (Felderer and Kirchner, 2013), substantive bias can be found in the web mode: Young and well-educated persons, and persons with high income are over-represented in the web survey as compared to the sample frame. Therefore, differences in response behavior between the samples are not necessarily caused by the sample mode, they can also be due to different sample compositions. To be able to causally attribute differences to survey modes, we perform propensity score matching by searching for those CATI respondents who are most similar to the web respondents by means of observed variables. By matching on the covariates, we eliminate or at least reduce the relationship between respondents' characteristics and survey mode (Ho et al., 2007; Rosenbaum and Rubin, 1983).

As all socio-demographic information used in prior models is known to affect the responses to different modes as well as data quality, all these variables and all possible two-way interactions are included in the logistic model which is employed to estimate the propensity scores. The final propensity model was chosen by AIC selection. We perform a one-to-one nearest neighbor propensity score matching, discarding respondents from both modes who are not within the common support of the propensity score. Figure 5.9 shows the estimated propensity scores for participation in the web survey for web and CATI respondents.

In total, 367 respondents from the web survey were matched to 367 respondents from the CATI survey. The 14 respondents from the web and 125 respondents from the CATI survey whose propensity scores were outside the common support were discarded from matching (see Table 5.7).

	CATI	Web
All	1,234	381
Matched	367	367
Unmatched	742	0
Discarded	125	14

Table 5.7: Cases used for matching.

Summary statistics of means and proportions of web and CATI variables before and after propensity score matching can be found in Table 5.8. As can be seen, the similarity of the two samples has been much improved for income, which is the most relevant variable. The samples become very similar for age, UB II, having multiple jobs, being currently employed, gender, and education. For employment type, the

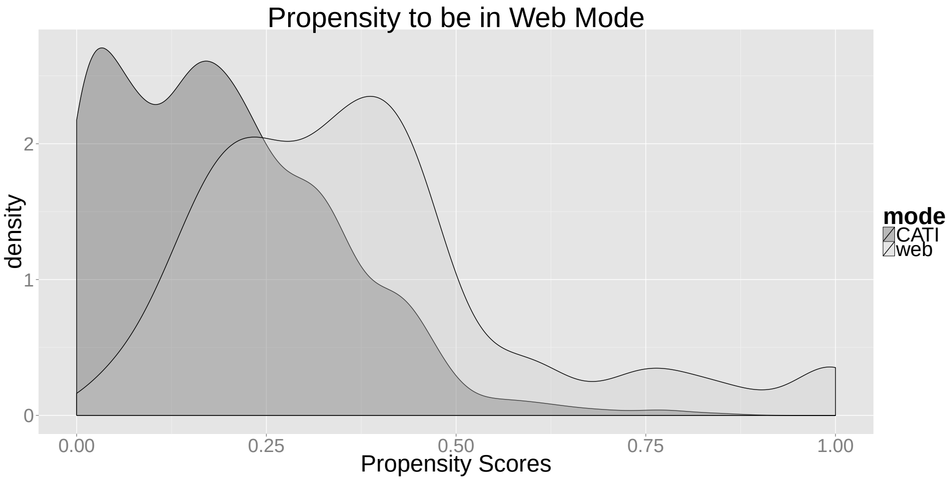


Figure 5.9: Propensities for being in the web mode for CATI and web respondents.

composition improves for some categories while becoming slightly less similar for others. The distributions differ slightly more for spell length.

Variable	samples before matching			matched samples			%
	Web	CATI	Diff.	Web	CATI	Diff.	Improved
Register income	2,230.30	1,971.57	258.72	2,247.68	2,294.53	-46.85	81.89
Age	40.71	42.13	-1.41	40.66	41.16	-0.50	64.76
Female	0.52	0.54	-0.02	0.52	0.54	-0.02	8.45
UB II	0.10	0.19	-0.09	0.10	0.10	0.00	97.04
Multiple jobs	0.09	0.11	-0.03	0.09	0.08	0.01	56.44
Currently employed	0.98	0.99	-0.01	1.00	1.00	0.00	100.00
Low education	0.15	0.30	-0.15	0.14	0.15	-0.01	92.74
Middle education	0.39	0.41	-0.02	0.40	0.39	0.01	46.57
Mini job	0.13	0.12	0.01	0.13	0.13	0.00	100.00
Part time employed	0.24	0.23	0.01	0.24	0.26	-0.02	-59.73
Vocational training	0.04	0.04	-0.00	0.04	0.03	0.01	-1,511.41
Spell length	308.55	316.06	-7.51	319.05	331.37	-12.32	-63.99

Table 5.8: Comparison of sample composition in web and CATI before and after matching.

For the analyses of the matched data, we include the same covariates that have been used before. Moreover, we add an indicator of survey mode as well as all possible interactions of the other covariates with survey mode. The interaction between survey mode and administrative income is of special interest, as this interaction indicates that the effect of income and income measurement error is causally moderated by survey mode. Final models are again found by AIC selection.

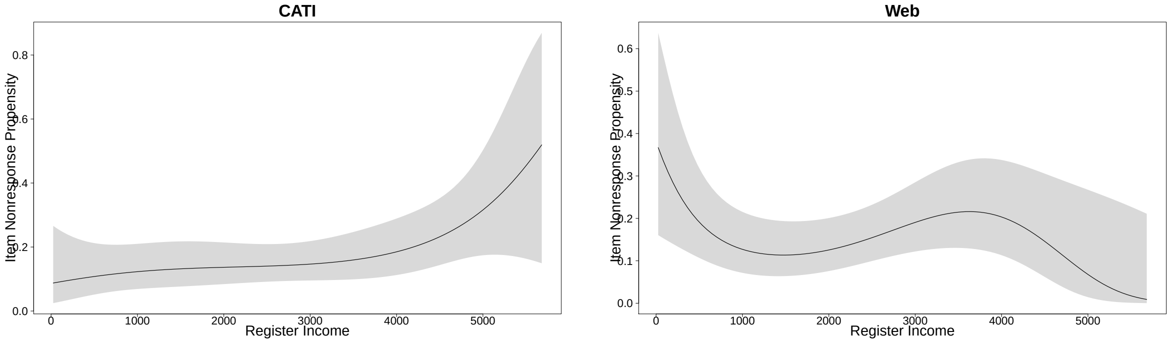
The final model includes a cubic effect of income, quadratic age and spell duration effects, and it contains the indicators for multiple jobs and German nationality. While the effect of spell duration does not differ between the modes, we find interactions

	Estimate	Std. Error
(Intercept)	-14.0854	601.1239
admin income	0.0610	0.0997
admin income^2	-0.0025	0.0043
admin income^3	0.0001	0.0001
age	-0.1087	0.0876
age^2	0.0015	0.0010
duration	-0.0126	0.0086
duration^2	0.0001	0.0001
multiple jobs	-1.9336	1.0808
german	14.5566	601.1209
web	11.2075	601.1279
admin income*web	-0.3074	0.1401
admin income^2*web	0.0143	0.0063
admin income^3*web	-0.0002	0.0001
age:web	0.2549	0.1356
age^2:web	-0.0030	0.0016
multiple jobs:web	2.1220	1.1862
german:web	-14.8054	601.1214

Table 5.9: Best model for item nonresponse in the combined sample, after AIC selection.

of mode and income, age, multiple jobs, and German nationality, though only the income interaction is significant at the 5% level (see Table 5.9). The interaction effect of income and survey mode indicates that there is a different response behavior that can be attributed to the survey mode. Figure 5.10 shows the effects of income on the item nonresponse propensity with all other variables kept at their means and modes, respectively. For the CATI mode, we find a non-linear increase of the item nonresponse propensity as income increases, while for the web mode the nonresponse propensity is decreasing with income. The decrease seems to be non-monotonic for the web survey. However, as this result is based on few observations only, it should be interpreted with caution. Overall, we find the effects to strongly differ by mode, but not in the way we would have expected from theory.

Measurement error is found to be affected by income, gender, age, education, spell duration, having multiple jobs, and past receipt of UB II. Age, spell duration, and multiple jobs are found to have different effects for different modes while the other variables show the same effect in both modes. As there is no interaction of mode and income selected to the model, we reject the hypothesis of differential income effect by mode (see Figure 5.11).



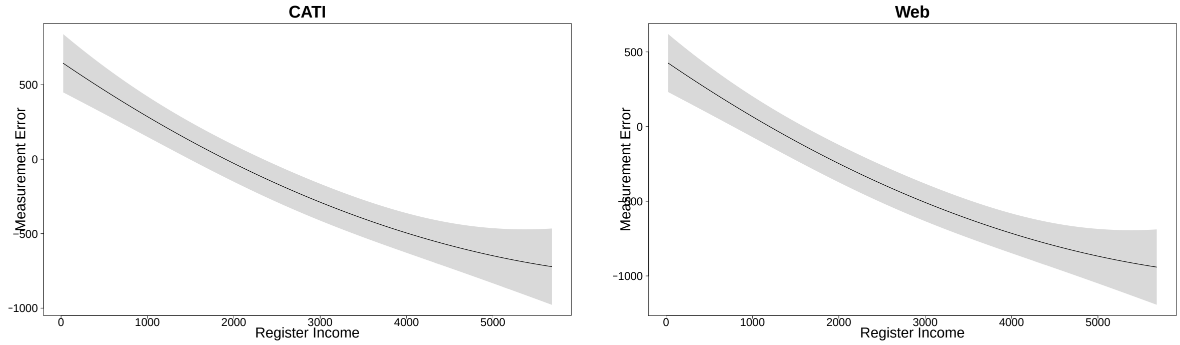
(a) CATI

(b) web

Figure 5.10: Effect of income on item nonresponse for CATI and web.

	Estimate	Std. Error
(Intercept)	375.6733	518.7056
admin income	-39.5014	6.4559
admin income^2	0.2691	0.1178
female	-264.4946	51.6275
age	63.8916	21.7112
age^2	-0.8216	0.2584
low education	-169.2001	79.6823
middle education	-114.2018	52.4539
duration	-6.3031	3.1568
duration^2	0.0120	0.0066
multiple jobs	-410.6823	119.7530
UB II	-147.6381	87.2746
web	1,350.1550	691.9428
age:web	-114.9758	29.4121
age^2:web	1.5908	0.3475
duration:web	5.5594	3.8646
duration^2:web	-0.0116	0.0081
multiple jobs:web	273.2557	169.4345

Table 5.10: Best model for measurement error in combined sample, after AIC selection.



(a) CATI

(b) web

Figure 5.11: Effect of income on measurement error for CATI and web.

5.5 Summary and Conclusion

We find that item nonresponse and measurement error in the income question are related to the true income in both modes. However, the effect of true income on income nonresponse is highly dependent on other socio-demographic characteristics and can have various forms. The often hypothesized general u-shaped effect of personal income on income item nonresponse is not found in our study. The causal analyses of mode on the item nonresponse propensity shows different effects of a person's income on item nonresponse between the modes. Whereas for the CATI mode the item nonresponse propensity increases with personal income, we find a decreasing effect for the web survey.

The finding of mean-reverting income measurement error – the tendency to report an income that is closer to the mean than the true income – which is broadly reported in the literature is replicated for both of our surveys. We do not find causal mode effects on the relationship between income and income measurement error. This, again, was not expected from social desirability theory. We rather expected to find less rounding towards the mean in the web survey. Therefore, social desirability might not be the main reason behind the mean reversion: respondents might just give some rough estimate of their income which can lead to the same pattern. If lack of knowledge is the main reason for measurement error, it makes sense that there are no mode differences. Different issues can affect measurement error for different respondents: whereas high-income respondents might just not know their exact income, low-income respondents should be more aware of what they earn. The correcting of low income towards some mean income can be due to social desirability or fear of disclosure in the CATI survey, while it may rather be exclusively due to fear of disclosure in the web survey. The motivation might be different between the modes but still lead to the same pattern. We note that our data do not allow for deeper analyses. In future studies, respondents could be asked for reasons not to report their income. Moreover, they may be asked whether they have estimated their income, and, potentially, which estimating strategy they have used.

The fact that both item nonresponse and measurement error in the income question are related to true income leads to problems for all disciplines interested in income effects, for instance when studying the returns of education or income inequality. Further analyses of the mechanisms behind item nonresponse and measurement error is needed to develop strategies to gain better data quality.

For the validation of the reported income, we needed to assume income to be stable over an employment spell and bonus payments to be negligible. To make reported and administrative income even more comparable, in a later web survey, "Herausforderungen am deutschen Arbeitsmarkt 2014", we asked the income question in a way that exactly matches the administrative records. That is, we asked not only for yearly income but also for the amount of bonus payments. Validation data for this survey will be available as of the beginning of 2016. These data will help understand whether the same relations found in this Chapter still hold for yearly (as opposed to monthly) income. Finally, they will inform whether the negligible bonus and stable income assumptions hold.

6 Conclusions

The aim of this thesis is to contribute to a deeper understanding of nonresponse and measurement error. Linking survey data and administrative records gives us the great opportunity to study both error sources individually.

One of our main focuses is the effect of *incentives* on these two error sources. For the variables *income* and *having a "mini job"*, we find that, compared to conditional post-paid lottery ticket incentives, unconditional prepaid cash incentives can be used to increase response rates and to decrease nonresponse bias as these measures bring more low-income cases and cases having a mini job into the sample (Chapter 3). Therefore, prepaid unconditional cash incentives are promising for surveys that otherwise *underrepresent* these groups. However, for surveys for which these groups are already well represented or even overrepresented, prepaid cash incentives might lead to a bias or increase an already existing bias. We also find some evidence that additional incentives within a survey can be used to reduce item nonresponse for some questions.

Besides these encouraging results, there are some concerns about monetary incentives that we did not cover in this thesis. One such concern addresses possible long-term consequences of any cash incentive. For instance, if people get used to receiving cash payments for their interviews, these incentives may set expectations, thereby decreasing the willingness to participate of formerly intrinsically motivated respondents when they are not offered such cash incentives. Another concern is that of respondents viewing monetary incentives as a compensation for the time spent on the survey rather than a "token of appreciation". Respondents who think of incentives as such compensation for their opportunity costs might not be willing to participate when receiving no or low cash incentives, and respondents of panel surveys might expect the payment to increase over the waves of the panel.

The economic literature reports some evidence that incentives crowd out intrinsic motivation in economic decision experiments (Gneezy et al., 2011; Gneezy and Rustichini, 2000; Frey and Oberholzer-Gee, 1997). However, this has not been shown for survey participation (Laurie and Lynn, 2009). Moreover, Mercer et al. (2015) find no evidence that the effect of a certain amount of incentive on response rates has declined over the years. All in all, the empirical evidence is limited, and more research is needed to understand the long-term effects of monetary incentives.

In a mode comparison, we find that the relationship between a person's true income and the propensity to report her income is different for the CATI and web mode, respectively. However, misreport in income is found to be mean-reversive for both modes. For our variable under study, there is no preferred data collection mode.

We have studied nonresponse and measurement error individually for this thesis. A possible next step could be to analyze how these two error sources interact, and how they jointly contribute to the total survey error. For the same survey as in Chapter 5, the CATI and web survey "Work and Consumption in Germany", in Felderer and Kirchner (2013) we study the effects of survey mode on nonresponse bias and measurement error both individually and in their joint effect on the sample means. The question in this analysis is whether both error sources reinforce each other, i.e, if their interaction leads to a higher TSE, or if they cancel out each other to some extent, thereby decreasing TSE. In our study, we follow the work of Kreuter et al. (2008) and Sakshaug et al. (2010) who disentangle the error sources and analyze the interaction of both. Whether TSE is dampened or reinforced by the interplay of nonresponse bias and measurement error depends on the direction and relative magnitude of both error sources. If nonresponse bias and measurement error have different signs, they can cancel out each other and lead to survey estimates that are quite accurate. In Felderer and Kirchner (2013), we study nonresponse bias by comparing survey respondents and nonrespondents using administrative data (as in Chapter 3). Measurement error bias is approached by comparing means which are computed based on the respondents' administrative records and means computed based on the respondents' survey data. In contrast to this thesis, we are not linking data on an individual level in that study. We rather draw comparisons on an aggregate level, i.e., we compare means. Thus, individual bias might be averaged out for the means and not be detected in the analyses. Comparing survey modes, we find an overrepresentation of women in the CATI mode, but no significant overrepresentation of women in the web mode. As there is basically no measurement error in gender, combined bias in the estimated proportion of women is only significant for the CATI mode. Similarly, older respondents are overrepresented in the CATI, and younger respondents are overrepresented in the web mode. As there is no measurement error in age, estimated mean age is biased downwards in the web mode, and biased upwards in the CATI mode.

Both surveys overrepresent employed individuals, especially individuals having a regular job (as opposed to having a "minijob"), and overrepresentation is higher in the web than in the CATI mode. Both modes show measurement error in employment type: regular employment is overreported, while having a "mini job" is underreported. These reporting errors are smaller in the web than in the CATI mode. We conjecture that this finding may be explained by more accurate reporting due to less social desirability concerns. Looking at TSE, we find that nonresponse bias is reinforced by measurement error bias for these variables. The comparatively lower bias due to nonresponse in the CATI survey is counteracted by higher measurement error. Because nonresponse bias is higher than measurement error for both sub-surveys, TSE in the CATI mode is still smaller than in the web survey. For income, we basically find no nonresponse bias in the CATI survey, whereas the web survey overrepresents high-income people while underrepresenting low-income people. When comparing the aggregated survey *and* administrative data, we find the same mean-reversion that we found for consenters on an individual level: too few respondents report to belong to the highest income category, while too many respondents claim to be in the low-income group. The high-income respondents are *overrepresented* in the web survey while *underreporting* their income at the same time. Therefore, nonresponse and measurement error counteract each other, leading to a TSE that is smaller than it would be if no measurement error was present. For the CATI survey, we find the opposite: measurement error *increases* the TSE. Even though measurement error is similar to the web mode, the fact that there is no nonresponse bias to be compensated, measurement error induces a bias that would not have been present otherwise.

In another study, Kirchner and Felderer (2015), we find that, despite this substantive total survey error in means within both surveys, regression coefficients are not biased very much. All in all, they capture the direction and relative magnitude of the underlying true parameters quite well, even in the presence of nonresponse bias and measurement error.

Administrative records offer the potential to study even more error sources than we could address in this thesis. For example, coverage error could be studied using register data of the FEA. We know that almost 90% of the workforce are captured by the system, but not all of them can be covered by our survey as the address data and phone numbers may not always be correct and up-to-date. Analyzing the relations between returning invitation letters or outdated phone numbers on the one hand and socio-demographics like education and income on the other hand could help to answer the question of systematic undercoverage of certain subpopulations.

All measures used to increase response rates can in principle affect nonresponse bias. Administrative records give a great opportunity to study this issue. For example, different invitation letters might systematically attract different kinds of people, thereby affecting biases.

Summarizing, administrative data keep many opportunities for future research which may further improve the survey process.

A

German Wording of the Intervention and Questions in the “Work and Consumption in Germany” Surveys

Intervention

No incentive, no request group
Abschließend möchten wir Ihnen noch ein paar Fragen zum Thema “Gesundheit und Lebensstandard” stellen.

Request only group
Abschließend möchten wir Ihnen noch ein paar Fragen zum Thema “Gesundheit und Lebensstandard” stellen. Wir möchten Sie bitten sich bei der Beantwortung dieser Fragen noch einmal **Zeit zu nehmen und in Ruhe nachzudenken**.

Additional incentive groups
Abschließend möchten wir Ihnen noch ein paar Fragen zum Thema "Gesundheit und Lebensstandard" stellen. Wir möchten Sie bitten sich bei der Beantwortung dieser Fragen noch einmal **Zeit zu nehmen und in Ruhe nachzudenken**. Als Dank für Ihre Mühen erhalten Sie von uns **zusätzlich x Euro**.
$x = (0.5, 1.0, 1.5, 2.0, 2.5, 3, 3.5, 4, 4.5)$

Questions Wordings

Health Block
Calories
The question was not taken from any survey.

Was denken Sie, wie viele Kalorien benötigt ein durchschnittlicher Erwachsener Ihres Geschlechts pro Tag?
. . . Kalorien

Weiß nicht
Keine Angabe

Doctor Visits
Questions about doctor visits are for example asked in PASS and SHARE.

Haben Sie im Jahr 2011 einen Arzt aufgesucht? Wenn ja, geben Sie bitte an, wie häufig.

Falls Sie im Jahr 2011 keinen Arzt aufgesucht haben geben Sie bitte 0 ein.
Dabei zählt jeder Arztbesuch, auch zum Abholen eines eigenen Rezeptes.

. . . Arztbesuche

Weiß nicht
Keine Angabe

Diseases
The question is taken from the third wave of the German part of the SHARE survey. The response options are ordered by the frequency they were named in the SHARE survey. The last three options were added for our survey.

Hat Ihnen ein Arzt schon einmal gesagt, dass Sie unter einer der folgenden Krankheiten leiden?

1. Herzinfarkt einschließlich Myokardinfarkt, Koronarthrombose oder andere Herzkrankheitern einschließlich Herzinsuffizienz
2. Bluthochdruck
3. Hohe Cholesterinwerte
4. Schlaganfall einschließlich Durchblutungsstörungen im Gehirn
5. Diabetes oder hohe Blutzuckerwerte
6. Chronische Erkrankungen der Lunge wie chronische Bronchitis oder Lungenemphysem
7. Asthma
8. Arthritis einschließlich Osteoarthritis oder Rheuma

9. Osteoporose
10. Krebs oder bösartiger Tumor, einschließlich Leukämie und Lymphdrüsenkrebs, ausschließlich kleinerer Hautkrebserkrankungen
11. Magengeschwür, Zwölffingerdarmgeschwür
12. Parkinson'sche Krankheit
13. Grauer Star
14. Oberschenkelhalsbruch oder Hüftfraktur
15. Migräne
16. Nahrungsmittelunverträglichkeit
17. Allergie

Keine

Weiß nicht
Keine Angabe

Life Expectancy
The question if taken from the German part of the SHARE survey.
Für wie wahrscheinlich halten Sie es, dass Sie [X] oder mehr Jahre alt werden?
Zur Beantwortung geben Sie bitte eine Zahl zwischen 0 und 100 ein. Sie können dabei alle Zahlen von 0 bis 100 verwenden.
Beispiel:
Wenn Sie keinesfalls glauben, dass Sie [X] oder mehr Jahre alt werden, geben Sie die 0 ein.
Wenn Sie sich ganz sicher sind, dass Sie so alt werden, dann geben Sie die 100 ein.
Mit den Werten dazwischen können Sie Ihre Einschätzung abstufen.

0. . .100 Prozent

Weiß nicht
Keine Angabe

Standard of Living Block
The question is taken from the PASS survey.

Income
Wie hoch war Ihr letztes monatliches Arbeitseinkommen?

Bitte geben Sie Ihr Bruttoeinkommen an, also Ihr Einkommen vor Abzug von Steuern

und Sozialversicherungsbeiträgen.
Wenn Sie im letzten Monat Sonderzahlungen hatten, z.B. Urlaubsgeld oder Nachzahlungen, rechnen Sie diese bitte nicht mit. Entgelt für Überstunden rechnen Sie dagegen mit. Für selbständige Tätigkeiten geben Sie bitte stattdessen Ihren monatlichen Gewinn vor Steuern an.
Wenn Sie momentan erwerbslos sind, geben Sie bitte Ihr letztes Monatseinkommen aus Ihrer letzten Erwerbstätigkeit an.

. . . Euro

War noch nie erwerbstätig
Weiß nicht
Keine Angabe

Unemployment Benefit
The question is taken from the PASS survey. Haben Sie oder Ihr Haushalt in den letzten 12 Monaten zu irgendeinem Zeitpunkt Arbeitslosengeld II bzw. "Hartz 4" bezogen?

Ja
Nein

Weiß nicht
Keine Angabe

Employment History
This question is taken from the PASS survey.
Wenn Sie jetzt einmal an alle Erwerbstätigkeiten denken, die Sie bisher ausgeübt haben, wie lange hat Ihre längste Phase der Erwerbstätigkeit gedauert, die Sie ohne Unterbrechung ausgeübt haben?
Bitte geben Sie die Dauer Ihrer längsten Erwerbstätigkeit an, Arbeitgeber- und Berufswechsel zählen dabei nicht als Unterbrechung.

Dauer in . . . Jahren oder/und
Dauer in . . . Monaten

Weiß nicht
Keine Angabe

Standard of Living
The question is taken from the SHARE survey.
Für wie wahrscheinlich halten Sie es, dass Ihr materieller Lebensstandard in fünf Jahren geringer sein wird als heute?

Zur Beantwortung geben Sie bitte eine Zahl zwischen 0 und 100 ein. Sie können dabei alle Zahlen von 0 bis 100 verwenden.
Beispiel:
Wenn Sie keinesfalls glauben, dass Ihr materieller Lebensstandard in fünf Jahren geringer sein wird als heute, geben Sie die 0 ein.
Wenn Sie sich ganz sicher sind, dass dies passieren wird, dann geben Sie die 100 ein.
Mit den Werten dazwischen können Sie Ihre Einschätzung abstufen.

0. . .100 Prozent

Weiß nicht
Keine Angabe

German Wording of the Income Question

Wie hoch war Ihr letztes monatliches Arbeitseinkommen? Bitte geben Sie Ihr Bruttoeinkommen an, also Ihr Einkommen vor Abzug von Steuern und sozialversicherungsbeiträgen.

Wenn Sie im letzten Monat Sonderzahlungen hatten, z.B. Urlaubsgeld oder Nachzahlungen, rechnen Sie diese bitte nicht mit. Entgelt für Überstunden rechnen Sie dagegen mit. Für selbständige Tätigkeiten geben Sie bitte stattdessen Ihren monatlichen Gewinn vor Steuern an.

B

Additional Tables and Graphs

	Cash Group			Lottery Group		
Variable	CI -	rel bias	CI +	CI -	rel bias	CI +
high income	-1.10	2.25	5.54	1.44	5.34	9.19
middle income	-3.60	-0.29	2.93	-5.18	-1.01	3.06
low income	-5.25	-1.86	1.43	-8.84	-4.55	-0.37
age $>=$ 60	-1.25	3.79	8.53	2.40	8.32	13.97
age 50-59	0.69	3.47	6.25	0.56	3.93	7.39
age 40-49	-0.60	1.82	4.23	-2.31	0.58	3.45
age 30-39	-6.07	-2.73	0.66	-6.49	-2.49	1.39
age $<$ 30	-12.03	-7.56	-3.21	-16.94	-10.82	-4.89
foreign	-6.61	-3.88	-1.20	-5.89	-2.41	1.00
mini job	-4.73	-0.32	3.95	-10.05	-4.60	0.80
employed	-0.91	1.31	3.46	-1.77	0.84	3.32
female	-0.42	1.10	2.63	-2.09	-0.31	1.47
UB II	-3.60	-1.64	0.23	-3.09	-0.53	2.02

Table B.1: Relative nonresponse bias estimates using administrative data. Relative nonresponse bias estimates and 0.025 % and 0.975% quantiles using administrative data.

	Cash Group			Lottery Group		
Variable	CI -	rel bias	CI +	CI -	rel bias	CI +
high income	-1.03	0.27	1.51	-0.58	1.02	2.53
middle income	-1.33	-0.15	1.03	-0.71	0.81	2.30
low income	-1.35	-0.09	1.09	-3.75	-1.95	-0.19
age >= 60	0.98	2.62	4.20	1.38	3.76	5.84
age 50-59	0.75	1.75	2.71	1.19	2.45	3.65
age 40-49	-0.31	0.61	1.49	-1.02	0.01	1.00
age 30-39	-0.80	0.28	1.31	-1.97	-0.63	0.68
age < 30	-8.07	-5.98	-4.16	-8.52	-6.08	-3.85
foreign	-7.51	-4.69	-2.19	-3.13	-0.22	2.24
mini job	-2.14	-0.53	0.95	-3.34	-1.61	-0.00
employed	0.31	1.12	1.96	-0.20	0.74	1.69
female	-0.51	0.00	0.53	-1.22	-0.63	-0.04
UB II	-2.01	-1.34	-0.69	-0.97	-0.16	0.64

Table B.2: Relative nonresponse bias estimates and 0.025 % and 0.975% quantiles using survey data.

			Proportion	Proportion
Variable	Time to Answer		item nonresponse	insufficient answers
	F-test	Kruskal-Wallis Test		
calories	0.456	0.934	0.235	**0.095**
doctor visits	0.985	0.906	**0.001**	0.684
diseases	0.184	0.302	0.28	
life expectancy	0.518	0.772	0.157	**0.098**
income	0.556	0.732	**0.063**	0.421
UB II	0.127	0.652	0.872	
employment	0.79	0.901	0.389	0.793
standard of living	0.938	0.933	0.696	0.767
sum over all questions	0.613	0.951	0.210	0.504

Table B.3: Results for tests of equal means in response times and proportions of item nonresponse between the experimental groups for all questions of the experiment.

Variable	Wrong answer in validation questions		
	F-test	Kruskal-Wallis Test	
income	0.203	0.215	0.313
UB II			
employment	0.685	0.962	

Table B.4: Results for tests of equal proportions of insufficient answers and wrong answers to validation questions between the experimental groups for all questions of the experiment.

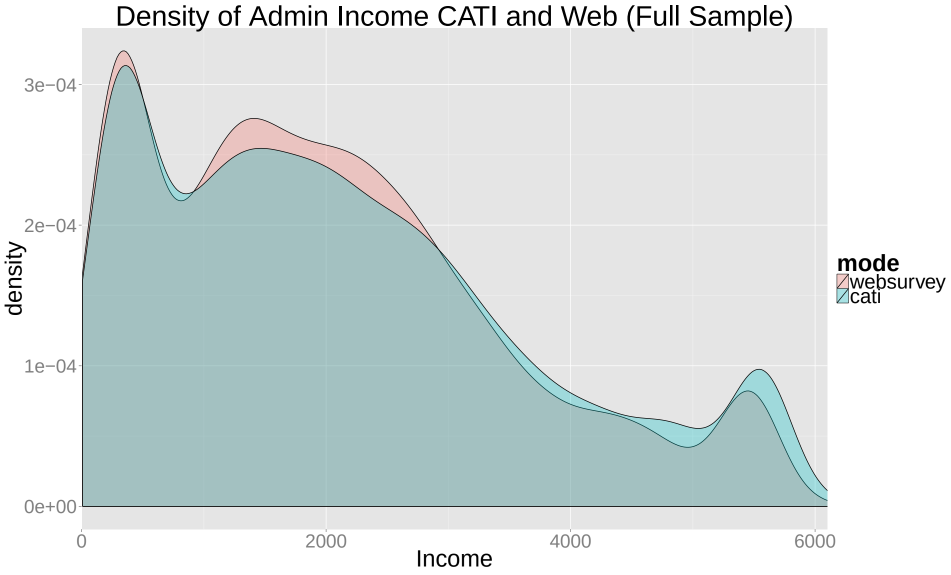


Figure B.1: Kernel density estimates of administrative income for all sample cases and both survey modes. The densities of the administrative income are not different for the telephone and web survey sample as the assignment to the two survey modes was random for administrative income. Differences in the administrative income distributions that are found for the *respondents* of both surveys are only due to differential participation in the surveys.

Bibliography

AAPOR (2015). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. The American Association for Public Opinion Research: AAPOR.

Abowd, J. M. and M. H. Stinson (2013). Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data. *The Review of Economics and Statistics 95*(5), 1451–1467.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Atkeson, L. R., A. N. Adams, and R. M. Alvarez (2014). Nonresponse and Mode Effects in Self–and Interviewer– Administered Surveys. *Political Analysis 22*(3), 304–320.

Azur, M. J., E. A. Stuart, C. Frangakis, and P. J. Leaf (2011). Multiple Imputation by Chained Equations: What is it and How Does it Work? *International Journal of Methods in Psychiatric Research 20*(1), 40–49.

Bentley, J. P. and P. G. Thacker (2004). The Influence of Risk and Monetary Payment on the Research Participation Decision Making Process. *Journal of Medical Ethics 30*, 293–298.

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 Datenreport. Methodische Aspekte zu Arbeitsmarktdaten.

Bethlehem, J. (1988). Reduction of Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics 4*(3), 251–260.

Bethlehem, J., F. Cobben, and B. Schouten (2011). *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons.

Biemer, P., D. Trewin, H. Bergdahl, and L. Japec (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics 30*(3), 381–415.

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly 74*(5), 817–848.

Biemer, P. P. and L. E. Lyberg (2003). *Introduction to Survey Quality*, Volume 335. John Wiley & Sons.

Bollinger, C. R. (1998). Measurement Error in the Current Population Survey: A Nonparametric Look. *Journal of Labor Economics 16(3)*, 576–594.

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

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

Bound, J. and A. B. Krueger (1991). The Extend of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? *Journal of Labor Economics 9*(1), 1–24.

Bradburn, N. M. and S. Sudman (1979). *Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research.* San Francisco CA 94103-1741: Jossey–Bass.

Bricker, J. and G. V. Engelhardt (2008). Measurement Error in Earnings Data in the Health and Retirement Study. *Journal of Economic and Social Measurement 33*(1), 39–61.

Bruine de Bruin, W., P. S. Fischbeck, N. A. Stiber, and B. Fischhoff (2002). What Number is "Fifty–Fifty"?: Redistributing Excessive 50 % Responses in Elicited Probabilities. *Risk Analysis 22*(4), 713–723.

Bruine de Bruin, W., B. Fischhoff, S. G. Millstein, and B. L. Halpern-Felsher (2000). Verbal and Numerical Expressions of Probability: "It's a Fifty–Fifty Chance". *Organizational and Human Decision Processes 81*(1), 115–131.

Büngeler, K., M. Gensicke, J. Hartmann, R. Jäckle, and N. Tschersich (2010). FDZ–Methodenreport 10/2010. IAB–Haushaltspanel im Niedrigeinkommensbereich: Welle 3 (2008/2009). *Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung.*

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

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC Press.

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta–Analysis. *Public Opinion Quarterly 57*(1), 62–79.

Cochran, W. G. (1968). Errors of Measurement in Statistics. *Technometrics 10*(4), 637–666.

Couper, M. P. (2008). *Designing Effective Web Surveys*, Volume 75. Cambridge University Press New York.

Curtin, R., E. Singer, and S. Presser (2007). Incentives in Random Digit Dial Telephone Surveys: A Replication and Extension. *Journal of Official Statistics 23*(1), 91–105.

Duncan, G. J. and D. H. Hill (1985). An Investigation of the Extent and Consequences of Measurement Error in Labor–Economic Survey Data. *Journal of Labor Economics 3*(4), 508–532.

Essig, L. and J. K. Winter (2009). Item Non–Response to Financial Questions in Household Surveys: An Experimental Study of Interviewer and Mode Effects. *Fiscal Studies 30*(3-4), 367–390.

Felderer, B. and A. Kirchner (2013, July). Nonresponse in Probability Web Surveys. A Validation Study. Presented at the 5th conference of the European Survey Research Association, Ljubljana, Slovenia.

Frey, B. S. and F. Oberholzer-Gee (1997). The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding–Out. *The American economic review*, 746–755.

Gneezy, U., S. Meier, and P. Rey-Biel (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives 25*(4), 191–210.

Gneezy, U. and A. Rustichini (2000). Pay Enough or Don't Pay at All. *Quarterly journal of economics*, 791–810.

Göritz, A. S. (2004). The Impact of Material Incentives on Response Quantity,Response Quality, Sample Composition, Survey Outcome and Cost in Online Access Ppanels. *International Journal of Market Research 46*, 327–346.

Göritz, A. S. (2006). Incentives in Web Studies: Methodological Issues and a Review. *International Journal of Internet Science 1*(1), 58–70.

Göritz, A. S. (2008). The Long–Term Effect of Material Incentives on Participation in Online Panels. *Field Methods 20*(3), 211–225.

Gottschalk, P. and M. Huynh (2010). Are Earnings Inequality and Mobility Overstated? The Impact of Nonclassical Measurement Error. *The Review of Economics and Statistics 92*(2), 302–215.

Gouldner, A. W. (1960). The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review 25*(2), 161–178.

Gouret, F. (2011). What can we learn from the fifties? Technical report, Technical report, CAEPS, Universitat de Barcelona.

Grady, C. (2001). Money for Research Participation: Does It Jeopardize Informed Consent? *American Journal of Bioethics 1*(2), 40–44.

Grant, R. W. (2002). The Ethics of Incentives: Historical Origins and Contemporary Understandings. *Econonomics and Philosophy 18*(1), 111–139.

Grant, R. W. and J. Sugarman (2004). Ethics in Human Subjects Research: Do Incentives Matter? *Journal of Medicine and Philosophy 29*(6), 717–738.

Groves, R. M., M. P. Couper, S. Presser, E. Singer, R. Tourangeau, G. P. Acosta, and L. Nelson (2006). Experiments in Producing Noresponse Bias. *Public Opinion Quarterly 70*(5), 720–736.

Groves, R. M., D. A. Dillman, J. L. Eltinge, and R. J. Little (2002). Survey nonresponse. *Survey Nonresponse*, 3–26.

Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology*, Volume 2. John Wiley & Sons.

Groves, R. M. and L. Lyberg (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly 74*(5), 849–879.

Groves, R. M. and E. Peytcheva (2008). The Impact of Nonresponse Rates on Nonresponse Bias. A Meta–Analysis. *Public Opinion Quarterly, 77*(2), 167–189.

Groves, R. M., E. Singer, and A. Corning (2000). Leverage–Saliency Theory of Survey Participation – Description and an Illustration. *Public Opinion Quarterly 64*(3), 299–308.

Halpern, S. D., J. H. T. Karlawish, D. Casarett, J. A. Berlin, and D. A. Asch (2004). Empirical Assessment of Whether Moderate Payments Are Undue or Unjust Inducements for Participation in Clinical Trials. *Archives of Internal Medicine 164*(7), 801–803.

Hansen, R. A. (1980). A Self–Perception Interpretation of the Effect of Monetary and Nonmonetary Incentives on Mail Survey Respondent Behavior. *Journal of Marketing Research 17*(1), 77–83.

Heerwegh, D. (2009). Mode Differences Between Face–to–Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research 21*(1), 111–121.

Heitjan, D. F. and D. B. Rubin (1991). Ignorability and Coarse Data. *The annals of statistics 19*(4), 2244–2253.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis 15*(3), 199–236.

Jäckle, A. and P. Lynn (2008). Respondent Incentives in a Multi–Mode Panel Survey: Cumulative Effects on Nonresponse and Bias. *Survey Methodology 34*(1), 105–117.

Jacobebbinghaus, P. and S. Seth (2007). The German Integrated Employment Biographies Sample IEBS. In *Schmollers Jahrbuch. Zeitschrift für Wirtschafts– und Sozialwissenschaften*, Volume 127, pp. 335–342. Duncker & Humblot, Berlin.

James, J. M. and R. Bolstein (1990). The Effect of Monetary Incentives and Follow–Up Mailings on the Response Rate and Response Quality in Mail Surveys. *Public Opinion Quarterly 54*(3), 346–361.

Jesske, B. and S. Schulz (2012). Methodenbericht Panel Arbeitsmarkt und Soziale Sicherung PASS. 5. Erhebungswelle – 2011. *FDZ Methodenreport*.

Kapteyn, A. and J. Y. Ypma (2007). Measurement Error and Misclassification: A Comparison of Survey and Administrative Data. *Journal of Labor Economics 25(3)*, 513–551.

Kim, C. and C. R. Tamborini (2012). Do Survey Data Estimate Earnings Inequality Correctly? Measurement Errors Among Black and White Male Workers. *Social Forces 90*, 1157–1181.

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. *Sociological Methods & Research 43(1)*, 39–72.

Kirchner, A. and B. Felderer (2015, August). The Effects of Nonresponse Error and Measurement Error on Estimates of Regression Coefficients. Presented at the Joint Statistical Meeting, Seatle, USA.

Koenker, R. and G. Bassett (1982). Tests of Linear Hypotheses and l" 1 Estimation. *Econometrica: Journal of the Econometric Society 50*(6), 1577–1583.

Kreuter, F., S. Presser, and R. Tourangeau (2008). Social Desirability Bias in CATI, IVR, and Web Surveys. The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly 72*(5), 847–865.

Kristensen, N. and N. Westergaard-Nielsen (2007). A Large–Scale Validation Study of Measurement Errors in Longitudinal Survey Data. *Journal of Economic and Social Measurement 32*, 65 – 92.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology 5*(3), 213–236.

Krumpal, I. (2013). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality & Quantity 47*(4), 2025–2047.

Laurie, H. and P. Lynn (2009). The use of respondent incentives on longitudinal surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys*, pp. 205–233. John Wiley & Sons.

Mack, S., V. Huggins, D. Keathley, and M. Sundukchi (1998). Do Monetary Incentives Improve Response Rates in the Survey of Income And Program Participation? In *Proceedings of the Section on Survey Methodology, American Statistical Association*, pp. 529–534.

Manski, C. F. and F. Molinari (2010). Rounding Probabilistic Expectations in Surveys. *Journal of Business and Economic Statistics 28*(2), 219–231.

Medway, R. (2012). *Beyond Response Rates: The Effect of Prepaid Incentives on Measurement Error*. Ph. D. thesis, Graduate School of the University of Maryland.

Mercer, A., A. Caporaso, D. Cantor, and R. Townsend (2015). How Much Gets You How Much? Monetary Incentives and Response Rates in Household Surveys. *Public Opinion Quarterly 79*(1), 105–129.

Miller, P. V. and C. F. Cannell (1982). A Study of Experimental Techniques for Telephone Interviewing. *Public Opinion Qumrteriy 46*(2), 250–269.

Moore, J. C., L. L. Stinson, and E. J. Welniak (2000). Income Measurement Error: A Review. *Journal of Official Statistics 16(4)*, 331–361.

Nordberg, L., U. Rendtel, and E. Basic (2004). Measurement Error of Survey and Register Income. In Ehling and Rendtel (Eds.), *Harmonisation of Panel Surveys and Data Quality*, pp. 65–88. Statistisches Bundesamt, Wiesbaden.

Oberschachtsiek, D., P. Scioch, C. Seysen, and J. Heining (2009). Stichprobe der Integrierten Erwerbsbiografien IEBS. *FDZ Datenreport 3*.

Pedace, R. and N. Bates (2000). Using Administrative Records to Assess Earnings Reporting Error in the Survey of Income and Program Participation. *Journal of Economic and Social Measurement 26*(3), 173–192.

Peytchev, A., S. Riley, J. Rosen, J. Murphy, and M. Lindblad (2010). Reduction of Nonresponse Bias Through Case Prioritization. *Survey Research Methods 4*(1), 21–29.

Philipson, T. (1997). Data Markets and the Production of Surveys. *The Review of Economic Studies 64*(1), 47–72.

Presser, S. and L. Stinson (1998). Data Collection Mode and Social Desirability Bias in Self–Reported Religious Attendance. *American Sociological Review 63*(1), 137–145.

Rässler, S. (2000). Ergänzung fehlender Daten in Umfragen/Imputation of Missing Data in Surveys. *Jahrbücher für Nationalökonomie und Statistik*, 64–94.

Riphahn, R. T. and O. Serfling (2005). Item Non–Response on Income and Wealth Questions. *Empirical Economics 30*(2), 521–538.

Rosenbaum, P. R. and D. B. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika 63*(3), 581–592.

Ruud, P. A., D. Schunk, and J. K. Winter (2014). Uncertainty causes rounding: an experimental study. *Experimental Economics 17*(3), 391–413.

Ryu, E., M. P. Couper, and R. W. Marans (2006). Survey Incentives: Cash vs. In–Kind; Face–to–Face vs. Mail; Response Rate vs. Nonresponse Error. *International Journal of Public Opinion Research 18*(1), 89–106.

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

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

Sakshaug, J. W. and F. Kreuter (2014). The Effect of Benefit Wording on Consent to Link Survey and Administrative Records in a Web Survey. *Public Opinion Quarterly 78*(1), 166–176.

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

Sakshaug, J. W., T. Yan, and R. Tourangeau (2010). Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi– Mode Survey of Sensitive and Non– Sensitive Items. *Public Opinion Quarterly 74*(5), 907–933.

Schräpler, J.-P. (2004). Respondent Behavior in Panel Studies: A Case Study for Income Nonresponse by Means of the German Socio–Economic Panel (GSOEP). *Sociological Methods & Research 33*(1), 118–156.

Schräpler, J.-P. (2006). Explaining Income Nonresponse – A Case Study by Means of the British Household Panel Study (BHPS). *Quality and Quantity 40*(6), 1013–1036.

Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.), *Survey Nonresponse*, pp. 163–177. John Wiley & Sons New York.

Singer, E. and M. P. Couper (2008). Do Incentives Exert Undue Influence on Survey Participation? Experimental Evidence. *Journal of Empirical Research on Human Research Ethics 3*(3), 49–56.

Singer, E., R. M. Groves, and A. D. Corning (1999). Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation. *Public Opinion Quarterly 63*(2), 251–260.

Singer, E., J. van Hoewyk, N. Gebler, T. Raghunathan, and K. McGonagle (1999). The Effect of Incentives on Response Rates in Interviewer–Mediated Surveys. *Journal of Official Statistics 15*(2), 217–230.

Singer, E. and C. Ye (2013). The Use and Effects of Incentives in Surveys. *The Annals of the American Academy of Political and Social Science 645*(1), 112–141.

Smith, T. W. (1991). An Analysis of Missing Income Information on the General Social Surveys. *Chicago: National Opinion Research Center, University of Chicago, GSS Methodological Report 71.*

Smyth, J. D., D. A. Dillman, L. M. Chistian, and M. Mcbride (2009). Open–Ended Questions in Web Surveys. Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality? *Public Opinion Quarterly 73*(2), 325–337.

Toepoel, V. (2012). Effects of Incentives in Surveys. In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences*, pp. 209–223.

Tourangeau, R. (1984). Cognitive Science and Survey Methods. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, 73–100.

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

Tourangeau, R. and T. W. Smith (1996). Asking Sensitive Questions. The Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly 60*(2), 275–304.

Tourangeau, R. and T. Yan (2007). Sensitive Questions in Surveys. *Psychological Bulletin 133*(5), 859–883.

Trappmann, M., B. Christoph, J. Achatz, C. Wenzig, G. Müller, and D. Gebhardt (2009). Design and Stratification of PASS: A New Panel Study on Research on Long Term Unemployment. *IAB Discussion Paper*.

Watson, N. and M. Wooden (2009). Identifying Factors Affecting Longitudinal Survey Response. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys*, Chapter 10, pp. 157–182. John Wiley & Sons.

Wertheimer, A. and F. G. Miller (2008). Payment for Research Participation: A Coercive Offer? *Journal of Medical Ethics 34*, 389–392.

Willimack, D. K., H. Schuman, B.-E. Pennell, and J. M. Lepkowski (1995). Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face–to–Face Survey. *Public Opinion Quarterly 59*(1), 78–92.

Eidesstattliche Versicherung

gemäß der Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5

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

München, 25. September 2015

Barbara Ingrid Maria Felderer