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

IAB-Bibliothek 348
Die Buchreihe des Instituts für Arbeitsmarkt- und Berufsforschung

Techniques for Asking Sensitive Questions in Labor Market Surveys

Antje Kirchner

Dissertationen
wbv

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

IAB

IAB·Bibliothek 348

Die Buchreihe des Instituts für Arbeitsmarkt- und Berufsforschung

Techniques for Asking Sensitive Questions in Labor Market Surveys

Antje Kirchner

Dissertationen

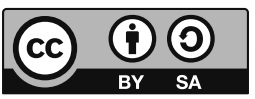
wbv

Bibliografische Information der Deutschen Nationalbibliothek

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

Dissertation an der Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München – eingereicht am 30.07.2013

Erstgutachterin: Prof. Dr. Frauke Kreuter
Zweitgutachter: Prof. Dr. Thomas Hinz
Verteidigung: 15.11.2013



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

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

ISBN 978-3-7639-4083-7 (Print)
ISBN 978-3-7639-4084-4 (E-Book)
Best.-Nr. 300849 www.iabshop.de www.iab.de

Preface: Scope of this Work

Acknowledgements and Project Idea

The initial idea to this project originated during a lunch break at the European Survey Research Association conference in 2009 by Mark Trappmann, head of the research department 'Panel Labour Market and Social Security' (PASS) at the Institute for Employment Research (IAB), and Ivar Krumpal, University of Leipzig.

The rationale of the project was to contribute to the improvement of data quality for information that is relevant to the PASS study, and known to be difficult to collect in labor market surveys due to respondents' incentives to misreport. Engagement in undeclared work and receipt of basic income support, a form of means-tested social security payment (more precisely, unemployment benefits II) are such information. It can be reasonably assumed that the receipt of welfare benefits and undeclared work might be perceived as socially undesirable, potentially stigmatizing or as behaving antisocially by those respondents possessing this sensitive trait or who are engaged in this behavior. Undeclared work even being pursued by the authorities if disclosed. Thus, it comes as no surprise that this behavior is often denied or underreported in population surveys and that estimates derived from these survey reports will be inaccurate.

The main idea of the project was to assess the use of special techniques to collect information on these sensitive topics in large-scale labor market surveys in Germany, for potential future implementation in the PASS, which is a low-income panel study in Germany focusing on dynamics of welfare benefit receipt and poverty (Trappmann et al., 2010). In order not to jeopardize the PASS study itself, an evaluation of these 'dejeopardizing techniques' (Lee, 1993)—the most prominent example is the randomized response technique (RRT; Warner, 1965; Fox and Tracy, 1986)—within the PASS study was not possible.

Thus, Mark and Ivar initiated a project called 'Measuring Undeclared Work in Labor Market Surveys' which I joined in February 2010, finalizing the project proposal for funding at IAB. Due to the relevance for PASS, some limitations with respect to the choice of the interview mode or the sample selection were pre-imposed.

While the main idea is from Mark and Ivar, we jointly worked out the exact design and details of the project. I was by and large responsible for the management of the entire data collection process. My particular contribution involved: 1) managing the request for proposal (RfP), 2) sampling and preparation of the frame data (including data confidentiality), 3) designing the experiments and the questionnaire, 4) training of the interviewers, 5) supervising fieldwork, 6)

cleaning and preparing data for the analysis. Furthermore, I conducted a large part of the data analysis and contributed to the drafting of the manuscripts.

During the course of the project, Ivar acquired further funding as a part of the project "Asking Sensitive Questions: Possibilities and Limits of Randomized Response and Other Techniques in Different Survey Modes" (Leipzig VO 684/11) funded by the German Research Foundation within the Priority Programme on Survey Methodology (1292). This allowed us to initiate a second project, 'Measuring Undeclared Work in Labor Market Surveys' with the same goals. However, this project aimed at evaluating another prominent dejeopardizing technique: the so-called the item count technique (ICT; Droitcour et al., 1991; Biemer et al., 2005). Hagen von Hermanni joined this second project as student assistant at the University of Leipzig and was responsible for a large share of the operational activities in Leipzig.

Due to the cooperations within the projects, parts of the work presented here are based on publications that were prepared in jointly with co-authors. For reasons of internal consistency, the entire dissertation is hence written in the 'we' form, even if chapters are single authored by myself. Parts of the ideas of the co-authored papers were also included in the introductory Chapter 1, as well as the concluding Chapter 4.

I presented all of the papers that form the basis of the individual dissertation chapters at different conferences and received valuable comments (such as Kirchner et al., 2011, 2012; Kirchner, 2013). Furthermore, for one of the joint papers (Kirchner et al., 2013), I was responsible for an initial draft which was then revised jointly, the analyses as well as the revise and resubmit. This paper on which Chapter 2 is based, is meanwhile published in *Zeitschrift für Soziologie*, while the paper that forms the basis of Chapter 3 is published in the *Journal of Survey Statistics and Methodology*. Ben Jann, University of Bern, joined the project team for this latter paper. Aside from his contribution to the preparation of the manuscript, he derived the likelihood functions and contributed the Stata program for the analyses.

Table 1 provides an overview of the cooperations:

Table 1: Cooperations, co-authors and publication type

Chapter	Based on Manuscript by	Publication Type
2	Kirchner, Krumpal, Trappmann, von Hermanni (2013), 'Measuring and Explaining Undeclared Work in Germany'	Published
3	Trappmann, Krumpal, Kirchner, Jann (2014), 'Item Sum: A Novel Technique for Asking Continuous Sensitive Questions'	Published
4	Kirchner (2013), 'Validating Sensitive Questions: A Comparison of Survey and Register Data'	Under Review

I really enjoyed working with Mark Trappmann, Ivar Krumpal, Hagen von Hermanni and Ben Jann on this project and learnt a lot about how to conduct survey research from all of them. I want to give Mark and Ivar a special thanks for letting me join this project and for letting me use the results for my dissertation. Further, this research would not have been possible without the support of the two funders, the Institute for Employment Research and the German Research Foundation.

Also, without this project, I would have not had the opportunity to work with Frauke Kreuter, who has been a wonderful supervisor of my thesis and mentor over the past three years. I also want to thank my second supervisor, Thomas Hinz, who gave me many opportunities to present the project ideas and results in his research colloquium in Konstanz. I am especially thankful to the members of the so-called FKRG (Frauke Kreuter Research Group): Stephanie Eckman, Barbara Felderer, Julie Korbmacher, Joe Sakshaug, Jennifer Sinibaldi and Frauke Kreuter herself. I really value our regular meetings, giving me room for discussion, questions, much feedback and most of all, a great education on how to conduct survey research.

I received valuable comments and encouragement from many other people in preparing this dissertation. I hope the few who I will leave out, will accept my apologies. I learnt a lot during my meetings with Helmut Küchenhoff and Thomas Augustin. My colleagues at IAB, especially the PASS group and the KEM group, always supported me, both substantively as well as with cookies. I am particularly grateful to my friends Inna Becher, Gerrit Müller, Hagen von Hermanni and Mihaela Anastasiade for their constant encouragement and patience with me. Last but not least, I'd like to express my gratitude to my (extended) family!

Contribution

This dissertation focuses on the main results of the projects outlined above, henceforth referred to as the RRT study (commissioned by the IAB) and the ICT study (commissioned by the University of Leipzig). All empirical analyses are based on data collected in these two German population surveys, conducted on the telephone in 2010. In order to evaluate the different techniques, the experimental set-ups and designs of both surveys—the RRT study (n = 3,211) and the ICT study (n = 1,603)—were kept as similar as possible.

Furthermore, the RRT study incorporated two experiments that were not replicated in the ICT study (due to budget and data confidentiality constraints). We developed and experimentally tested a method to collect information on continuous sensitive information (quantity of undeclared work): the item sum technique (IST). Additionally, we included an experiment to allow an in-depth analysis of the RRT, using administrative records for validation of survey reports

(relaxing the so-called 'more-is-better' assumption, see Chapter 1). Table 2 provides an overview of the studies, the topics and the experiments used in this dissertations.

Table 2: Overview of experiments within each study

Experiments	RRT Study (IAB)	ICT Study (Leipzig)	Chapter	Experiments
Undeclared Work:				
Estimates of Prevalence	✔	✔	2	RRT & ICT
Substantive Analyses	✔		2	RRT
Estimates of Quantity (Hours and Income)	✔		3	IST
Unemployment Benefits:				
Validation	✔		4	RRT & Frame Data
Substantive Analyses	✔		4	RRT & Frame Data

This dissertation contributes to the existing literature in several ways: *First*, while the RRT has been applied to various topics, to our knowledge, no large scale population survey (in Germany) exists that applies and evaluates this special technique to collect data on undeclared work or welfare benefit receipt (Boockmann et al., 2010).[1] Given the overall promising evidence regarding the RRT (see Chapter 1.3.3), the first contribution is rather methodological in nature and aims to evaluate the performance of the RRT in the context of labor market surveys.

Second, we are also not aware of any methodological studies evaluating the prominent, yet more recently developed, ICT (Tourangeau and Yan, 2007) in the context of labor market surveys on undeclared work. Given the promising evidence (see Chapter 1.3.4), the functioning of the ICT is investigated compared to direct questioning.

The *third* contribution within the RRT study is of a substantive nature. Using two samples with differing incentive and opportunity structures, we investigate individual motivations to engage in undeclared work—more precisely, the utility, cost, opportunity and norm hypothesis—and provide insights into who underreports

1 The two existing studies investigating these topics were conducted by Van der Heijden et al. (2000) and Lensvelt-Mulders et al. (2006). The studies assess fraud in welfare and unemployment benefits as well as social security and disability insurance fraud in the Netherlands using the RRT. The experiments in the first study were conducted in three selected Dutch cities, allowing only limited inference. While the second study included one item capturing 'undeclared work' and dealt with social security fraud, the focus of the study was not 'undeclared work' per se. Furthermore, the study covered only a very specific part of the general population, namely individuals who had been eligible for disability benefits.

welfare benefit receipt. These analyses were conducted using logistic regression analyses (Jann, 2011).

Fourth, we developed the so-called item sum technique (IST) to collect data for continuous sensitive characteristics. Although RRT schemes tailored to continuous sensitive characteristics have been proposed in the literature (cf. Himmelfarb and Edgell, 1980; Eichhorn and Hayre, 1983; Gjestvang and Singh, 2007), there is only little evidence on how these techniques perform in practice. The majority of RRT applications are typically used for the collection of dichotomous sensitive variables. Due to its complexity, it can be expected, that an RRT scheme for continuous variables is even more difficult to implement than standard—binary—RRT and imposes an additional cognitive burden on the respondents. Contrary to that, the ICT is less complex in implementation and imposes less cognitive burden on the respondent. We are not aware of any implementation beyond the collection of binary data for the ICT. Thus, another contribution of our study is a generalization of the ICT for the collection of data on quantitative sensitive variables by means of the IST. We applied our novel technique for gathering information on earnings from and hours engaged in undeclared work. We evaluate whether the IST outperforms standard direct questioning and for whom this technique works particularly well. We further demonstrate how IST data can be analyzed and what a potential survey design should look like for a successful implementation.

Fifth, our particular RRT study design provides a unique opportunity to investigate the so-called 'more-is-better' assumption, which is so often relied upon in the literature (Lensvelt-Mulders et al., 2005a). The opportunity to validate the RRT is very rare in the literature on sensitive questions (Lensvelt-Mulders et al., 2005a; Wolter, 2012). We combine information from administrative records and survey data on welfare receipt, and thus know the true percentage of respondents who have received transfer payments, and hence the percentage of people who should have reported receipt. This permits us to validate the reported percentage against the known true rate for the responding cases and to analyze underreporting in depth (also at an individual level).

Structure of the Dissertation

The outline of the dissertation closely mirrors the goals stated above.

Chapter 1 provides an outline of the major problems associated with the collection of data on undeclared work. It gives insights into how respondents approach the task of answering survey questions, and the problem of measurement error in the presence of sensitive questions (Tourangeau and Rasinski, 1988; Groves, 1991;

Tourangeau et al., 2000; Groves et al., 2009). Furthermore, the main ideas of the randomized response technique and the item count technique will be introduced—providing examples of the particular variants we have implemented—along with their estimators. Chapter 1 concludes with an overview of the research questions and an outline of the main dissertation chapters (Chapter 2 to Chapter 4).

Chapter 2, 'Measuring and Explaining Undeclared Work in Germany,' aims at evaluating two research questions. The first research question deals with whether we can obtain valid estimates of undeclared work using the RRT or the ICT (given the 'more-is-better' assumption). The second research question addresses the issue of which factors contribute to the explanation of undeclared work (using individual-level survey data).

Chapter 3, 'Item Sum: A Novel Technique for Asking Continuous Sensitive Questions,' is dedicated to the development of our new technique. The main research question we address in this chapter, is whether the IST can outperform standard direct questioning. For this evaluation, we collected data on hours engaged in undeclared work and earnings from undeclared work. Solely for the purpose of this dissertation, the other research question is to examine whether the IST performs differently in different subgroups, depending on differential cognitive abilities.

The final substantive Chapter 4, 'Validating Sensitive Questions: A Comparison of Survey and Register Data,' analyzes underreporting of welfare benefit receipt in surveys, relaxing the 'more-is-better assumption.' Relying on administrative data, we know the 'true' answer of the respondent and compare response bias in the RRT and direct questioning condition. More specifically, this chapter addresses two research questions: First, whether item specific response bias in surveys can be reduced by means of the randomized response technique with respect to a) the true value in the administrative data and b) direct questioning (DQ) in the survey data. And second, which subgroups are especially affected by response error.

This dissertation ends with an overall conclusion in Chapter 5, summarizing the main results, discussing the limitations of the studies, and implications for future research.

Contents

List of Tables

List of Figures

1 Introduction

According to recent publications by the *Institut für Angewandte Wirtschaftsforschung (IAW)*, the size of the shadow economy in Germany is at an all time low since the 1990s (IAW, 2013). Many of these studies estimating the size of the shadow economy rely on macroeconomic approaches. While these approaches are often useful, their estimates depend heavily on the underlying assumptions. Over the past years, it has thus been much debated which of the different approaches yields the most accurate estimates (Thomas, 1999; Thießen, 2011). To give one example of the resulting discrepancies: Macro-level estimates regarding the magnitude of shadow economic activities range from 14.7% to 16.3% of the gross domestic product in 2000/01, while micro-level approaches estimate the magnitude at 1.3% for the same period (Feld et al., 2007, p. 7). The fact that shadow economic activities in general, or undeclared work in particular, remain unobserved in most instances (Koch, 2005) does not provide a satisfactory explanation for this observed discrepancy. Further, drawing inferences based on macroeconomic approaches often does not allow for a precise distinction of the different aspects of shadow economy. Aside from undeclared work, shadow economy, for example, also comprises criminal activities such as the trade of illegal goods and outputs. Thus, in order to obtain an estimate of the magnitude of undeclared work, additional approaches at the microeconomic level are commonly used (Merz and Wolff, 1993; Mummert and Schneider, 2001; Pedersen, 2003; Feld and Larsen, 2005, 2008; EC – European Commission, 2007).

Approaches at the macro-level are very inclusive as well as non-reactive and often lead to upwardly biased prevalence estimates, while approaches at the micro-level yield downwardly biased estimates (Schneider, 2003; Breusch, 2005; Koch, 2005; Feld et al., 2007; Pickard and Sardà, 2011). Given that undeclared work is a socially undesirable behavior and will be pursued by the authorities if disclosed, it comes as no surprise that this behavior is often denied or underreported in population surveys (Feld and Larsen, 2008). Although both approaches provide prevalence estimates for undeclared work, survey data at the micro-level have one major advantage: They allow an investigation of relationships at the individual level, and thus individual motivations for engaging in undeclared work. In order to model such relationships accurately, however, it is necessary that such data do not contain any systematic error, such as underreporting (Hausman, 2001).

To combat misreporting on sensitive topics, survey designers have developed various data collection strategies aiming at eliciting more truthful and accurate answers from respondents by increasing the anonymity of the question-and-

answer process. The main idea of these 'dejeopardizing techniques' (Lee, 1993) is to increase respondent anonymity and reduce respondents' concerns about honestly reporting undeclared work in a survey. The most prominent examples of such techniques (Lee, 1993) are the randomized response technique (RRT; Warner, 1965; Fox and Tracy, 1986) and the item count technique (ICT; Droitcour et al., 1991; Biemer et al., 2005).

Up to date, no population survey in Germany exists that particularly addresses the problem of social desirability associated with asking questions concerning undeclared work. Thus, the main focus of this dissertation is to analyze and discuss the potential of special data collection strategies in the context of undeclared work (Boockmann et al., 2010, p. 100). More precisely, we assess whether we can reduce underreporting of undeclared work by means of RRT or ICT and thus obtain more accurate, i.e., 'increased,' prevalence estimates compared to a direct questioning approach. Our empirical analyses are based on data collected in two German population surveys (see Chapter 2.1).

The goal of this chapter is to outline existing research gaps. We will first discuss the different concepts and definitions of 'shadow economy' and 'undeclared work' in Section 1.1. Section 1.2 will provide an overview of how these concepts can be measured using macro-level approaches, while Section 1.3 describes individual-level approaches—particularly focusing on survey-based measures—and the challenges associated with collecting data on undeclared work in surveys. This chapter concludes with an overview of the identified research gaps and the research questions of the dissertation (Section 1.4).

1.1 Concepts: 'Shadow Economy' and 'Undeclared Work'

According to the dual economy approach, the economy can be split into an official and an unofficial sector (Schneider and Enste, 2007). The official sector comprises all economic activities, that are included in the gross national product. The underground economy is that part of the economy that is not included in the gross national product. It subsumes those activities that contribute to the added economic value in a country, but are hidden from the authorities.[1] Despite attempts by the EU to standardize what is to be included in the national accounts (thus implicitly defining the underground economy), definitions of national accounts differ internationally. Also, due to the fact that the underground economy comprises "numerous [hidden] economic activities, it is dicult to provide

1 The terms unofficial economy and underground economy or sector will be used interchangeably throughout the text.

a formal definition" (Schneider and Enste, 2007, p. 6). For a detailed overview see Schneider and Enste (2007) or Schneider and Enste (2000).

Similarly to the concepts provided by the Organisation for Economic Co-operation and Development (OECD, 2002, p. 37 ff.), Schneider and Enste (2007, p. 11) distinguish four different sectors within the underground economy: the household sector, the informal sector, the irregular sector and the criminal sector. While the household and the informal sector belong to the so-called 'self-sufficient economy,' the latter two sectors belong to the so-called 'shadow economy.' The main distinction between both categories is the criterion of legality. This distinction according to the criterion of legality is merely one possibility for differentiating sectors within the underground economy. The form of transaction itself is also often considered (Boockmann et al., 2010, p. 14): While some activities, such as neighborhood help or voluntary work, as well as trade of (stolen) goods are typically nonmonetary transactions, other activities such as undeclared work can be of a purely monetary nature.

'Do-it-yourself' activities (Buehn et al., 2009) as one example for the *household sector*, as well as neighborhood help or voluntary activities as one example for the *informal sector*, are by definition legal. They provide an added economic value but cannot be captured in the national accounts, because they are not recorded anywhere and the precise value is usually not defined. Shadow economic activities, on the other hand, are by definition illegal and comprise the irregular and the criminal sector. The *irregular sector* is equivalent to what is commonly understood by undeclared work, or shadow economy in the narrower sense. Again, an added economic value is created, that is not captured in the national accounts because it is concealed from the authorities, for tax or regulatory burdens. This is essentially equivalent to what Evers (1987, p. 353) refers to as the informal sector. It is characterized by a (comparatively) small-scale market-based production and distribution resulting in goods and services that are in principle (economically) legal, but illegally hidden activities. Last but not least, the *criminal sector*, according to Schneider and Enste (2007), includes illegal activities such as trade with stolen goods and drugs, fraud or smuggling. The difference compared to undeclared work, is that both the production itself as well as the produced goods and services are illegal.

Depending on the exact research question and research method, the exact definition and distinction of the concepts of shadow economy or undeclared work vary. Even though some authors seem to use the same terminology throughout their work, the underlying concepts sometimes vary and warrant caution in hasty comparisons (Janisch and Brümmerhoff, 2004). Scientists, therefore, often rely on one broad, commonly used definition (and operationalization) of shadow economy that is based on statistical criteria: It comprises all value added production,

economically legal or illegal, that is not captured in the national accounts (Schneider and Enste, 2007; Boockmann et al., 2010). Thus, it is not part of the estimation of the gross domestic product (GDP). Undeclared work, i.e., shadow economy in the narrow definition, would then pertain to all market-based, illegally hidden productions of goods and services that are in principle legal.

These statistical or fiscal approaches are less theoretically but rather instrumentally motivated. They also seem to provide a direct means to estimate the magnitude of undeclared work in Euro, hours and hourly wages at the macro-level—and thus provide a pragmatic basis for a definition. However, since the gross national product implicitly captures parts of shadow economic activities, these approaches are often misleading (OECD – Organisation for Economic Co-operation and Development, 2002; Janisch and Brümmerhoff, 2004, p. 13).

The lack of one precise definition, the imprecise use of the terminology, as well as the problematic differentiation of the four sectors, render international comparisons very difficult and result in diverging estimates (see Thomas, 1999; Schneider et al., 2002; Pedersen, 2003; Renooy et al., 2004; Koch, 2005; Enste and Schneider, 2006b). To give one more example: While neighborhood help is part of the household sector in Germany, and not part of the shadow economy, it is illegal in Denmark. Thus it belongs to the irregular sector (Boockmann et al., 2010, p. 14). These legislative and fiscal differences often make it impossible to differentiate or compare self-sufficient and shadow economic production at the macro-level, since these concepts vary internationally (Jessen et al., 1988; Pedersen, 2003).

The focus of the following dissertation is exclusively on undeclared work in Germany, i.e., shadow economic activities in the narrow sense. The definition of undeclared work in Germany is fairly straightforward: It is based on an act that was passed in 2004. This "Law to intensify the fight against black activities and accompanying tax evasion" (Feld et al., 2007, p. 1) provides the legal basis to combat undeclared work, tax evasion and illegal employment (SchwarzArbG, 2004). Unlike other countries that often only subsume tax evasion and social security fraud under undeclared work, the German framework is more inclusive (Boockmann et al., 2010).

Violations regarding handicraft regulations as well as abuse of public benefits, due to (motivated) underreporting of income—even if the employment itself is reported to the social security authorities—, are within the scope of undeclared work (§1 Sec. 1 SchwarzArbG, 2004; Schneider and Enste, 2007, p. 11). The German law also includes a negative definition of undeclared work: Those activities that are not meant to generate sustainable profits and are only marginally paid, are not within the scope of undeclared work (§1 Sec. 3

SchwarzArbG, 2004). Thus all neighborhood help or voluntary activities that are based on courtesy, i.e., according to the definition above belong to the 'self-sufficient economy,' are excluded.

To summarize: A simple, unified definition of undeclared work, that captures all legal aspects, that allows a distinction with respect to the 'self-sufficient economy' and that is understood by all respondents in a survey, is a challenge. Thus we define undeclared work as any productive activity that

- generates labor income,
- is economically legal in the sense that it could be transferred to the official sector,
- is concealed from the authorities, either to evade taxes and/or social security contributions, to undermine legal regulations, such as health standards or minimum wages, or administrative burden (Boockmann et al., 2010, p. 15; see also Renooy et al., 2004; EC – European Commission, 2007).

1.2 Measurement of Undeclared Work: Macro-level Approaches

Similar to the discourse regarding the definition of the concepts of shadow economy and undeclared work, there is much disagreement on how to measure these concepts in order to obtain valid estimates of the magnitude and developments over time (Thomas, 1999; Pedersen, 2003; Koch, 2005; Enste and Schneider, 2006b; Thießen, 2011). In general, there are two broad classes of measurements: *indirect (macro-level)* and *direct methods (individual-level)* (Pedersen, 2003; Feld et al., 2007; Schneider and Enste, 2007). As already indicated, these different methods result in very heterogeneous estimates.

Indirect methods are mostly based on macroeconomic indicators. They comprise a variety of methods, such as discrepancy approaches, monetary approaches, physical input approaches and hidden variable approaches (Boockmann et al., 2010, p. 63; Renooy et al., 2004; Weiß, 2008). Each method has its strengths and weaknesses with respect to capturing certain shadow economic activities. Further, each method relies on different assumptions, which is why these methods are sometimes also referred to as 'indicator' approaches (Schneider and Enste, 2007). These methods serve as a point of reference only and are not main focus of this work. They demonstrate why we rely on survey data for our analyses. We will thus only briefly summarize the main ideas and assumptions. The interested reader is referred to Feige (1990) or Schneider and Enste (2000) for more detailed information.

Discrepancy Approaches: Discrepancy approaches make use of the fact that there are multiple ways to measure the same construct in an economy. The basic

assumption is that that incentives to underreport (labor) income outweigh incentives to underreport expenditures (Boockmann et al., 2010, p. 63).

To provide one example, based on the national accounts definition: The expenditure measure of the gross national product should be equal to its income measure (Janisch and Brümmerhoff, 2004). Using this approach, differences that are found between the two measures are then attributed to shadow economic activities, i.e., people consuming more than they officially should have money for (Feld et al., 2007). Thus, this income gap approach provides an estimate for the magnitude and development of shadow economic activities. However, the expenditure-side components typically suffer from measurement error due to (c)omissions. Purely statistical discrepancies, due to problems with data quality are assumed to be stable over time. Therefore, "these estimates may be crude and not very reliable" (Schneider and Enste, 2007, p. 17).

This method, however, allows the derivation of estimates by industrial sector, and can easily be extended to a lower level by comparing household incomes with household expenditures (microeconomic approach) or differences between the official and the actual income rates (see Section 1.3).

Other examples of the discrepancy approaches include a comparison of the official and the actual labor force (Enste and Schneider, 2006a). It is typically assumed that the labor-force participation is constant and—all other things being equal—a decline in labor participation in the official economy is interpreted as an indicator for an increase in undeclared work. This approach, however, neglects moonlighting, which is understood as engaging in undeclared work while having an 'official' job.

Monetary Approaches: Monetary approaches, on the other hand are based on the assumption that shadow economic activities are settled by paying in cash in order to remain untraceable. If—all other things being equal—either the ratio of cash flow and economic power changes, or the demand for cash rises beyond a level that could be explained by economic growth, a change or rather an increase in the size of shadow economic activities can be inferred.

One of the earliest techniques within this framework is based on a comparison of the ratio of currency and demand deposits at time t_1, with a ratio in a base year t_0 (Gutmann, 1977). The stock of money is comprised of the components currency and demand deposits. It is assumed that there was no shadow economic activity in the base year t_0. This difference can then be used to estimate the growth of currency that is held for illegal purposes, i.e., the size of the shadow economy.

To obtain a more accurate measure of illegal activities, one can use the transaction approach developed by Feige (1979). This technique essentially compares the 'normal' and the 'actual' development of the cash demand, i.e., the total nominal GNP and the official GNP. It makes use of Fisher's quantity equation (Fisher, 1922), again assuming a base year with no shadow economy, as the reference. Further assumptions are made with respect to the velocity of money and the (constant) relationship of the transactions (Schneider and Enste, 2000; Feld et al., 2007). The transaction approach is among the most frequently used techniques, as is the currency demand approach (Schneider et al., 2002, p. 27).

The currency demand approach (Kirchgässner, 1983; Tanzi, 1983; Langfeldt, 1984; Karmann, 1990; Schneider and Enste, 2000) simulates currency demand with and without tax variables. These variables include, for example, interest rates and per capita income, as well as other indicators relating to undeclared work (burden of taxation, regulatory burden, or tax morale). Two estimates of currency demand are then derived, and, accounting for the velocity of money, the discrepancy of both informs about the magnitude of the shadow economy (Boockmann et al., 2010, p. 81). This approach is already a transition to more recently developed model-based approaches, that are also partly based on the currency demand approach.

Physical Input Approaches: The electricity consumption approach belongs to the class of physical input methods. It uses electricity as indicator of economic activity. More precisely, the Kaufmann-Kaliberda method (1996) assumes "electric-power consumption to be the best physical indicator of overall economic activity" (Schneider and Enste, 2007, p. 22). Assuming an elasticity of one between GDP and electricity—and again, a base year without shadow economy—we can then obtain an estimate of the GDP. Subtracting the official GDP from this estimate provides an estimate of the magnitude of shadow economic activities (Schneider and Enste, 2000; Schneider et al., 2002).

A similar approach by Lackó (1998, 1999) models undeclared work as well as do-it-yourself activities primarily as a function of household electricity consumption. Since it is unknown how kilowatt-hours translate into shadow economic production, this relationship has to be estimated as well.

Model-based Approaches: The previously described approaches consider for the most part only single indicators, such as electricity, labor supply or currency demand, for estimating the size of the shadow economy. Since none of these indicators can fully capture the phenomenon, model-based approaches,

> also known as hidden variable approaches, were developed (Frey and Weck-Hannemann, 1984; Schneider, 2007). In these approaches, shadow economy is understood as a latent or unobserved variable in a measurement model that is based on multiple causes and multiple observable indicators.
>
> In the '(dynamic) multiple-indicators and multiple-causes' ([DY]MIMIC) approach, the unknown coefficients are derived from statistical theory and estimated in a set of structural equations (Breusch, 2005; Schneider, 2007). Macroeconomic variables, such as real GDP growth, monetary indicators, such as currency demand, and labor market indicators, such as working hours, are among the most frequently used indicators for the shadow economy and are then linked to causal variables such as tax burden, regulation density and many more (Schneider, 2007, 2008; Boockmann et al., 2010; Thießen, 2011; Buehn, 2012). Which indicators and causes are included in the model is often determined by data availability for the units of analysis in any given period (Boockmann et al., 2010, p. 90).

There is a clear trend towards using currency demand or the (DY)MIMIC approaches, which more and more replace discrepancy, transaction- and physical input approaches. This is the result of various developments: The significance of indicators in order to explain shadow economic activities used in these earlier approaches is decreasing, results derived with these methods are implausible, data availability with respect to other indicators is improving, and last but not least, more sophisticated statistical models have been developed (Schneider, 2009).

Turning to the empirical results of two of these methods for Germany, the currency demand as well as model-based approaches estimate the size of the shadow economy in 2005 at 15.5% to 16.0% of the GDP (Enste and Schneider, 2006b, 188). This is approximately equivalent to 340 to 350 billion Euro. According to the authors, this overall estimate is composed of (and the following estimates are derived from survey data): 3.1% undeclared work, 3.1% to 4.6% as a "correction factor" for underreporting in surveys; 3.0% to 4.0% material input; 4.3% to 4.8% of illegal or criminal activities, as well as an additional 1.0% to 2.0% of activities that are already captured implicitly in the GDP. The estimates provided by Feld and Schneider (2010, p. 125) for the year 2006 are essentially identical to those of 2005, with one exception, that 'undeclared work' and the 'correction factor' are combined.

This is a nice demonstration that both shadow economy and undeclared work, in popular parlance often used interchangeably, cannot be used as equivalent to one another. What is also obvious from these estimates, is that undeclared work—including the correction factor—accounts for the largest part—approximately

40%—of all shadow economic activities in Germany (Enste and Schneider, 2008, p. 113). Contradicting these findings, according to recent results from a study conducted by Thießen (2011, p. 200), this share of undeclared work is supposed to be much lower, amounting only to 1.3% of the GDP. This finding is particularly interesting from an economic policy perspective: The share of labor that could in principle be legalized using suitable reforms is much smaller than assumed.

To summarize the main limitations of indirect methods:

- According to Williams (2010, p. 251), results derived with these indirect approaches should be treated with "utmost caution" due to the results being largely determined by the underlying assumptions, which shall not be discussed any further at this point (see Thomas, 1999; Pedersen, 2003; Koch, 2005, 2008; Boockmann et al., 2010). In his study, Thießen (2011), demonstrates how much estimates differ depending on the specific approach and the assumptions used, explaining the tremendous variation in the observed results (see also Koch, 2005).
- These approaches often do not allow a clear cut differentiation between the concepts or the different components of shadow economic activities. Feld et al. (2007) note, for example, that the main difference between the last three classes of approaches discussed above and the first one, i.e., the income gap method, is substantial. While the first approach captures shadow economic activities including pure tax evasion, i.e., an inclusion of undeclared income from capital, the latter approaches fail to account for this.
- The most serious limitation, however, is one that all of these approaches have in common: They only provide aggregate estimates for the magnitude of the shadow economy and undeclared work in a given economy (mostly expressed as the share of the GDP). If scientific research is interested in inquiring individual incentives to engage in undeclared work, these methods do not provide any insights.

1.3 Measurement of Undeclared Work: Individual-level Approaches

So-called direct methods, on the other hand, can help to gather insights at the individual level. Furthermore, these methods allow a distinction between various shadow economic activities, such as undeclared work (the irregular sector) and other illegal activities (criminal sector). Again, several measurement approaches are subsumed within direct methods (Schneider and Enste, 2007): the use of administrative records, tax auditing, as well as the use of population surveys.

The first individual-level approach derives prevalence estimates based on administrative data (e.g., by the German 'finance control unit' (*Finanzkontrolle Schwarzarbeit*, FKS)). Conducting analyses based on these records—or using these as a sampling frame to conduct surveys—essentially leads to highly selective samples that are not representative for the general population (Feld et al., 2007, p. 8). Using administrative records, information is available only for those individuals who have been charged with engaging in undeclared work, which would thus "notoriously underestimate the size of hidden populations and the extent of deviant activities" (Lee, 1993, p. 45).

Compliance or discrepancy methods, such as tax auditing, typically suffer from the same shortcomings regarding selectiveness by auditing only tax payers and commonly a non-random sample of these taxpayers. Furthermore, they rely on a comparison of actual earnings reported in an audit and income declared for tax purposes. Using these methods, however, it remains unclear if respondents will report their income from undeclared work as 'actual earnings' or not (Pedersen, 2003, p. 21). Only if that is the case are the differences meaningful and can be interpreted.

In order to obtain an estimate of the magnitude of undeclared work representative of the general population, survey-based measures provide a third means of measurement within these individual-level approaches. They are widely used and are designed to measure a variety of shadow economic activities (Pedersen, 2003; EC – European Commission, 2007).[2] Aside from conceptual problems mentioned above, another particular challenge of surveys concerning undeclared work (as one such underground activity) is the sensitivity of the topic. Since we are dealing with a prototype of sensitive questions—this behavior is not only socially undesirable, but also illegal—it is plausible to assume that respondents will either underreport or conceal their engagement. In the worst scenario, respondents are offended and might drop out of the survey and refuse further participation (Tourangeau and Smith, 1996, p. 276). Also, if cooperation is systematically related to the variable of interest, i.e., only those individuals who do not engage in undeclared work decide to participate in the survey, surveys will not provide reliable estimates of the size of undeclared work or individual motivations.

Systematic survey reports, i.e., underreporting or item nonresponse due to social desirability concerns, result in a lower validity of the prevalence estimates ('social desirability bias,' see Tourangeau and Yan, 2007). Especially if respondents have

2 Though population or sample surveys only capture some components of shadow economic activities: Typically surveys do not account for material input potentially used for undeclared work. Neither do they assess undeclared work from companies for companies, but rather only undeclared work by 'private' individuals (which might be identical to a company).

confidentiality concerns or are worried about their anonymity, potential social and legal consequences might foster systematic answering strategies (Fox and Tracy, 1986). The more threatening or sensitive a survey question is perceived to be by a respondent, the more likely it is that these strategies dominate (Bradburn et al., 2004; Lensvelt-Mulders et al., 2005a).

The question arises of how to approach these concerns in order to obtain honest responses and a valid measurement of undeclared work in general population surveys.

1.3.1 The Response Process and Sensitive Questions

The goal of each survey is to obtain accurate data and estimates. This means estimates that are reliable and valid, i.e., "stable over replications and close in value to the true value of a statistic" (Groves, 1989, p. 16, see also Biemer and Lyberg, 2003). The actual accuracy of a survey estimate, however, can be jeopardized at various stages of the survey process. Aside from errors of nonobservation (e.g., coverage error, sampling error, nonresponse error), measurement error is an important source of error especially when dealing with sensitive questions and concerns of social desirability or 'threat' posed by a question.

In general, nonsampling errors can be be categorized into a systematic component (bias) and a random component (variance) that affect survey error (Biemer, 2010). Observations are said to be biased if the expected value of the errors over response and sampling distributions are unequal to zero ($E(\varepsilon_i) \neq 0$), e.g., if there is a systematic tendency to misreport. Following the notation in Groves (1991) or in Biemer and Lyberg (2003, p. 40), "nonsampling errors are systematic if $E(\varepsilon_i) = B$ where $B \neq 0$." The observed response can thus be modeled as a combination of systematic and random errors, for respondent *i*. The standard framework models the observed response as a combination of the respondent's true score (X_i), the respondents' bias (B)—an overall tendency across respondents to misreport—and an individual random error component (e_i) (slightly adapted from Tourangeau et al. (2000, p. 266)):

$$Y_i = X_i + B + e_i$$

The following chapters focus particularly on the systematic component of measurement error, more precisely underreporting, when collecting data on sensitive topics such as undeclared work. In general, measurement error can arise from various sources (Groves, 1989): the interviewer, the respondent, the questionnaire and the mode of data collection. When dealing with sensitive questions, the respondent as one main 'source of error' is of particular importance. In order to understand how respondents approach the task of answering a sensitive

question and the potentially resulting inaccuracies (Lee, 1993), it seems advisable to briefly describe the response process in general.

One prominent model for the response process is suggested by Tourangeau (1984, see also Tourangeau and Rasinski, 1988): Respondents first assess the question content (comprehension); second, they identify and recall information relevant to answering the question (retrieval). On the third stage of this model, respondents evaluate their 'potential answer' (judgement) before eventually reporting it to the interviewer (response). Along any of these stages, multiple errors may occur. While most of the problems associated with a survey response, such as forgetting or other memory problems, are not particular to sensitive questions, incentives to intentionally misreport due to the 'threatening' nature of the question, on the other hand, are a great concern for the researcher. More generally, it is often assumed that traits that are positively valued—such as voting behavior, donations or energy conservation—are overreported, while undesirable traits—such as illicit drug use, abortion, social fraud, or plagiarism—will be underreported (Groves, 1989; Lee, 1993; Bradburn et al., 2004).

There is a wide range of topics referring to undesirable traits, taboos, illegal activities or unsocial opinions that can be considered sensitive in a survey interview (see also Tourangeau et al., 2000; Kreuter et al., 2008; Krumpal, 2013). But what exactly makes a topic 'sensitive' for a respondent (Lee, 1993; Bradburn et al., 2004; Tourangeau and Yan, 2007)?

Using a theoretical approach, the level of 'threat' or 'sensitivity' of a question as perceived by the respondent can be established along three dimensions (Tourangeau and Yan, 2007). First, the question content can be perceived as intrusive, e.g., it can be about a taboo topic or involves something that the respondent is ashamed of. Second, respondents might be concerned about a risk of disclosure, i.e., admitting to a certain behavior might entail legal or social sanctions, such as the presence of parents during an interview when surveying teenagers about their drinking behavior. Third, social desirability concerns might be triggered when a respondent is asked to admit to having violated a social norm or possess socially undesirable traits, i.e., to avoid social embarrassment and to project a positive self-image (Groves, 1989). There are also empirical approaches to establish whether 'social desirability' or generally 'sensitivity' is a problem: Rating-scales, such as asking the respondents or other coders (e.g., experts) to evaluate the social desirability are often used, as are empirical indicators post-hoc assessing sensitivity by analyzing the amount of item nonresponse, response times, or misreporting (Lee, 1993; Groves et al., 2004; Tourangeau and Yan, 2007; Krumpal, 2013).

However, one has to keep in mind, that social desirability is subjective. It is heavily influenced by characteristics of the item under study and it is typically only an issue if a respondent actually has the sensitive trait or behavior (Groves, 1989; Kreuter et al.,

2008). Furthermore, approaches assessing the amount of item nonresponse might go wrong, since sometimes refusing to answer a survey question might reveal more for the respondent than 'simple' misreporting (Groves et al., 2004, p. 240).

Survey methodologists have suggested a range of guidelines to combat measurement error from (item) nonresponse and misreporting ('under' as well as 'overreporting') due to the sensitive nature of a question (for an extensive overview see Barton, 1958; Lee, 1993; Bradburn et al., 2004; Tourangeau and Yan, 2007; Groves et al., 2009; Krumpal, 2013).[3] Aside from confidentiality assurances (Singer et al., 1995), particular care is to be put in the *wording* and the question format. Strategies to make respondents more at ease when responding to sensitive questions include so-called forgiving wording or loading, e.g., depicting the sensitive characteristic or behavior as something normal; extending the reference period, e.g., asking about past behavior first and then moving to the more recent period; embedding the question, e.g., starting with a gentle, general introduction to the topic and then moving to specific behaviors or traits; paraphrasing whenever possible and avoiding the use of the sensitive terminology; or providing the respondent with a response scale that suggests that the behavior is 'normal,' thus implicitly suggesting a different distribution. However, Bradburn et al. (2004, p. 80) argue, that as questions become more threatening to a respondent, respondents "are more likely to overstate or understate behavior, even when the best question wording is used." Furthermore, recent studies demonstrate that differences in wording of sensitive questions have no significant effect on individual responses and cannot reduce or eliminate social desirability bias (Krumpal and Näher, 2012).

Another strategy involves the administration of the questions themselves, i.e., to guarantee a *private data collection setting*. Usually, the main distinction when studying social desirability effects among different survey modes is whether the questions are interviewer- or self-administered (Groves et al., 2004; Schnell, 2012). Results suggest that the use of a less intrusive interview mode, such as self-administration, can substantially improve reporting of sensitive information and reporting accuracy (Tourangeau et al., 2000; Tourangeau and Yan, 2007; Kreuter et al., 2008). Conflicting evidence on self-administered surveys is presented in a recent study by Lelkes et al. (2012). The authors provide evidence that while reports under completely anonymous conditions sometimes increased reports of socially desirable attributes, accuracy was consistently reduced and survey satisficing increased. Also, whether or not other people are present during the interview changes perceived privacy.

3 The following list of strategies is not comprehensive and only provides examples. Tactics such as the 'bogus pipeline' will not be further discussed (Tourangeau and Yan, 2007).

Last but not least, researchers use so-called *dejeopardizing methods* to elicit sensitive information in surveys indirectly. The randomized response technique (RRT; Warner, 1965; Fox and Tracy, 1986) and the item count technique (ICT; Droitcour et al., 1991; Biemer et al., 2005) are the most prominent and most frequently used of these indirect surveying techniques. They aim to reduce respondents' concerns about honestly reporting sensitive information (from respondents who have the sensitive trait) by increasing the anonymity of the question-and-answer process. The main idea is that nothing regarding the 'true' status (mostly of binary nature, i.e., 'Yes' or 'No') of a respondent can be immediately inferred from their answer to the survey question.

1.3.2 Measurement of Undeclared Work: Selected German Surveys

As outlined in the previous section, several strategies have been introduced to counter systematic misreporting for socially (un)desirable topics (Barton, 1958; Lee, 1993; Tourangeau et al., 2000; Bradburn et al., 2004; Tourangeau and Yan, 2007; Krumpal, 2013). Before introducing our experiments in greater detail, we will provide a brief overview of the strategies used in German surveys on undeclared work and/or shadow economy and the estimates obtained in these studies (for a comprehensive overview see Boockmann et al., 2010).

Mode of data collection: Most studies on undeclared work (in Germany) are conducted in a face-to-face (f2f) mode (Pedersen, 2003; Feld and Larsen, 2005; EC – European Commission, 2007; Feld and Larsen, 2008; Feld and Schneider, 2010), while only few surveys are implemented in a telephone mode (Mummert and Schneider, 2001). Feld and Schneider (2010, p. 112) specifically argue, and Pedersen (2003) demonstrates, that the f2f mode proved more successful in their study compared to a telephone setting, as it provided higher estimates of undeclared work. The latter study had originally been designed as a telephone survey. The low prevalence estimates in the pretest conducted in a telephone mode, however, led the researchers to conclude that there is tremendous bias, which caused the mode switch to f2f. Mogensen et al. (1995, as cited in Pedersen, 1998, p. 169) report similar difficulties, as well as high drop out rates and interviewer effects for their telephone survey in Denmark. A German survey of people age 18 and older conducted in 1997 by Lamnek et al. (2000, p. 76, 135) combines a f2f mode with a self-completion section for questions on undeclared work. However, given increasing survey costs, it is worthwhile to assess surveying methods that are viable in the telephone mode.

Use of definitions: The studies conducted by the Rockwool Foundation (Pedersen, 2003; Feld and Larsen, 2005) generally provide respondents with a broad definition of undeclared work[4], while other studies do not provide respondents with any definition at all (Schneider et al., 2002; Schneider, 2008). The Eurobarometer 2007 uses the same—rather inclusive—definition in all countries (EC – European Commission, 2007).[5] To decide not to provide any definition or to use broad definitions that do not fully match the national context can cause confusion among respondents. Furthermore, these diverging practices lead to results that are not comparable across and sometimes even within countries.

Operationalizations and Question wording: Only one German study explicitly asks respondents if they engage in the 'shadow economy' (Schneider, 2008, p. 94), as opposed to other studies that typically focus on certain components of shadow economy (such as engagement in undeclared work, consumption of goods and services provided undeclared, etc.). All of these studies approach the problem of social desirability and disclosure differently (for a comprehensive overview of strategies in general see Lee, 1993; Bradburn et al., 2004, p. 81). While some studies try to decrease perceived item sensitivity by extending the reference period from 'in the previous 12 months' to 'have you ever...' (Lamnek et al., 2000), other studies avoid the use of the sensitive phrase altogether. These studies paraphrase and refer to undeclared activities instead of 'illicit' or 'black labor' which would be a more literal translation for the exact German term 'Schwarzarbeit' (EC – European Commission, 2007). Another strategy, for example, is to "embed" the question (Lee, 1993, p. 78): The Eurobarometer, for example, initially asks respondents about undeclared work and attitudes in 'general,' before moving to behavior 'specific' to the respondent. More precisely, it first asks respondents to provide an estimate of the share of the population engaged in undeclared activities (in Germany). It then proceeds to asking respondents about whether relatives or friends engage in undeclared activities (also known as 'other people' approach Barton (1958)), before turning to the respondents' own potential engagement (starting with demand and then asking about supply) (Williams, 2010, p. 252).

Schneider et al. (2002, p. 44) use the so-called 'everybody approach' (Barton, 1958): Undeclared work is described as something normal, suggesting that it

4 The definition includes reciprocal favors and in kind payment, also by friends (Feld and Larsen, 2005).

5 "Respondents were asked to report as undeclared work all remunerated activities which are in principle legal, but circumvent declarations to tax authorities or social security institutions" (EC – European Commission, 2007, p. 8). Illegal activities were not intended to be reported. The Eurobarometer, however, did ask respondents to report undeclared work paid for with money as well as in kind.

is a frequent and socially accepted behavior, making it easier for respondents to admit to such.

Merz and Wolff (1993, p. 181) use yet a different strategy, deriving indirectly whether respondents engage in undeclared work. When respondents (age 14 and older) confirm having had a secondary occupation within the last three months of the interview, follow up questions regarding costs associated with this occupation are asked. The knowledge of whether respondents have expenditures or costs associated with this side job, e.g., for material, social security contributions, or taxes, then allows researchers to derive implicitly whether this job is undeclared. This approach, however, focuses only on moonlighting and assumes the primary occupation is legal.

Similar to the estimates derived from macroeconomic approaches discussed above, the latest estimates regarding undeclared work in Germany based on survey data vary tremendously: The share of respondents admitting to having engaged in undeclared work, ranges from 3.0% (age 15 and older) in the Eurobarometer 2007 (EC – European Commission, 2007, p. 115) to 11.1% and 7.2% of respondents (age 18 to 74) in the studies conducted by Feld and Larsen in 2005 and 2006 (Feld and Schneider, 2010, p. 123).

To summarize the main strategies to tackle misreporting, researchers can increase perceived privacy in an interview setting using (Lee, 1993; Groves et al., 2004): a more private interview mode (Kreuter et al., 2008) also avoiding interviewer and/or bystander effects (Aquilino et al., 2000), confidentiality assurances in a sensible measure (Singer et al., 1995), an adapted wording of the survey questions or the format of the survey response (Barton, 1958; Tourangeau and Smith, 1996), or last but not least, specific survey techniques (Warner, 1965; Droitcour et al., 1991) that introduce a probabilistic relationship between the survey question and the survey response (e.g. the RRT or the ICT).

While the former strategies have been applied in surveys on undeclared work in general population surveys, dejeopardizing techniques have not been used in the collection of data. The following sections will briefly introduce the main ideas of the RRT and the ICT, their estimators and their limitations since the application of these special techniques is the focus of this dissertation.

1.3.3 The Randomized Response Technique

1.3.3.1 The General Idea

The RRT method was originally developed by Warner in 1965 to reduce response bias arising from privacy concerns, and has been proposed and implemented

in many different variants (Horvitz et al., 1967; Greenberg et al., 1969; Boruch, 1971; Greenberg et al., 1971; Moors, 1971; Kuk, 1990; Mangat and Singh, 1990; Mangat, 1994). The basic idea, common to all RRT variants, is to conceal a respondent's answer by using a randomizing device (e.g., coins, cards, dice, spinner), whose outcome is only known to the respondent. For an overview of different RRT designs, estimators and applications see Fox and Tracy (1986), Umesh and Peterson (1991), Lensvelt-Mulders et al. (2005a), Lensvelt-Mulders et al. (2005b), or Tourangeau and Yan (2007).

For example, in the so-called forced-response variant of the RRT (Boruch, 1971) the respondent might be requested to flip three coins without revealing the result to the interviewer. The respondent is then instructed to answer 'Yes' if the outcome is 'tails' for all coins ($p = 0.125$), answer 'No' if the outcome is 'heads' for all coins ($p = 0.125$), or answer the sensitive question truthfully with 'Yes' or 'No' if the outcome is mixed ($p = 0.75$), i.e., 'tails' for one of the coins and 'heads' for the others or vice versa. The essential feature of the RRT is that the connection between the observed answer and the sensitive question is only probabilistic. Due to this randomization or chance component, no inference can be drawn at the individual level regarding the sensitive characteristic: Neither the interviewer nor the researcher can infer anything about the true status of any individual from his or her response. Since the randomization mechanism—and thus the probability distribution and the misclassification—is known by the researcher, estimation of the population prevalence of the sensitive characteristic under study (Fox and Tracy, 1986) is possible, as are regression analyses analyzing randomized response dependent variables (Maddala, 1983, p. 54 ff.).

In general, we do not know anything about the 'true' status of a respondent and have no means of validating individual responses or the overall prevalence estimate. The success of the RRT or more generally these special techniques, i.e. if we can elicit more truthful self-reports and observe less misreporting, is hence often established in reference to a baseline: typically a standard direct questioning condition. Due to lack of validation data, researchers commonly rely on this so-called 'more-is-better' assumption: These special techniques are assumed to outperform direct questioning, if they elicit higher prevalence estimates for questions that are assumed to be subject to underreporting ('more-is-better' assumption, for an overview see Umesh and Peterson, 1991; Lensvelt-Mulders et al., 2005a; Tourangeau and Yan, 2007).

So far, many surveys applying the RRT have been conducted in a face-to-face mode, some in a telephone mode and more recently also in a self-administration mode using the web (Lensvelt-Mulders et al., 2005a; Tourangeau and Yan, 2007; Holbrook and Krosnick, 2010a; Coutts and Jann, 2011). Empirical studies indicate

that in many cases the RRT yields higher prevalence estimates of a sensitive behavior than direct questioning and less biased estimates ('more-is-better' assumption). Successful experimental studies comparing the RRT to direct self-reports (in various survey modes) of sensitive or stigmatizing behavior analyze topics, such as: illicit drug use (Weissman et al., 1986), social security fraud (Van der Heijden et al., 2000), abortion (Lara et al., 2004, 2006), eating disorders (Lavender and Anderson, 2009), animal diseases (sheep scab) (Cross et al., 2010), dental hygiene (Moshagen et al., 2010), antisemitism (Krumpal, 2012), sexual behavior in sub-Saharan Africa (Anglewicz et al., 2013), or the use of cognitive-enhancing drugs (Dietz et al., 2013). Franke et al. (2013), in their study on the use of drugs for cognitive and mood enhancement, found that the RRT outperformed anonymous self-reports. However, in other recent experimental studies analyzing topics, such as voter turnout (here: 'less-is-better' Holbrook and Krosnick, 2010a),[6] freeriding, marijuana use or infidelity (Coutts and Jann, 2011), plagiarism (Coutts et al., 2011), as well as criminal convictions (Wolter, 2012), the RRT proved less successful compared to direct questioning.

Information on the exact implementation of the RRT in this dissertation can be found in Chapter 2.1.1.3.2.

1.3.3.2 Estimators

Assuming that the probability distribution of the randomization procedure (here: forced-choice RRT design) is known, the population prevalence as well as standard errors standard errors (s.e.) and confidence intervals (C.I.) can be estimated as follows: The observed sampling distribution of 'Yes' responses $\hat{\Phi}$ is used as an estimator for the unknown population parameter Φ. The overall proportion of positive responses (Φ) is the sum of the proportion of 'forced' 'Yes' responses (p_1 = 'forced' 'Yes' = 0.125; recall: p_2 = 'forced' 'No' = 0.125), and the product of the (unknown) population parameter π multiplied by the probability of having to respond truthfully (p_3 = truthful response = 0.75) in the first place: $\Phi = p_1 + p_3 * \pi$. Rearranging this equation yields the population estimate for the prevalence of the sensitive characteristic $\hat{\pi}_{RRT}$ (Lensvelt-Mulders et al., 2005b):

$$\hat{\pi}_{RRT} = \frac{\hat{\Phi} - p_1}{p_3} \tag{1.1}$$

An estimate of the sampling variance of $\hat{\pi}_{RRT}$ is given as:

6 This study implemented two surveying modes: telephone and web. In neither mode did the RRT provide lower prevalence estimates.

$$Var(\hat{\pi}_{RRT}) = \frac{\hat{\Phi} * (1-\hat{\Phi})}{n * (p_3)^2} \tag{1.2}$$

where n is the overall sample size.

Since the misclassification design is known (p_1, p_2 and p_3), we can use a binary logistic regression model with an adapted likelihood function. Assuming a vector of explanatory variables X, we want to estimate the effect of each variable on the sensitive characteristic. Let the observation be $y = 1$ if the person has a sensitive trait, and $y = 0$ otherwise. The ordinary multiple logistic regression model for binary data in the direct questioning condition is then defined as (for all subsequent equations see Maddala, 1983, p. 55):

$$Prob(y_i = 1) = \frac{exp^{\beta' X_i}}{1 + exp^{\beta' X_i}}, \tag{1.3}$$

where β is a vector of unknown regression parameters. Maximizing the likelihood function yields a parameter estimate for β:

$$L = \prod_{y_i = 1} \left(\frac{exp^{\beta' X_i}}{1 + exp^{\beta' X_i}} \right) \prod_{y_i = 0} \left(\frac{1}{1 + exp^{\beta' X_i}} \right) \tag{1.4}$$

Adapting the likelihood function to account for the misclassification yields:

$$L = \prod_{y_i = 1} \left(p_1 + p_3 \frac{exp^{\beta' X_i}}{1 + exp^{\beta' X_i}} \right) \prod_{y_i = 0} \left(p_2 + p_3 \frac{1}{1 + exp^{\beta' X_i}} \right) \tag{1.5}$$

Remember that $Prob(y_i = 1) = p_1 + p_3 * (\pi_i)$ and $Prob(y_i = 0) = p_2 + p_3 * (1 - \pi_i)$. If we assume that the probability $\pi(X_i)$—which is the the individual probability of giving a positive answer to the sensitive item conditional on a set of covariates X_i —can be written as $\pi(X_i) = exp^{\beta' X_i} / 1 + exp^{\beta' X_i}$, simplifying and modifying equation 1.5, the log-likelihood for randomized response data becomes (Van der Heijden et al., 2000, p. 259):

$$log\, L = \sum_i n_{i1} log(p_1 + p_3 \pi(X_i)) + \sum_i n_{i0} log(p_2 + p_3 [1 - \pi(X_i)]) \tag{1.6}$$

where n_{i1} (n_{i0}) is defined in terms of the total number of respondents for whom a 'Yes' ('No') response is observed in the RRT design. Using an iterative procedure (more precisely, the Newton-Raphson updating algorithm), this log likelihood can be maximized over the regression parameters β and also yields an estimate of the asymptotic covariance matrix of these estimates. The Stata routine 'rrlogit' (Jann, 2011 following Maddala, 1983) is one possibility to fit a maximum-likelihood logistic regression for RRT data (see also Scheers and Dayton, 1988; Van der Heijden et al., 2000, p. 259; Van den Hout and van der Heijden, 2002).

Average marginal effects (AME) can then be derived using the Stata command 'margins' (Williams, 2012). Standard errors are calculated using the Delta-Method (Oehlert, 1992). Using AME eases interpretation and allows for a comparison across models: The AME expresses the average effect of the independent variable x on Prob(y = 1) (Mood, 2010, p. 75). In other words, an increase of x by one unit increases the probability of y = 1 by AME (percentage) points (pts) (Best and Wolf, 2012, p. 384).

1.3.3.3 Limitations

One major drawback of the RRT is that—due to the 'random noise' component introduced to the data (p_1 and p_2)—a larger sample size is needed to achieve the same level of precision as a comparable direct questioning (Warner, 1965; Fox and Tracy, 1986). This is particularly relevant with respect to survey costs which in turn directly influences the choice of the survey mode. These additional costs are justified only if the RRT increases the amount of truthful self-reports and reduces bias significantly.

If respondents understand the RRT procedure and appreciate the induced privacy protection (Landsheer et al., 1999), they should be inclined to provide more honest answers to sensitive questions than under direct questioning.[7] This is another limitation inherent to the RRT: Evaluations of the technique typically rely on the 'more-is-better' assumption.

Another drawback of the RRT is that almost all empirical implementations of the RRT focus on binary sensitive variables. Although RRT schemes tailored to continuous sensitive characteristics have been proposed in the literature (cf. Himmelfarb and Edgell, 1980; Eichhorn and Hayre, 1983; Gjestvang and Singh, 2007; Peeters et al., 2010), there is little evidence on how these techniques perform in practice. Due to its complexity, it can be expected, however, that an RRT scheme for continuous variables is even more difficult to implement than standard RRT and imposes an additional cognitive burden on the respondents.

7 Some studies report serious problems with RRT, such as a substantial proportion of respondents not understanding or not trusting the procedure and providing self-protective 'No' answers irrespective of the outcome of the randomizing device (see Holbrook and Krosnick, 2010a).

1.3.4 The Item Count Technique

1.3.4.1 The General Idea

The second prominent technique that will be used throughout the dissertation is the item count technique (ICT). The item count technique is also known as the 'unmatched count technique' (Ahart and Sackett, 2004; Dalton et al., 1994), the 'block total response' (Smith et al., 1975; Raghavarao and Federer, 1979), or the 'list experiment' (Kuklinski et al., 1997; Corstange, 2009). Again, the main idea is to scramble the individual response in order to protect respondents' privacy. Similar to the RRT, the ICT allows for an estimation of the population prevalence of the sensitive characteristic, as well as for regression analyses (Blair and Imai, 2012).

In the ICT (single-list) design, two subsamples of respondents are generated via randomization. One of the subsamples is confronted with a long list of items (LL) containing a number of innocuous (non-sensitive) questions plus the sensitive question of interest. The other subsample receives a short list (SL) that only contains the innocuous questions. For example, the following list of questions could be used (Table 1.1):

Table 1.1: The item count technique: single-list design

Item	Short List	Long List
Do you use public transportation on more than 5 days per week?	X	X
Are you covered by liability insurance?	X	X
Did you grow up in the countryside?	X	X
Have you engaged in any undeclared work for a private person this year?		X

Respondents in each subsample are simply asked to indicate the number of items that apply to them (the total number of 'Yes' answers), without answering each question individually. Unless a respondent indicates that all or none of the items apply, it remains unknown whether the respondent engaged in the sensitive behavior or not. The population estimate of the sensitive behavior is then derived from difference in means between both subsamples.

One of the main disadvantages of the ICT, however, is its relative inefficiency: The variance of the estimator is dependent on the variance of the innocuous questions and their number. Droitcour et al. (1991) recommends using three to five innocuous items per list in order to achieve sufficient privacy protection.

In order to avoid bias from so-called 'ceiling effects' (Kuklinski et al., 1997), i.e. absolute transparency regarding the sensitive characteristic when all items apply to a respondent, the general piece of advice given is the following (Glynn, 2013, p. 163): Both, high-prevalence and low-prevalence innocuous items should be avoided and the number of innocuous items should be sufficiently high. These recommendations, however, lead to an increased variance in the number of applicable items and to an increased sampling variance of the estimator. In order to minimize the risk of ceiling effects, yet, still keep the variance minimal, Glynn (2013) suggests that it can be advantageous to choose innocuous items that are negatively correlated. Furthermore, in the implementation introduced above—the single-list design—only one subsample actually receives the sensitive question. Here, the effective sample size is further reduced (divided in half) compared to direct questioning, again contributing to an increase in the variance of the estimator (Droitcour et al., 1991).

In order to increase the statistical efficiency of the ICT estimator, the so-called double-list implementation was suggested (Droitcour et al., 1991). The double-list version generates two independent item count estimates of the sensitive characteristic: Using an additional list of innocuous items, the sensitive item will be used in both subsamples. The first subsample will receive the sensitive question in the first block (long list 1; LL1), while the second subsample receives the short list (short list 1; SL1). For the second list, comprising another set of innocuous items, the first subsample will receive only the short list (short list 2; SL2), while this time respondents in the second subsample receive the long list (long list 2; LL2) including the same sensitive item that was used in LL1 in the first subsample.

Table 1.2 schematically sketches the ICT double-list design.

Table 1.2: The item count technique: double-list design

	Subsample 1	Subsample 2
List 1	LL1: SL1 & 'undeclared work'	SL1
List 2	SL2	LL2: SL2 & 'undeclared work'

Tourangeau and Yan (2007), in their meta-analysis on the ICT report less consistent results for the ICT than for the RRT. Again, relying on the 'more-is-better' assumption, 'successful' experimental studies comparing the ICT to direct self-reports of sensitive or stigmatizing behavior examine topics, such as: unethical

employee behavior (Dalton et al., 1994), racial attitudes ('affirmative action') (Kuklinski et al., 1997), employee theft (Wimbush and Dalton, 1997), risky sexual behavior (LaBrie and Earleywine, 2000), hate crime victimization (Rayburn et al., 2003), shoplifting (Tsuchiya et al., 2007), eating disorders (Lavender and Anderson, 2009), voter turnout (here: 'less-is-better' Holbrook and Krosnick, 2010b),[8] or freeriding, marijuana use or infidelity (Coutts and Jann, 2011). This success was not always replicated. The ICT was less successful in the following studies analyzing topics, such as intravenous drug use or receptive anal intercourse (Droitcour et al., 1991), engagement in counterproductive behavior (Ahart and Sackett, 2004), drug use (Biemer et al., 2005), or plagiarism (Coutts et al., 2011). Tsuchiya and Hirai (2010) as well as Glynn (2013) explicitly analyze why the ICT sometimes performs so poorly.

Compared to the RRT, the ICT has the advantage that it does not require a randomizing device and that the procedure is much easier to administer. Thus, only a moderate cognitive burden is imposed on the respondent, likely increasing the respondent's ability and trust to comply with the interview protocol and to provide more honest self-reports (Lavender and Anderson, 2009; Coutts and Jann, 2011). Only few empirical studies exist that compare the performance of the ICT relative to that of the RRT: These studies suggest that the ICT outperforms the RRT in reducing social desirability bias in survey measures of sensitive attributes (see Lavender and Anderson, 2009, mode: paper-and-pencil, population: student population; Holbrook and Krosnick, 2010a, b, mode: telephone, population: national sample of American adults; Coutts and Jann, 2011, mode: web, population: German "Sozioland" access panel).[9] One disadvantage of the ICT is its low statistical power. Estimates obtained from the ICT typically have larger standard errors than estimates from the RRT based on the same sample size (see Coutts and Jann, 2011).

1.3.4.2 Estimators

Droitcour et al. (1991) and Biemer et al. (2005) provide a thorough overview regarding the analyses of ICT data which will be briefly presented below. Remember that for the single-list design, the mean difference of answers between the two subsamples provides an estimate for the population prevalence $\hat{\pi}_{ICT}$ of the sensitive behavior (Droitcour et al., 1991).

8 This result was only found for the telephone mode but could not be confirmed using an online mode, indicating that the online mode might be less susceptible to social-desirability bias.

9 Though Coutts et al. (2011) report in their web-survey that both, the RRT and the ICT performed poorly with respect to plagiarism among students.

$$\hat{\pi}_{ICT} = \bar{x}_{LL} - \bar{x}_{SL}, \tag{1.7}$$

where $\bar{x}_{LL}$ is the mean estimate in the long list and $\bar{x}_{SL}$ the mean estimate in the short list subsample.

Furthermore, as long as the samples are independent, the variance of $\hat{\pi}_{ICT}$ can be estimated as the sum of the sampling variances of the two group means, that is:

$$Var(\hat{\pi}_{ICT}) = Var(\bar{x}_{LL}) + Var(\bar{x}_{SL}) \tag{1.8}$$

Turning to the double-list estimator, we obtain two independent estimates for the population prevalence:

$$\hat{\pi}_1 = \bar{x}_{LL1} - \bar{x}_{SL1} \tag{1.9}$$

and

$$\hat{\pi}_2 = \bar{x}_{LL2} - \bar{x}_{SL2} \tag{1.10}$$

Taking the mean of these estimates yields the overall population estimate (Biemer et al., 2005; Coutts et al., 2011):

$$\begin{aligned} \hat{\pi}_{DL} &= \frac{\hat{\pi}_1 + \hat{\pi}_2}{2} \\ &= \frac{(\bar{x}_{LL1} - \bar{x}_{SL1}) + (\bar{x}_{LL2} - \bar{x}_{SL2})}{2} \\ &= \frac{(\bar{x}_{LL1} - \bar{x}_{SL2}) + (\bar{x}_{LL2} - \bar{x}_{SL1})}{2} \end{aligned} \tag{1.11}$$

The reformulated last equation 1.11 yields for the expression in the first (second) bracket the overall sum in the first (second) subsample over the differences of a respondents' long-list and short-list answer.

Compared to the single-list variance estimate, the double-list sampling variance estimator is more efficient (Biemer et al., 2005; Coutts et al., 2011):

$$\begin{aligned} Var(\hat{\pi}_{DL}) &= \frac{Var(\hat{\pi}_1) + Var(\hat{\pi}_2) + 2COV(\hat{\pi}_1, \hat{\pi}_2)}{4} \\ &= \frac{Var(\bar{x}_{LL1} - \bar{x}_{SL2}) + Var(\bar{x}_{LL2} - \bar{x}_{SL1})}{4} \end{aligned} \tag{1.12}$$

For a detailed overview of ICT estimators modeling ICT data as a function of covariates, the reader is referred to Blair and Imai (2012) as well as Glynn (2013).

The R-package 'list,' for example, in its current form only allows for single-list implementations of the ICT (Blair and Imai, 2012).

1.3.4.3 Limitations

Remember that one of the main disadvantages of the ICT estimator is its relative inefficiency depending on the variance of the innocuous questions and their number. Thus, as in the case of RRT, a larger sample size is needed—even if the double-list design is used—to achieve the same precision as a comparable direct questioning method (Tsuchiya et al., 2007). Similar to the RRT this argument is relevant with respect to survey costs and the choice of the survey mode.

While seemingly granting more anonymity compared to the RRT (Lavender and Anderson, 2009; Coutts and Jann, 2011), the ICT is susceptible to so-called 'floor' and 'ceiling effects' (Kuklinski et al., 1997): Whenever either no item or all items apply to a respondent, anonymity is no longer granted. The researcher can then directly make inferences about the sensitive trait. Thus, incentives to misreport persist in singular instances even if the innocuous items are chosen following the design advice from previous studies summarized in Glynn (2013, p. 163). Another limitation particular to the ICT relates to the assumption of 'no-design-effects' (Tsuchiya and Hirai, 2010; Blair and Imai, 2012). This assumption relates to the fact that responses can be influenced by the question format itself, i.e., here the sum of applicable items as opposed to a response to the individual items.

Like the RRT, other limitations that also hold for the ICT are the often relied upon 'more-is-better' assumption and the focus on binary sensitive variables. We are not aware of any implementation beyond the collection of binary data for the ICT.

1.4 Summary of Research Gaps and Research Questions

The previous section argued that macro-level approaches cannot sufficiently differentiate between the shadow economy and undeclared work in particular, providing estimates that are too inclusive. Furthermore, these approaches cannot provide insights into individual motivations to engage in undeclared work. While survey-based approaches allow for that last possibility, they provide a lower boundary of prevalence estimates of undeclared work due to potential underreporting.

Comparing existing prevalence estimates of survey-based and indirect, macro-economic approaches, leads us to the assumption that commonly applied strategies to 'dejeopardize' questions in surveys concerning undeclared work do not suffice to eliminate response bias and that the results of existing surveys on

undeclared work in Germany are still downwardly biased to an unknown extent. None of the existing studies on undeclared work has analyzed the benefits of using the randomized response or the item count technique avoid errors from (item) nonresponse and misreporting (here: 'underreporting') in the context of undeclared work in general population surveys (Boockmann et al., 2010, p. 100). Studying the RRT and the ICT implemented in a telephone mode, as opposed to a face-to-face mode, might prove worthwhile from a practicability point of view given the potential cost savings and a potentially increased sense of privacy for the respondents (Weissman et al., 1986; Groves, 1989; Schnell, 2012; Krumpal, 2012).

The main question is thus whether the quality of sensitive information collected by means of direct questioning in labor market surveys can be improved, using the RRT or the ICT. All empirical work in this dissertation is based on data collected in two experimental studies (see Chapter 2.1).

The *first research question* addressed in Chapter 2—'Measuring and Explaining Undeclared Work in Germany'—concerns the performance of these two special data collection techniques—the RRT and the ICT—in the specific context of undeclared work. The question is whether we can elicit more truthful self-reports of undeclared work using the RRT or the ICT compared to direct questioning ('more-is-better' assumption) and obtain higher prevalence estimates. Furthermore, we are interested in investigating which factors contribute to the explanation of undeclared work (using individual-level survey data). To address this latter research question, we briefly introduce a theoretical foundation for the explanation of undeclared work and our hypotheses that are to be tested. Using logistic regression models, we examine individual characteristics fostering engagement in undeclared work using the RRT data.

The *second research question*, is presented in Chapter 3—'Item Sum: A Novel Technique for Asking Continuous Sensitive Questions'—and addresses the challenge that both techniques—the RRT and the ICT—have typically been applied to the collection of binary data. We thus developed a technique that is suitable for collecting information on sensitive continuous variables in a telephone survey: the item sum technique (IST). The IST is a generalization of the item count technique for collecting information on sensitive continuous variables. The main research question we address in this chapter is whether the IST can outperform standard direct questioning. For this evaluation, we collected data on hours engaged in undeclared work and earnings from undeclared work. Furthermore, we examine whether the IST performs differently in different subgroups, depending on differential cognitive abilities.

The *third research question* addressed in Chapter 4—'Validating Sensitive Questions: A Comparison of Survey and Register Data'—deals with the 'more-

is-better' assumption, which is so often relied upon in the literature on sensitive questions (Lensvelt-Mulders et al., 2005a). Whether or not these dejeopardizing techniques really outperform direct questioning can only be assessed in the presence of validation data. By nature, the data collected on undeclared work cannot be validated. However, other sensitive labor market information that is also known to be underreported can be used for this purpose. One topic that satisfies these requirements is the receipt of basic income support, a form of means-tested social security payment or welfare (more precisely, unemployment benefits II (UB II)—colloquially also referred to as 'Hartz IV'). It can be assumed to be socially stigmatizing and sensitive and is known to be subject to underreporting (Kreuter et al., 2010, 2013).

The RRT study is particularly designed to allow an in-depth analysis of the performance of the RRT and to relax the 'more-is-better' assumption using validation data on 'welfare benefit receipt.' We make use of the fact, that our samples in the RRT study are drawn from a frame containing additional information regarding the receipt of welfare benefit. Stratifying by this indicator, the sample is constant with respect to the dependent variable welfare benefit receipt, avoiding individual linkage of survey and register data, but nonetheless permitting validation. Combining this information from administrative records and survey data on welfare receipt, we know the true percentage of respondents who have received transfer payments for basic income support, and hence the percentage of people who should have reported receipt. This permits us to validate the reported percentage against the known true rate for the responding cases and to evaluate underreporting in the RRT and the direct questioning condition. Furthermore, this allows us to analyze individual-level factors contributing to reporting accuracy—something that is impossible with aggregate data relying on the 'more-is-better' assumption.

To analyze which technique performs better with respect to the true value as well as an assessment of mechanisms that might increase reporting accuracy are the central issues addressed in this chapter. Only a small number of validation studies evaluating the RRT exist (Lensvelt-Mulders et al., 2005a; Wolter, 2012). To our knowledge, all of these validation studies have implemented the RRT in a face-to-face mode. We thus evaluate the RRT in the context of a telephone mode.

2 Measuring and Explaining Undeclared Work in Germany

As outlined above, the following chapter will investigate data collection strategies that were developed solely for the purpose of asking sensitive questions in surveys: the randomized response technique and the item count technique.[1] This chapter serves two main purposes: a methodological one and a substantive one. The methodological purpose is to assess data quality and measurement error with respect to undeclared work. The substantive goal is to estimate the amount of undeclared work and analyze its determinants.

The contribution to the existing literature is threefold. *First*, there are no population surveys in Germany particularly addressing the problem of social desirability associated with asking questions concerning undeclared work. Thus, one focus of this chapter is to analyze the potential of these special techniques, in the context of undeclared work (Boockmann et al., 2010, p. 100). More precisely, we assess whether we can reduce underreporting of undeclared work by means of RRT or ICT and thus obtain 'increased', i.e., more accurate prevalence estimates compared to a direct questioning approach ('more-is-better' assumption, see Tourangeau and Yan, 2007).

Second, all techniques will be evaluated according to their performance in different sub-populations. The RRT study provides us with an opportunity to evaluate the functioning of the RRT in two very different subpopulations by oversampling recipients of welfare benefit in relation to employees. It is known from prior studies that the 'success' of the RRT depends on respondents understanding and trusting the procedure (for the difficulty of using RRT in populations with limited language skills, see Landsheer et al., 1999). Compared with a general population, individuals who receive welfare are known to usually have a lower educational background, as well as a migrant background (Aldashev and Fitzenberger, 2009). This explains why general problems of communication in surveys are more frequent and might jeopardize the applicability of the RRT. The RRT study is thus conducted in a setting with high external validity particularly with regard to other future telephone surveys of low-income populations.

Complementing the uni- and bivariate analyses regarding the prevalence of undeclared work (overall and in subpopulations), the *third* contribution relates to a substantive question: Which factors contribute to the explanation of undeclared work (using individual-level survey data). Relying on theories concerning undeclared work, we will empirically test existing hypotheses. Thus complex relationships

1 This chapter is based on a paper by Kirchner et al. (2013) and reprinted with modifications by permission of Lucius & Lucius Verlagsgesellschaft mbH.

between undeclared work and socio-demographic characteristics, opportunity and incentive structures, as well as normative attitudes can be analyzed at an individual level. These analyses were conducted using logistic regression analyses using the Stata routine 'rrlogit', with an adapted likelihood function (Jann, 2011).

The chapter outline is as follows: Section 2.1 describes all relevant study specific details. Prevalence estimates are presented in Section 2.2.1, while the results of the multivariate regression models—including the theoretical foundation and hypotheses to be tested—are presented in Section 2.2.2. The chapter ends with a discussion and conclusion of our results (Section 2.3).

2.1 Study Details: The Experiments

Data utilized in this dissertation are drawn from two nation-wide telephone surveys that experimentally assigned respondents into the experimental conditions: direct questioning, randomized response, and item count.

Aside from differences in the sampling design (see below), the two studies differ only in the experimental set-up, i.e., the randomized response and the item count technique. Using a split-ballot design in both studies, errors from specification, frame, nonresponse as well as errors from data processing can be assumed to be constant in both experimental groups and therefore will not be discussed any further.

The first survey—the RRT study—was commissioned by the Institute for Employment research (IAB) an independent institute of the German Federal Employment Agency (FEA), and was carried out by the ForschungsWerk institute (Nuremberg) from October 18th to December 10th, 2010. In approximately the same period (25.10.2010 to 22.12.2010), the second survey—the ICT study, commissioned by the University of Leipzig—was carried out by the Usuma institute (Berlin). Overall, a total of 3,211 interviews were completed in the RRT study, and a total of 1,603 in the ICT study.[2]

The following section will provide a general overview of the study designs, while the remaining chapters will provide more detailed information, such as specific operationalizations, whenever warranted (see Chapter 3 and 4).

2 While in total 1,606 interviews were completed in the ICT study, three interviews turned out to be invalid and could not be used for further analyses.

2.1.1 The RRT Study

2.1.1.1 Sampling and Data Collection

We chose a particular sampling design that would enable us to analyze two samples with different incentive and opportunity structures for undeclared work. The RRT study is a dual-frame survey, using sampling frames that are maintained by the German Federal Employment Agency (IAB Unemployment Benefit II History (LHG) V6.03.01 and (XLHG) V01.06.00-201007; IAB Employment Histories (BeH) V08.04.00, Nuremberg 2010).

These frames consist of all registered unemployment benefit (II) recipients as well as all employed persons who are subject to social security contributions. The register includes all employees who are subject to social security contributions, that is, all people with income from dependent work in a specific month. Self-employed and civil servants are not covered because they do not pay social security contributions.

The first random sample was drawn from the FEA registers of basic income support recipients. It consists of people aged 18 to 64 who were known to have received basic income support in June 2010 (a form of means-tested—essentially welfare—benefits called 'Unemployment Benefit II' paid up to the age of 64, henceforth referred to as UB II or benefit recipients sample). The second random sample was drawn from the register of employees maintained by the FEA. It consists of people aged 18 to 70 who were employed in December 2009 (henceforth referred to as employee sample). For both samples the latest available registers were used. Nonetheless, employment status or UB II eligibility may have changed by the date of the interview. That is, a part of the respondents in the employee sample or the benefit recipient sample was no longer employed or receiving benefits when being interviewed.

The FEA registers only contain telephone numbers for about forty percent of the employees and ninety percent of benefit recipients (landline and mobile phone numbers). Furthermore, it is known from past surveys that some of these numbers are out of date. We therefore tried to complete the numbers using public telephone directories. Nonetheless, 31.8% of the employee sample and 8.3% of the benefit recipient sample remained without telephone number. During fieldwork, 17.5% of the phone numbers in the employee sample and 17.2% of the phone numbers in the benefit recipient sample turned out to be invalid and could not be replaced by a working phone number from the public directories. Among those who could be contacted, about 26% agreed to participate in the survey. However, due to the large proportion of missing or invalid phone numbers, the overall response rates were 16.3% and 18.8% in the two samples respectively

(RR1 according to AAPOR 2011).[3] Table 2.1 provides an overview of the sample sizes and response rates.

Every selected individual received a personalized advance letter, inviting them to participate in the labor market survey "Living and Working in Germany." Each advance letter contained a toll free number, as well as a link to a homepage to access in case of further questions. Furthermore, individuals were informed about the topic of the survey (labor market developments), survey sponsorship, data privacy and the voluntary nature of participation. However, neither the main substantive topic (undeclared work and welfare receipt) nor the conduct of methods experiments were mentioned in the advance letter.

Table 2.1: RRT study: sample sizes and response rates

	Employees	Benefit recipients	Total
Gross sample	9,996	8,999	18,995
Net sample (with phone number)	6,820	8,250	15,070
Invalid phone number	1,196	1,422	2,618
Non-contact	813	1,564	2,377
Refusal	3,094	3,173	6,267
Ineligible (deceased, moved abroad, outside age range, speaks no German)	104	493	597
Interview completed	1,613	1,598	3,211
Response Rate (AAPOR RR1)	16.3%	18.8%	17.5%

2.1.1.2 Experimental Design

In both samples—employees and benefit recipients—about one third of the respondents in the gross sample was randomly assigned to direct questioning (DQ; n = 1,145), and the remaining two-thirds to the randomized response technique (RRT; n = 2,066). Prior to fieldwork, regression analyses were conducted showing that randomization into experimental groups within subsamples worked appropriately and that sample composition did not differ significantly. Unit nonresponse affected both experimental conditions equally. Table 2.2 provides an overview of the assignments to the experimental conditions.

3 Overall, 275 respondents completed the interview only partially (subsumed under 'refusal' in Table 2.1), i.e., all respondents who started the experimental section. Out of those 275, 124 dropped out prior to the experimental condition, while the majority of the remaining participants dropped out upon entering the experimental condition (n = 95), and 22 when confronted with questions regarding undeclared work. 34 participants dropped out after the experimental condition.

Table 2.2: RRT study: experimental conditions

Study	Assigned Condition	N	Realized Condition	N
RRT study	DQ	1,145	DQ	1,145
	RRT	2,066	RRT	1,792
			DQ_RRT	274

The unequal assignment to the experimental conditions was necessary to achieve approximately the same level of statistical precision in the RRT condition (Warner, 1965; Cohen, 1988). The loss of statistical efficiency in the RRT condition is due to the additional random noise component.

Some of the respondents originally assigned to the RRT refused the application of the randomized response technique and were subsequently asked to respond to the relevant survey questions directly (DQ_RRT; n = 274). Out of those 274 respondents (13.3% of the original RRT sample), 201 reported that they 'were not in the mood for RRT' and 73 reported that they 'did not have coins at hand.' While it can be shown in analyses not presented here that respondents in the direct questioning and the randomized response split do not differ significantly in substantive variables, further tests for the 'non-compliers' or 'defiers' show that these groups significantly differ in some respects from the RRT compliers (see Section 2.1.3). Thus all further analyses will be conducted separately. Section 2.1.3 provides some information regarding the final sample composition.

2.1.1.3 Questionnaire

2.1.1.3.1 General Information

Respondents in both studies—the RRT study and the ICT study—received identical questionnaires, differing only in the experimental sections. Both surveys used the same set-ups with respect to: assurances of confidentiality and anonymity, introduction to the topic, definitions (undeclared work) and further explanations, if needed.

The questionnaire initially collected information concerning individual employment histories and labor market issues. Following this section, the experiments were implemented collecting data using either standard direct questioning methods or the randomized response technique.

Within the experimental section, respondents first received a question regarding welfare benefit receipt. This question concerning receipt of unemployment benefit II (yes/no) was asked prior to the questions regarding undeclared work (yes/no). The item sum technique (asking follow-up questions on

hours engaged in undeclared work and income from undeclared work) succeeded the RRT section. For the exact wording of the introductions to each experimental condition, see Appendix A.1 (RRT) and A.2 (IST).

Following the experiments, the questionnaire concluded with a section asking about attitudes towards undeclared work and labor market participation as well as socio-demographic characteristics, such as age or education. These additional data are used to form the explanatory variables in subsequent analyses. Average questionnaire completion time was 18.1 minutes.

In the course of questionnaire development, 31 cognitive paper-and-pencil interviews were conducted to identify substantive and methodological problems (e.g., with respect to RRT, or IST instructions). After adjusting the questionnaire accordingly, a second pretest (n = 63) was fielded by the data collection institute to test the instruments as well as all relevant filters and randomizations.

Interviewers in the RRT and the ICT study received identical, study-specific training. Furthermore, they received additional training in order to successfully learn how to implement the experiments and reply to respondent questions or concerns in a standardized way.

2.1.1.3.2 The RRT Implementation

Following the logic outlined in Section 1.3.3, we opted to minimize two respondent hazards in our specific RRT-design: Neither a positive ('Yes') nor a negative ('No') answer should risk suspicion. We decided to implement a symmetric forced-response RRT variant (Boruch, 1971). According to a study conducted by Lensvelt-Mulders et al. (2005b) the forced choice implementation of the RRT has the highest statistical efficiency among different RRT designs and is usually best understood (Landsheer et al., 1999; De Schrijver, 2012). In this RRT variant, neither a positive nor a negative response reveals anything about the true status of the respondent.

More precisely, depending on the outcome of a randomization device, respondents are instructed to reply according to a set of rules: They are asked to either give a truthful answer to the sensitive question, i.e., 'Yes' or 'No,' a forced 'Yes,' or a forced 'No'—irrespective of their true status. As a rule of thumb Lensvelt-Mulders et al. (2005b) suggest that the probability of providing a forced 'Yes' should be approximately the same as the expected prevalence of the sensitive item under investigation while the probability to tell the truth should be between 0.7 and 0.8. Since the main item under study was 'undeclared work,' with an assumed prevalence of about 10% to 12%, we chose the probabilities of a forced 'Yes'/'No' and 'the truth' accordingly.

In our study, respondents in the RRT condition were asked first to gather paper and pencil in order to note the rules. Respondents were then asked to flip

three coins prior to each question in the RRT section. Due to the telephone mode, privacy with respect to the outcome of the coin flip was easily provided. Should a respondent nonetheless accidentally reveal the outcome, interviewers were trained to ask respondents to flip the coin again without revealing the outcome. The exact rules implemented to provide an answer were the following:

> *"Please, always reply 'No' if the outcome of the coins is heads only. If the outcome of the coin flip is all tails, always reply 'Yes,' irrespective of the true answer. Only if the outcome of the coin flip is mixed, i.e., heads and tails, answer truthfully with 'Yes' or 'No' to the respective question."*

Remember that only the respondent is aware of the outcome of the coin flip and that p_1 = 'forced' 'Yes' = 0.125, p_2 = 'forced' 'No' = 0.125, and p_3 = truthful response = 0.75. The exact instructions used in the questionnaire can be found in Appendix A.1 (translated from German).

The study conducted by Landsheer and colleagues (1999) further shows that a thorough understanding of the method is crucial in order for respondents to actually provide truthful, i.e., potentially socially undesirable, answers. Thus, to ensure respondent understanding of the technique, a minimum of one 'test' example—in which the true answer had been reported by the respondent earlier in the questionnaire—was provided to everyone in this experimental condition so as to familiarize the respondents with the RRT. If this 'test example' was answered incorrectly, or the interviewer was under the impression that the technique had not been fully understood, another standardized example was provided to the respondent. Only when full understanding of the rules had been assured did the main RRT section begin.

2.1.2 The ICT Study

2.1.2.1 Sampling and Data Collection

The ICT study is a based on a general population sample, more precisely, an ADM telephone sample "Easy Sample" (see Häder and Gabler, 1998). We randomly selected German landline numbers, proportional to the population size and regionally stratified by communities. Using random digit dialing (RDD), we also covered individuals whose telephone numbers are not published in telephone books or online. We randomly selected respondents aged 18 to 70 within households, using a 'Kish Selection Grid' (adapted from Kish, 1965, p. 399).

Overall 1,606 interviews were completed and the overall response rate was 15.2% (RR1 according to AAPOR 2011).[4] Unlike in the RRT study, we did not encounter any interview break-offs in the ICT study. Table 2.3 provides a more detailed overview of the sample size and the response rate.

Table 2.3: ICT study: sample sizes and response rates

	General Population Sample	in %
Gross sample	28,128	
Ineligible phone numbers	17,533	
Net sample of phone numbers	10,595	100.0
Unknown eligibility	7,447	70.3
Non-contact	362	3.4
Refusal	908	8.6
Other (speaks no German)	272	2.6
Interview completed	1,606	15.2
Response Rate (AAPOR RR1)		15.2

Of 10,595 telephone numbers, approximately 70% were of unknown eligibility. 3.4% of eligible households could not be contacted, while 8.6% refused participation or did not have sufficient knowledge of the German language (2.6%). Due to the sampling and data collection using RDD, selected individuals in the ICT study did not receive an advance letter announcing the study.

2.1.2.2 Experimental Design & ICT Implementation

Facing the tradeoff of privacy protection and statistical efficiency (see Chapter 1.3.4), we implemented the double-list ICT using three innocuous items (see also Ahart and Sackett, 2004; Coutts and Jann, 2011).

Respondents were randomly assigned to either the control condition—direct questioning—or the treatment condition—item count. Table 2.4 provides an overview of the assignment to the experimental conditions.

4 Due to rounding, the total adds up to 100.1%.

Table 2.4: ICT study: experimental conditions

Study	Assigned Condition	N	Realized Condition	N
ICT study	DQ	500	DQ	500
	ICT 1	550	ICT 1	550
	ICT 2	553	ICT 2	553

Following the same logic as in the RRT study, we also assigned respondents unequally to either condition in order to increase statistical power in the ICT condition. This resulted in 500 complete interviews in the direct questioning condition as well as 550 completes in ICT 1, and 553 in ICT 2 (remember that two subsamples are necessary for the ICT, thus ICT 1 and ICT 2). Due to the double-list implementation of the ICT, this yields a total of 1,103 respondents in the ICT condition. Turning to the 'assigned' and the 'realized' columns, we see that we did not encounter any problems regarding respondents refusing the application of ICT and breaking randomization. Section 2.1.3 provides more information regarding the final sample composition.

2.1.2.3 Questionnaire: General Information

As mentioned above, respondents in both studies—the RRT study and the ICT study—received identical questionnaires, differing only in the experimental sections. Following an initial section concerning labor market biographies, we implemented the experiments and collected data using standard direct questioning or the item count technique.

The ICT experimental section only contained questions relating to undeclared work (yes/no). For the exact wordings of the introduction and the lists in the ICT conditions, please see Appendix A.3. Following our experiments, the interview concluded with fewer, but identical, questions as the RRT study. Average questionnaire completion time was 14.1 minutes in the ICT study.

In the course of questionnaire development, 31 cognitive paper-and-pencil interviews were conducted to identify substantive and methodological problems (e.g., with respect to ICT instructions). After adjusting the questionnaire accordingly, a second pretest (n = 42) was fielded by the data collection institute to test the instruments and all relevant randomizations.

2.1.3 Sample Composition

Table 2.5 provides an overview of the sample composition of both studies. We conducted χ^2 and two-sided t-tests to assess group differences, a summary of these results will be reported in the text.

Table 2.5: Socio-demographic characteristics (ICT and RRT study)

Characteristic	DQ	ICT 1	ICT 2	DQ	RRT	DQ_RRT
	n (percent)	n (percent)	n (percent)	n (percent)	n (percent)	n (percent)
Gender						
Female	274	266	280	631	941	137
	(54.8)	(48.4)	(50.6)	(55.1)	(52.5)	(50.0)
Male	226	284	273	514	851	137
	(45.2)	(51.6)	(49.4)	(44.9)	(47.5)	(50.0)
Age						
Age ≤ 34	112	127	126	385	592	88
	(22.4)	(23.1)	(22.8)	(33.6)	(33.0)	(32.1)
Age 35 to 50	196	206	205	421	650	105
	(39.2)	(37.5)	(37.1)	(36.8)	(36.3)	(38.3)
Age ≤ 50	192	217	222	339	550	81
	(38.4)	(39.4)	(40.1)	(29.6)	(30.7)	(29.6)
Highest Formal Training						
Pupil/No Degree	55	61	64	250	336	70
	(11.0)	(11.1)	(11.6)	(21.9)	(18.8)	(25.9)
Vocational Training	318	340	331	724	1,198	174
	(63.7)	(62.0)	(59.9)	(63.5)	(67.0)	(64.4)
Tertiary Degree	126	147	158	166	253	26
	(25.3)	(26.8)	(28.6)	(14.6)	(14.2)	(9.6)
Employment Status						
No Employment	141	183	183	379	541	123
	(28.2)	(33.3)	(33.2)	(33.1)	(30.2)	(44.9)
Marginally Employed	44	41	35	150	245	31
	(8.8)	(7.5)	(6.3)	(13.1)	(13.7)	(11.3)
Employed > €400	315	326	334	616	1,006	120
	(63.0)	(59.3)	(60.5)	(53.8)	(56.1)	(43.8)
Migrant Background (Either Respondent or Parent Born Outside of Germany)						
No	390	415	429	799	1,287	181
	(78.0)	(75.5)	(77.6)	(70.0)	(72.0)	(66.5)
Yes	110	135	124	343	500	91
	(22.0)	(24.5)	(22.4)	(30.0)	(28.0)	(33.5)

Continued on Next Page

Characteristic	DQ	ICT 1	ICT 2	DQ	RRT	DQ_RRT
Household Size						
Average (s.e.)	2.7	2.7	2.5	2.7	2.7	2.5
	(1.4)	(1.2)	(1.4)	(1.5)	(1.4)	(1.5)
Household Income (Based on Needs-adjusted Equivalence Income)						
Up to €799	33	48	40	354	542	128
	(9.4)	(12.0)	(10.7)	(33.4)	(32.2)	(53.3)
€800 to €1,500	139	142	134	398	597	72
	(39.6)	(35.5)	(35.7)	(37.6)	(35.4)	(30.0)
More than €1,500	179	210	201	307	547	40
	(51.0)	(52.5)	(53.6)	(30.0)	(32.4)	(16.7)
Region of Residence						
West	408	446	446	890	1,365	195
	(81.6)	(81.1)	(80.6)	(77.7)	(76.2)	(71.2)
East	92	104	107	255	427	79
	(18.4)	(18.9)	(19.4)	(23.3)	(23.8)	(28.8)

The composition of the ICT study differs significantly ($p \leq 0.05$) with respect to gender and household size: more precisely, between ICT 1 and ICT 2. For the RRT study, the DQ split and the (realized) RRT split do not differ significantly from each other. Comparing the RRT and the DQ_RRT split, i.e., the 'compliers' and the 'defiers,' these differ significantly ($p \leq 0.10$) from each other with respect to formal training, employment status, migrant background as well as household size or household income. Further analyses, presented in Chapter 3.2, indicate however, that both splits do not differ with respect to central attitudes towards undeclared work. All subsequent analyses for the RRT study will be conducted separately or account for this noncompliance.

Table 2.5 also shows the expected differences in sample composition between both studies: i.e., the register based samples which oversample recipients of basic income support (RRT study) and the general population sample (ICT study). This is most obvious turning our attention to the household income (need based, accounting for household size): The lower income strata is heavily overrepresented in the RRT study.

Further analyses not presented here, replicating Table 2.5 by subsample of the RRT study, confirm our initial expectations regarding differential education, migrant background and language skills (significantly poorer in the benefit recipient sample according to interviewer assessments). While the employee sample and the ICT general population sample are more alike with respect to education (approx. 42% with a higher secondary degree), in the benefit recipient

sample only 20% of the respondents report having an upper secondary degree (see also Aldashev and Fitzenberger, 2009). The same pattern emerges for migrant background, i.e., benefit recipients having a higher share (approx. 35%), while the employee sample and the ICT sample do not differ.

2.1.4 Undeclared Work: The Dependent Variables

In accordance with the existing research practice (Pedersen, 2003; Feld and Larsen, 2005; EC – European Commission, 2007; Schneider and Enste, 2007; Williams, 2009; Boockmann et al., 2010), respondents received a definition of undeclared work just prior to the beginning of that section in the RRT and ICT survey. Both surveys used the following introduction and definition (translated from German):

> *"We would now like to ask you a few questions regarding your experience with undeclared work.*
>
> *By undeclared work we mean any paid labor that is hidden from (tax) authorities, e.g., to avoid paying social security contributions or taxes. Criminal activities, such as drugtrafficking, are NOT included in the definition of undeclared work."*

Following this introduction, respondents were then asked to answer specific items relating to undeclared work. Both surveys differentiated between undeclared work for a private individual and undeclared work for a company (see EC – European Commission, 2007; Pfau-Effinger, 2009; Boockmann et al., 2010). Table 2.6 provides information regarding the exact wordings, sample sizes in each condition, as well as item nonresponse.

Table 2.6: Wording of the items measuring undeclared work, sample sizes and item nonresponse (translated from German)

Item	Operationalization	Study	Valid Responses	Item Nonresponse	N
Person	Have you engaged in any undeclared work for a private person this year?	RRT	DQ: 1,142	3	1,145
			RRT: 1,790	2	1,792
			DQ_RRT: 272	2	274
		ICT	DQ: 498	2	500
			ICT: 1,103	0	1,103
Company	Have you engaged in any undeclared work this year for a company, which paid you without reporting your income to the authorities?	RRT	DQ: 1,142	3	1,145
			RRT: 1,789	3	1,792
			DQ_RRT: 273	1	274
		ICT	DQ: 497	3	500
			ICT: 1,103	0	1,103

Response options to both questions were either 'Yes' or 'No' in the RRT study or rather the number of applicable items in the ICT study. As briefly outlined in Section 1.3.1, we expect to see very little evidence of item nonresponse in either experimental condition. Refusing to answer "may be more awkward than simply underreporting" undeclared work (Groves et al., 2004, p. 240). These items will form the main dependent variables of interest throughout the remainder of this chapter.

2.2 Empirical Results

2.2.1 Comparing Randomized Response, Item Count and Direct Questioning

This section addresses the first research question, analyzing whether we can improve measurement of undeclared (under the 'more-is-better' assumption) and how the RRT performs in different subpopulations.

Figure 2.1: Undeclared work for a private person: prevalence estimates (in %) and 95% confidence intervals

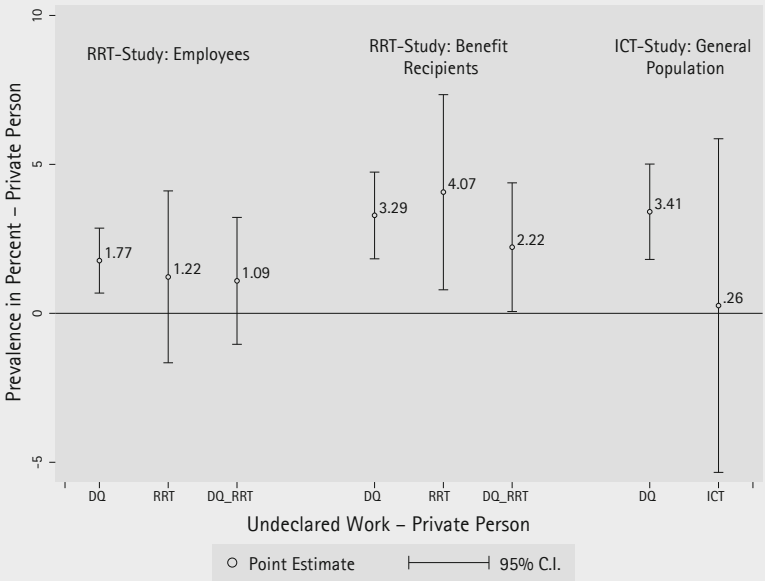


Figure 2.1 displays the prevalence estimates (in %), and the corresponding 95% confidence intervals, for the item 'undeclared work for a private person' (y-axis) by study for each experimental condition (x-axis). Contrary to our initial expectations, the prevalence estimates obtained using the RRT or the ICT are not consistently

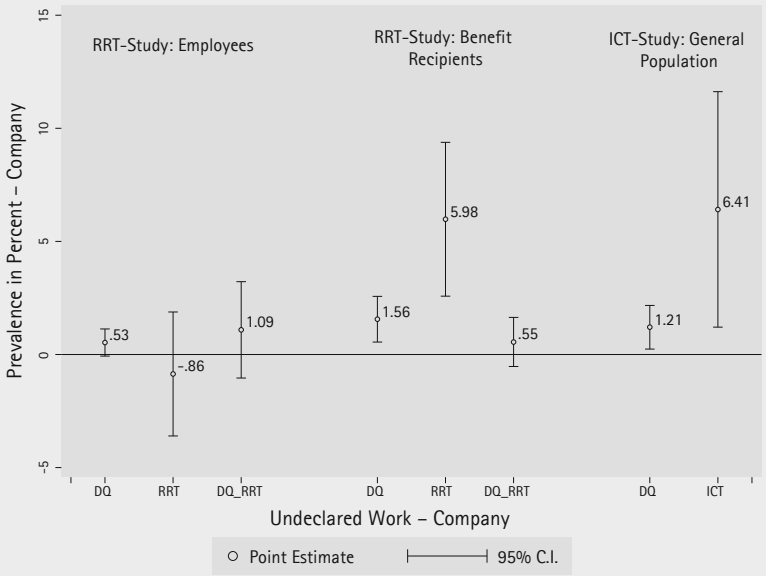
higher compared to the direct questioning conditions in either study. Only for the benefit recipients sample in the RRT study are results as expected, with the RRT eliciting more (truthful) reports of undeclared work (DQ: 3.29%; RRT: 4.07%). Estimates in the employee sample of the RRT study, as well as estimates in the ICT study are even lower in comparison to the traditional direct questioning (RRT study: 1.77% vs. 1.22%; ICT study: 3.41% vs. 0.26%). Neither differences between experimental and control groups are statistically significant in either study.[5] Those respondents who refused the application of the RRT do not differ significantly from the respondents in the corresponding RRT condition (employees: 1.09%; benefit recipients: 2.22%).

Overall, Figure 2.1 also reveals that respondents in the benefit recipients sample of the RRT study report slightly higher levels of undeclared work for a private person, compared to respondents in the employee sample. Taking all evidence into account, estimates obtained for the general population in the ICT study fall within the range of the former two estimates. Further, Figure 2.1 shows that the 95% confidence intervals are much broader in the ICT condition compared to those in the RRT condition. While there is evidence that ICT estimates are more inefficient compared to RRT estimates (Holbrook and Krosnick, 2010a, b; Coutts and Jann, 2011), in our study it is primarily due to the smaller sample size in the ICT condition.

Figure 2.2 again presents prevalence estimates (in %) and the corresponding 95% confidence intervals, for all experimental conditions, however, for the item 'undeclared work for a company.' As in Figure 2.1, respondents in the benefit recipient sample of the RRT study display higher prevalence rates in the RRT condition compared to direct questioning (DQ: 1.56%; RRT: 5.98%). This finding is significant at the 1% level. In the employee sample, differences in the RRT and direct question condition are nonsignificant: Again, contrary to our expectations the prevalence estimates obtained in the RRT condition are even lower compared to direct questioning (DQ: 0.53%; RRT: -0.86%). While it is not only lower, the prevalence estimate in the RRT condition is even negative (though nonsignificantly different from zero). Turning to the ICT study, Figure 2.2 shows that the prevalence estimates obtained via ICT are significantly higher at the 5% level compared to direct questioning (DQ: 1.21%; ICT: 6.41%). For this item, those respondents who refused the application of the RRT differ (significantly) from the respondents in the corresponding RRT condition and are more similar to those of the DQ condition (employees: 1.09%; benefit recipients: 0.55%).

5 Empirical significance levels (p-values) are based on two-sided z-tests (RRT) using the rrlogit routine, as well as t-test for differences in mean estimates (ICT).

Figure 2.2: Undeclared work for a company: prevalence estimates (in %) and 95% confidence intervals



Similarly to the results for the first item, estimates for undeclared work for a company in the benefit recipients sample are consistently higher compared to the employee sample of the RRT study. While results in the ICT study were slightly higher compared to those of the benefit recipient sample for the first item, they are now in the same range.

Summing up the results for the first research question: neither in Figure 2.1 nor in Figure 2.2 do we see the expected results. Both methods, the ICT and the RRT (benefit recipients), provide more 'valid' results only for the item 'undeclared work for a company' ('more-is-better' assumption). In other instances, neither the RRT nor the ICT result in consistently higher prevalence estimates compared to direct questioning. In the light of these results, the question arises of why we observe higher prevalence estimates in some instances, but not in others?

First, one potential explanation might be that undeclared work for a company is perceived as the more sensitive item of the two, given that it is a more 'quasi-institutionalized' form of employment. Looking into the German penalty code, in general, any offense against the German SchwarzArbG is treated as a regulatory offense, but also as a criminal offense. It can be penalized with fines up to €300,000 (penalties and back duties) for 'employers' and result in up to five years of prison for both 'employers' and 'employees' (§8 SchwarzArbG, 2004). Aside from the quantity of the offense, the underlying intention is crucial for the criminal sentence (Feld et al., 2007): It makes a difference if the nature of the offense is

'unwitting' tax evasion (e.g., concealing of capital income) or a more deliberate, organized form of tax evasion (combined with abuse of authority or fraudulent counterfeit) (Zoll, 2011, 2012). If conducted in a professional or organized form, the possible sentence increases from a minimum of one year of imprisonment to a maximum of 10 years (§370a Abgabenordnung; see also Feld et al., 2007). Another potential explanation in that regard might be the fact that undeclared work for private persons or households is socially much more accepted, less often concealed and more often only prosecuted as a regulatory offense, compared to undeclared work for a company, i.e., in the 'official economy' (Pfau-Effinger, 2009, p. 89).

Second, for respondents in the benefit recipient sample, yet another concern adds to perceived item sensitivity: Not only are benefit recipients prosecuted for engaging in undeclared work if discovered (§9 SchwarzArbG, 2004), but also for social security fraud (§263 StGB). This offense can lead to up to five years of imprisonment or a penalty (Zoll, 2011, 2012). Thus it is reasonable to assume that the items measuring undeclared work are perceived as even more sensitive for benefit recipients.

In accordance with other studies (Lensvelt-Mulders et al., 2005a) that argue that these special techniques perform better, the more sensitive an item is perceived to be by respondents, our results are as expected: These special techniques outperform direct questioning 1) in the benefit recipients sample compared to the employees and 2) for the item 'undeclared work for a company' compared to 'undeclared work for a private person.'

However, we are not solely interested in the relative comparisons of the different data collection techniques. We are also interested in how our prevalence estimates relate to those of other studies. To ensure comparability, it is sensible to limit our comparison to the most recent German studies. Since those studies do not differentiate between undeclared work for a private person and for a company, we have to aggregate our estimates to obtain an 'overall prevalence' of undeclared work. Furthermore, studies that qualify for this comparison should have approximately the same reference periods and a similar population. Two studies, with reference periods of one year, surveying the German general population, meet these criteria: the Eurobarometer, referring to a population aged 15 and older (Williams, 2009) and the study by Feld and Larsen (2008) with a population aged 18 to 74.

Estimating an overall prevalence, i.e., combining individual reports for the two kinds of undeclared work, direct questioning yields 1.9% in the employee sample, 3.8% in the benefit recipient sample, and 4.2% in the ICT general population survey. The matching estimates in the experimental conditions have to be derived from the indirect data: For the employee sample, the RRT estimate is at least 1.2%, while the same estimate for the benefit recipient sample is approximately

9.9%. The ICT estimates have a lower bound of at least 6.4% undeclared work.[6] Adjusting for the fact that both studies, the RRT and the ICT study, were fielded in the months of October and November 2010, the overall estimate for 2010 will likely be even higher than our actual estimates. The Eurobarometer estimates the prevalence of undeclared work for the year of 2007 at 3% (EC – European Commission, 2007, p. 19; Williams, 2010, p. 254); an estimate that closely relates to what we observe in our direct questioning conditions. The estimated prevalence in the study conducted by Feld and Larsen (2008) is 7.2% (in 2006).

2.2.2 Who Engages in Undeclared Work?

This section addresses the second research question, analyzing individual characteristics that foster engagement in undeclared work and the main motivations of respondents to do so. Drawing on existing theories explaining engagement in undeclared work, we will empirically test these hypotheses with our data. Given the lower statistical power of the ICT study, and the lack of available data analysis routines to analyze ICT double-list data, we will conduct all subsequent analyses using solely data from the RRT study. The R package "list: Statistical Methods for the Item Count Technique and List Experiment" is only available for the analysis of single-list ICT variants (Blair and Imai, 2012).

Before we do so, we will briefly review some theories of undeclared work and tax evasion/honesty and present the hypotheses to be tested. The hypotheses are usually derived from behavioral theories, e.g., using the framework of rational choice theory and subjective expected utility theory of the income-reporting decision (Becker, 1968; Voss and Abraham, 2000; Mehlkop and Becker, 2004; Eifler, 2009).

2.2.2.1 Theoretical Foundations and Hypotheses

One of the very first theoretical approaches explaining deviant behavior regarding paying taxes is the standard model of tax avoidance and tax evasion (Allingham and Sandmo, 1972). It particularly focuses on the influence of tax rates. The authors also investigate the impact of the probability of a tax audit, i.e., of being detected (the model assumes that this probability is known to the individuals), and the influence of the degree of punishment if individuals are discovered.

Bordignon (1993, p. 345), however, criticises that: 1) this standard deterrence model cannot account for the empirically observed 'compliance': The share of

6 The overall prevalence rates in the control groups can be obtained by simply adding the single estimates and subtracting the intersection (to obtain the set union). For the RRT we obtain an overall estimate according to the conditional probabilities (however, we did censor the negative prevalence estimate to the lower bound of zero). For the ICT, we rely on those (higher) estimates of the second item only, since an overall estimate cannot be computed.

individuals abiding by the laws and regulations regarding reporting of income is too high and cannot be explained with the standard model. 2) Some of the hypothesized relationships, e.g., that the level of evaded taxes is negatively related to the tax rate, are implausible and are not supported by empirical findings (Yitzhaki, 1974; Andreoni et al., 1998). 3) Last but not least, the standard model does not withstand an empirical test using either survey data or experiments (Dell'Anno, 2009).

Thus, over the past decades, these initial models were extended, allowing more integrated approaches and realistic assumptions. Extensions of the model incorporate characteristics of the specific legal tax code, enforcements and regulations (Yitzhaki, 1974). Also, assumptions that do not hold, such as 'complete' information regarding the probability of tax audits were relaxed. Another development (see Andreoni et al., 1998) relates to the inclusion of 'soft factors,' such as social norms and nonmonetary costs that arise from social stigma and loss of reputation (Benjamini and Maital, 1985), tax morale, moral obligations and intrinsic motivations to comply with norms (Gordon, 1989; Myles and Naylor, 1996), fairness considerations with respect to the tax burden as well as the satisfaction with the manner in which taxes are spent (Spicer and Lundstedt, 1976; Bordignon, 1993).

Accounting for individual restrictions and opportunity structures (cf. also Renooy, 1990), these integrated approaches provide the starting point for an explanation of deviant behavior. Dell'Anno (2009) essentially integrates two major directions of research: the perceived risk of detection by individuals and (subjective) expected utility from behavioral economics, as well as individual tax morale, defined as intrinsic motivation to pay taxes. If individuals are caught cheating on taxes, psychological costs due to stigma and loss of reputation arise that have to be included in the model (see also Gordon, 1989). A thorough review of the theoretical developments can be found in Cowell (1990) as well as Andreoni et al. (1998).[7]

Based on these arguments, we draw upon classical and more recent behavioral theories to explain delinquent behavior such as undeclared work (Becker, 1968; Voss and Abraham, 2000; Mehlkop and Becker, 2004; Schneider and Enste, 2007). According to these theories, the individual definition of a situation and the subsequent behavior, are influenced by an individual's prospect of success and the expected utility given 'success.' Furthermore, the anticipated probability of being detected engaging in criminal behavior, as well as the associated costs given detection contribute to this subjective evaluation. Other equally important

7 Similar to Dell'Anno (2009, p. 993), we apply these theories of tax avoidance and tax evasion to undeclared work. It can be reasonably assumed that we are dealing with similar types of individuals engaging in shadow economic activities and that the same arguments apply in both instances. Further theoretical approaches can be found in Renooy (1990), Hessing et al. (1993), Lamnek et al. (2000), Mummert and Schneider (2001), Wenzel (2004), Feld and Larsen (2005), as well as Schneider and Enste (2007, p. 54 ff.).

factors include prevailing social norms in an individual's environment and external restrictions, such as time and budget restrictions (Mehlkop, 2011).

From this rational choice perspective, we thus assume that a variety of factors contribute to the explanation of engagement in undeclared work (participation decision), namely: utility from undeclared work, i.e., expected monetary gains and perceived costs of undeclared work, opportunity structures as well as norms and values, i.e., general approval or disapproval of undeclared work.

Monetary gains that an individual receives from undeclared work depend on the potential 'additional' income (Tanzi and Shome, 1993, p. 811). Two mechanisms are plausible in order to generate this additional income: First, undeclared work can substitute official employment. An individual working undeclared then saves taxes and social security contributions and thus generates more income (provided that she does not receive social security benefits which would offset this additional income). Since these social security contributions would, in theory, benefit the individual herself, we will instead focus on the marginal tax rate. For employed individuals this ranges—depending on the annual income—from 0% (income below €8,000) to 42% (above €52,500). These gains from substitution are even more pronounced for recipients of welfare benefits (UB II). Any additional income (if declared) is deduced: Above the basic exempt amount of €100, the marginal benefit 'deduction' or 'withdrawal' rate (i.e., the amount that will be credited against the benefits) ranges between 80% and 90% of each additional Euro that is earned by an individual on a monthly basis.[8] Our data does not allow the derivation of an accurate measure of the marginal tax rate or the marginal 'deduction' rate, thus the sample status[9] in combination with the gross labor income will be used a proxy for gains by substitution. We hypothesize that individuals receiving unemployment benefits II, as well as individuals with a high labor income (Andreoni et al., 1998) will be more likely to engage in undeclared work, all other things being equal.

Another 'gain' from undeclared work is by means of classical 'moonlighting,' i.e., additional income from a complementary side job. This scenario is plausible if the preferred scope of work (amount of hours worked) cannot be achieved in the official labor market. Sometimes regulations, such as general agreements on pay grades (*Rahmentarifvertrag*), restrict individual preferences by limiting the number of working hours, or even force individuals to work part-time. The preferred

8 At the time of the survey unemployment benefit II recipients were allowed to keep 20% of their additional income between €101 to €800 and 10% of each Euro earned above €801 to a maximum of €1,500.

9 The survey contains another randomized response variable indicating current or past receipt of benefits. Thus, we rely on the sample indicator as a proxy (see Chapter 4). Given that respondents in the benefit recipient sample were known to have received benefits just prior to the main data collection, and further, that the benefit receipt among the respondents in the employee sample during the data collection period is as low as 3.8%, we assume that this proxy can be used.

number of working hours per week in relation to the actual number of hours worked by individuals is used as an indicator to capture this effect. We hypothesize that individuals who are (externally) constrained in their preferred number of working hours, will be more likely to engage in undeclared work, all other things being equal.

Monetary costs on the other hand are a combination of the two factors 'perceived risk of detection' and the 'expected penalty' given detection (see Andreoni et al., 1998; Pedersen, 2003; Feld and Larsen, 2005). Formally, the expected value of these costs is derived by multiplying both factors. We hypothesize that the higher the expected value of this product for a given offense, the lower the probability of engagement in undeclared work, all other things being equal. This proxy was generated using the following information: First, respondents were asked to estimate how many out of 100 individuals who work undeclared, would be discovered by the authorities. Second, respondents were given a hypothetical situation of an individual engaging in undeclared work (for six months, full-time, earning overall €12,000 with charges being pressed against this person). We asked respondents to provide an estimate for the monetary penalty that the hypothetical individual would be facing (we explicitly excluded back duties in our example).

Until now, the focus of our explanation to engage in undeclared work has been exclusively on the individual's decision to supply labor. However, the decision to participate in undeclared work also depends on the demand. *Opportunity structures*, such as type of job and how well an individual is connected or integrated, are thus essential for the explanation of undeclared work. Previous studies show, for example, that those types of employment that are more 'independent,' such as self-employment, lead to a higher propensity to engage in undeclared work (Andreoni et al., 1998; Feld and Larsen, 2005; Williams, 2009). Our measure of occupation, or rather a proxy for independence and opportunities either on the current or the previous job, relies on the occupational status adapted from the Erikson, Goldthorpe and Portocarero social class scheme (EGP; Erikson et al., 1979). Another aspect that influences opportunities to engage in undeclared work is individual integration and networks. While individuals who are socially integrated have more social obligations, and thus potentially less time to engage in undeclared work, they also have more opportunities and an exchange of information for doing so (Granovetter, 1974; Wolff, 1991; Williams, 2010). To capture these effects, we collected data on the amount of support a respondent would have in finding a job. This is measured as the number of people in one's network willing to provide this support resource. Another indicator for integration is active membership in an organization or club. We hypothesize that respondents who are more socially integrated, or have occupations that are characterized by more independence and opportunities, have a higher propensity to engage in undeclared work.

Finally, *normative considerations* such as informal normative expectations or internalized social norms are relevant in explaining deviant behavior such as engagement in undeclared work (Gordon, 1989; Tanzi and Shome, 1993; Cullis and Lewis, 1997; Andreoni et al., 1998; Falk, 2003; Feld and Larsen, 2005; Cialdini, 2007). On the one hand, if respondents (as well as their friends or acquaintances) disapprove of undeclared work, they will be less likely to rationally consider the potential tradeoff between official and undeclared work and the potential utility of engaging in undeclared work. Andreoni et al. (1998, p. 846), for example, argue for the "human need for consistent self-representation" and if honesty norms are internalized, for an individual tendency for norm-congruent behavior. Having said that, psychological costs due to noncompliance to one's own moral standards can be incorporated into the model. On the other hand, compliance with these internalized moral standards can be countered or weakened by "social information" (Cialdini, 2007, p. 264). This could be the case if an individual's social network mostly approves of undeclared work—or rather if the individual anticipates approval ("injunctive social norm")—, or if undeclared work is perceived as a common behavior that is accepted as something 'normal' ("descriptive social norm").

(Social) norms regarding undeclared work and values transmitted in one's social networks (societal norms of behavior) are captured using several indicators: 1) an estimate of the prevalence of undeclared work among one's friends and acquaintances, and 2) an index measuring attitudes towards undeclared work. This additive index is composed of seven attitudinal questions related to tax morality, perceived regulation density, or intolerance of undeclared work. Response formats for each of the seven items were identical using a 4-point rating scale (if necessary, recoded for the analyses to express positive attitudes).[10] High values on our scale indicate a positive, approving attitude towards undeclared work. We hypothesize that the higher the expressed approval of or the higher the share of undeclared work in one's own network, the higher the individual propensity to engage in undeclared work.

A detailed overview of all relevant operationalizations can be found in the Appendix A.5.

To summarize all hypotheses to be tested:

1. Individuals receiving welfare benefits, earning a high labor income as well as individuals who are limited in their preferred number of working hours have a higher individual propensity to work undeclared. (Utility Hypothesis)
2. Individuals with a higher perceived risk of detection and higher expected penalties have a lower individual propensity to work undeclared. (Cost Hypothesis)

10 Comparing sum scores with factor scores (derived using maximum likelihood estimation) shows a high correlation ($r = 0.95$; irrespective of whether we listwise delete observations with missing values or use imputation techniques). Cronbach's Alpha is 0.56 for the seven items.

3. Individuals who are socially more integrated (network resources for finding a job or active membership in an organization), or who are more independent in their occupation have a higher individual propensity to work undeclared. (Opportunity Hypothesis)
4. Individuals who express approval towards undeclared work, or who report a higher perceived share of undeclared work among their friends and acquaintances have a higher individual propensity to work undeclared. (Norm Hypothesis)

Furthermore, all models control for demographic characteristics such as gender, age, formal training (highest degree), migrant background and region of residence (East/West) (see Wolff, 1991; Merz and Wolff, 1993; Andreoni et al., 1998; Mummert and Schneider, 2001; Schneider and Enste, 2007; Lago-Peñas and Lago-Peñas, 2010).[11]

2.2.2.2 Empirical Evidence

All subsequent analyses to empirically test our hypotheses were conducted using logistic regression models (using rrlogit Jann, 2011 with a modified likelihood function, see Maddala, 1983). In order to address the second research question, two models will be estimated: Model 1) modeling undeclared work for a private person, as well as model 2) modeling undeclared work for a company, each as a function of the covariates introduced above. In order to relax assumptions regarding the exact nature of the relationships and given the relatively low power of the RRT estimator, we categorized most independent variables into three or four disjunct categories, each using empirical terciles or quartiles.[12] Further, to increase statistical power, we pooled data from both samples, i.e., the employee and the benefit recipient sample. Due to the low overall prevalence, we refrained from adjusting our models to include survey weights and did not model any interaction effects[13].

In order to deal with item nonresponse in the independent variables, we have generally imputed missing values using the hotdeck method (Mander and Clayton, 1999) with the exception of the 'attitudes towards undeclared work' index. For this index, each missing value (on an item level) was imputed using a mean imputation. Hotdeck imputation methods replace missing data with an randomly selected

11 Including controls seems appropriate given noncompliance to the randomization.

12 Using alternative thresholds for categorizations does not alter our substantive conclusions.

13 Including interaction terms would allow us to relax the assumption that the parameter estimates are identical under the DQ and the RRT condition. Running two separate models for each experimental condition to investigate differential relationships largely supports these assumptions, see Appendix A.7); yet also reveals some estimation problems: In the 'DQ company model' three covariate patterns (self-employed, manual supervisor and skilled manual, tertiary degree) predict the outcome variable perfectly. This problem of separation is essentially a problem of sample size (Zorn, 2005, p. 161). Several strategies to deal with this problem exist (Zorn, 2005, p. 161 f.). 1) Omitting these covariate patterns. 2) Supplementing the data with 'artificial' data. 3) Using exact logistic regression. Applying the first, most commonly used strategy—omission—, results are displayed in Table A.5 of Appendix A.7).

observed record of complete data. We formed adjustment cells based on theoretical considerations regarding association of auxiliary variables with each missing variable to be imputed as well as the mechanism fostering item nonresponse (Little and Vartivarian, 2005). More precisely, item-nonresponse was affecting items as follows: item 'perceived risk of detection' 4.3%; 'expected penalty' 10.0%; 'prevalence of undeclared work in one's network' 5.7%; 'networks job search' 4.7%; for all remaining variables it was lower than 2.7%. Sensitivity analyses replicating the analyses using listwise deletion, support our main conclusions.[14]

Table 2.7 displays average marginal effects (AME) based on derivatives resulting from our two logistic regression models (Stata version 12.1, rrlogit, Jann, 2011) and the 95% confidence intervals (Long, 1997; Allison, 1999; Bartus, 2005; Mood, 2010). Empirical significance levels (p-values) are based on two-sided z-tests.

According to the *utility hypothesis*, we expect to see an increased probability of undeclared working by individuals with either a high marginal tax or benefit 'deduction' rate, as well as those individuals who are constrained in their amount of working hours. While individuals who cannot realize their preferred working hours in the official economy are more likely to engage in undeclared work compared to individuals perceiving it as adequate in both models, effects are modest and statistically nonsignificant.

Regarding the marginal tax and benefit 'deduction' rates, we observe that the direction of the effects is as predicted. In both models, benefit recipients with an income above €800 (90% deduction rate) are on average most likely to engage in undeclared work, while those with an income of €800 or less (80% deduction rate) have the second highest probability. Merely the results of employees with a high income (higher marginal tax rates) and those with a lower income (lower marginal tax rates) across models are not consistent with our prediction.

Testing all possible contrasts of our combined utility indicator (sample and income)[15], for the first model (private person), only the differences between benefit recipients with an income above €800 compared to employees with an income up to €800 (4.0%pts, $p = 0.05$) or of €800 and above (3.4%pts, $p = 0.02$) are statistically significant.

---

14 Results displaying these models can be found in Appendix A.6. The main effects are essentially the same and only slight differences between the models can be observed: While the AME of the items 'high risk perception' (-3.4%pts, $p = 0.11$), 'manual supervisor and skilled manual workers' (2.4%pts, $p = 0.15$) as well as 'organizational membership' (1.9%pts, $p = 0.10$) are almost identical compared to the model based on imputed values, these effects are statistically nonsignificant under this model.

15 Our analyses include mainly categorical variables. Table 2.7 only displays results for both models with respect to one reference category. However, we replicated our models using all possible contrasts. AME's for these contrasts are not displayed, but can easily be obtained from Table 2.7 using subtraction. We will refer to these AME's and the p-values in the text only, if they are statistically significant. Due to rounding, reported AME's can differ up to 0.001 from those derived from Table 2.7 by subtraction.

Table 2.7: Logistic regression models analyzing undeclared work (average marginal effects and 95% confidence intervals)

Y: Undeclared Work for a ...		Model 1: Private Person	Model 2: Company
		AME	AME
		[95% C.I.]	[95% C.I.]
Methods Effect			
Experimental Condition (ref. DQ)	RRT	**0.024***	**0.047*****
		[0.002,0.045]	[0.023,0.072]
	DQ_RRT	-0.008	-0.007
		[-0.040,0.025]	[-0.046,0.032]
Utility Hypothesis			
Utility (ref. UB II w. Income ≤ €800)	Employee w. Income ≤€800	-0.030	-0.010
		[-0.067,0.006]	[-0.039,0.019]
	Employee w. Income > €800	-0.025	**-0.049***
		[-0.054,0.005]	[-0.095,-0.004]
	UB II w. Income > €800	0.010	0.010
		[-0.018,0.037]	[-0.022,0.043]
Pref. Working Hours (ref. Adequate (≤ 2))	Inadequate (≥ 3)	0.008	0.004
		[-0.016,0.031]	[-0.023,0.031]
Cost Hypothesis			
Risk Perception (ref. €0 to €120)	Low Risk (€121 to €600)	0.007	0.003
		[-0.015,0.029]	[-0.023,0.029]
	Medium Risk (€601 to €2,500)	-0.002	-0.001
		[-0.025,0.022]	[-0.025,0.024]
	High Risk (> €2,500)	**-0.029**⁺	-0.002
		[-0.063,0.005]	[-0.026,0.023]
Opportunity Hypothesis			
Occupational Status (ref. (Semi)Unskilled Manual)	N/A (Never Employed)	-0.016	-0.004
		[-0.050,0.017]	[-0.029,0.022]
	Low-High Controllers	-0.001	0.008
		[-0.031,0.030]	[-0.024,0.039]
	Routine Non-Manual	-0.006	-0.006
		[-0.033,0.022]	[-0.033,0.020]
	Self-Employed	0.020	-0.020
		[-0.024,0.064]	[-0.089,0.050]
	Manual Supervisor & Skilled Manual	**0.025**⁺	-0.014
		[-0.001,0.051]	[-0.048,0.020]
Networks Job Search (ref. Nobody)	1 to 4 Persons	**0.031***	0.005
		[0.002,0.059]	[-0.021,0.031]
	5 to 10 Persons	**0.038****	0.017
		[0.011,0.064]	[-0.005,0.040]
	≥ 11 Persons	**0.045****	0.011
		[0.014,0.076]	[-0.017,0.040]

Continued on Next Page

Y: Undeclared Work for a ...		Model 1: Private Person	Model 2: Company
Membership in an Organization (ref. None)	At Least 1	**0.018**$^{+}$	0.009
		[-0.001,0.037]	[-0.010,0.027]
Norm Hypothesis			
Undeclared Work in Network (ref. Nobody)	1 to 3%	**0.055***	0.011
		[0.013,0.096]	[-0.021,0.042]
	4 to 10%	**0.061****	0.013
		[0.022,0.101]	[-0.018,0.045]
	≥ 11%	**0.078*****	**0.045****
		[0.040,0.117]	[0.017,0.072]
Approval of Undeclared Work (ref. Disapproval 7 to 14)	Some Disappr. (15 to 16)	0.008	0.011
		[-0.032,0.048]	[-0.017,0.040]
	Some Appr. (17 to 19)	0.022	0.001
		[-0.014,0.058]	[-0.028,0.031]
	Approval (20 to 28)	**0.047****	0.018
		[0.014,0.080]	[-0.008,0.044]
Controls			
Gender (ref. Male)	Female	-0.001	**-0.019**$^{+}$
		[-0.021,0.019]	[-0.040,0.002]
Age (ref. 35 to 49)	Age ≤ 34	0.012	**0.030***
		[-0.009,0.034]	[0.007,0.052]
	Age ≥ 50	0.015	-0.039
		[-0.010,0.040]	[-0.096,0.019]
Formal Training (ref. Vocational Training)	Pupil/No Degree	0.005	-0.002
		[-0.020,0.029]	[-0.024,0.019]
	Tertiary Degree	**-0.064***	-0.025
		[-0.125,-0.003]	[-0.085,0.035]
Migr. Background (ref. None)	Yes	-0.006	-0.003
		[-0.027,0.015]	[-0.023,0.017]
Residence (ref. West Germany)	East Germany	-0.004	-0.001
		[-0.026,0.019]	[-0.023,0.022]
Model Fit			
N		3204	3205
LR Chi^2 (df)		125.161 (31)	90.014 (31)
Pseudo R^2		0.07	0.06
AIC		1733.906	1595.924
BIC		1922.143	1784.171

95% confidence intervals in brackets; $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Turning to the second model (company), benefit recipients—irrespective of their income—are, on average, significantly more likely to engage in undeclared work only when compared to employees with an income of more than €800 (≤€800: 4.9%pts, p = 0.03; >€800: 5.9%pts, p < 0.01). Employees with an income of up to €800 are significantly less likely to engage in undeclared work for a company compared to those with a higher income (3.9%pts, p = 0.07).

Overall, we find only weak empirical evidence for our utility hypothesis: The only statistically significant differences are found between benefit recipients and employees.

The *cost hypothesis* states that individuals with a higher risk perception (perceived risk of detection and perceived penalty) are less prone to work undeclared compared to those with a lower risk perception. While this hypothesis is supported by our first model (private person)—individuals with a high perceived risk (>€2,500) are, on average, less likely to engage in undeclared work compared to individuals with a very low (€0–€120: -2.9%pts, $p = 0.09$) or a low risk perception (€121–€600: -3.6%pts, $p = 0.04$)—we do not observe the same effect in our second model (company). All AME's in the second model are approximately zero.

According to the *opportunity hypothesis*, we expect to see that more 'independent' occupations as well as larger networks and integration will foster engagement in undeclared work. Our results suggest that manual supervisors and skilled manual workers are on average more likely to engage in undeclared work for a private person than individuals who are semi-skilled or unskilled manual workers (2.5%pts, $p = 0.06$). A similar result holds when comparing them with individuals who have never been employed before (4.2%pts, $p = 0.03$) as well as with routine non-manual workers (3.1%pts, $p = 0.04$). Due to the low statistical power, even average marginal effects that are well beyond zero—e.g., for self-employed compared to semi-skilled or unskilled manual workers (2.0%pts, $p = 0.37$)—are statistically nonsignificant. Our first model shows that, on average, larger labor-market-related networks—with at least one person compared to knowing nobody to help with the job search (>3.1%pts, $p \leq 0.04$)—as well as active membership in an organization (1.8%pts, $p = 0.06$) foster engagement in undeclared work. None of these effects within the opportunity hypothesis can be replicated in the second model (company), i.e., none of the expected effects are statistically significant.

Our results strongly support the *norm hypothesis*: a low perceived prevalence of undeclared work in one's network (≥ 1%: 5.5%pts, $p \leq 0.01$), as well as an approving attitude towards undeclared work (score 20–28: 4.7%pts, $p < 0.01$), increases the probability to engage in undeclared work for a private person. The last relationship does not only hold in comparison to the reference category: An approving attitude (score 20-28) also increases the average probability of engaging in undeclared work with reference to both middle categories (score 15–16: 3.9%pts, $p < 0.01$; score 17–19: 2.5%pts, $p = 0.02$). The same can be said about perceived prevalence: Those individuals with a high perceived prevalence (≥ 11%) are on average more likely to engage in undeclared work compared to those who suspect a lower share (1-3%: 2.4%pts, $p = 0.07$).

These empirical results for the norm hypothesis can also be supported with our second model: Individuals who suspect a large share of undeclared workers in their networks (≥ 11%) are significantly more likely themselves to engage in undeclared work compared to individuals who have a perceived prevalence of zero (4.5%pts, $p < 0.01$). Again, this effect is significant also with respect to the two middle categories (1–3%: 3.4%pts, $p = 0.01$; 4–10%: 3.1%pts, $p = 0.01$).

Regarding the *control variables*, our models show that there are no gender differences in the first model, while women on average have a significantly lower probability of engaging in undeclared work for a company (AME = -1.9%pts, $p = 0.07$). Compared to individuals aged 35 to 49, younger respondents (34 and younger) as well as older respondents (aged 50 and older) do not differ significantly in model 1, while individuals aged 34 and younger significantly more likely engage in undeclared work for a company compared to the reference category (≤ 34: 3.0%pts, $p = 0.01$; ≥ 50: -6.8%pts, $p = 0.02$). Furthermore, individuals with a tertiary degree are on average significantly less likely to engage in undeclared work for a private person, compared to individuals with some form of formal training (-6.4%pts, $p = 0.04$) or individuals who have no degree or are students (-6.9%pts, $p = 0.04$). The type of degree does not make a difference for undeclared work for a company. Other controls such as migrant background or place of residence do not exert a significant influence in either model.

The *methods effect* is as expected in both models (model 1: 2.4%pts, $p = 0.03$; model 2: 4.7%pts, $p < 0.01$). Controlling for other covariates and pooled across both samples, respondents in the RRT condition are significantly more likely to report engagement in undeclared work for a private person or for a company compared to respondents in the DQ condition. This effect is slightly larger in the second model compared to the first model. Respondents not complying with the RRT request are more similar to respondents in the DQ condition with respect to reporting of engagement in undeclared work in both models. These differences are not significant.

Summing up the results for our second research question: We find more empirical support for our hypotheses regarding the explanation of undeclared work for a private person and less for a company. This could be due to the lower overall prevalence in the second model, which leads to smaller AMEs and in turn to less statistical power to detect effects. Nonetheless, the direction of the effects is predominantly the same across both models. Especially our expectations with respect to the opportunity and norm hypotheses were supported by our data and are significantly related to undeclared work for a private individual, as well as for a company.

2.3 Discussion and Conclusion

This chapter had two main goals: a methodological one and a substantive one. Methodologically, we were interested in whether we can obtain more accurate self-reports—by those respondents engaged in undeclared work—of undeclared work in large scale population surveys on the telephone. Given the associated problems due to the sensitive nature of the questions, we used special data collection techniques, that scramble the individual response process and increase anonymity. Substantively, we were interested in investigating individual motivations and opportunity structures fostering engagement in undeclared work.

The first central result is that neither the randomized response, nor the item count technique outperform standard direct questioning. In two out of six tests we saw the expected result, i.e., the RRT or the ICT producing significantly (on the 5%-level) higher prevalence estimates compared to direct questioning. In the remaining four tests we did not find any significant effects. Our results are in accordance with the literature (Lensvelt-Mulders et al., 2005a), showing that these techniques perform the better the more sensitive the item is potentially perceived to be by a respondent (higher expected penalties, i.e., undeclared work for a company or the benefit recipients sample).

Furthermore, Coutts and Jann (2011) argue that we can never know for sure if respondents actually picked up a coin and/or responded according to the (RRT) instructions. Our negative prevalence estimate in one of the RRT conditions does suggest that not all respondents complied with the instructions. At this point, we can only speculate about the potential reasons for noncompliance. Taking all evidence into account and analyzing complete and partial interviews, a total of 369 respondents refused the application of the RRT (15.8% of complete and partial interviews). This high noncompliance rate as well as increased costs due to a more complex and longer data collection process, an increased respondent burden (interview time and cognitive demands), more intensive interviewer training and the necessity of a much larger sample in order to achieve the same level of statistical precision warrant against the quick use of these techniques (in a telephone survey).

A closer look at the data of the ICT study also reveals so-called 'ceiling-effects' (Glynn, 2013): Whenever all items, the sensitive and the nonsensitive item, apply to a respondent, no anonymity is granted anymore. Thus, careful designing of the survey instrument (i.e., the innocuous questions) and even more extensive pretesting seem advisable. Recent studies argue that the ICT outperforms the RRT for obvious reasons, despite the ICT being even somewhat less efficient statistically (Corstange, 2009; Holbrook and Krosnick, 2010a, b; Coutts and Jann,

2011). While respondents in the RRT sometimes believe in some kind of trick, do not understand the rules or consciously 'cheat,' the ICT seems easier to administer, more convincing and more applicable in large scale surveys. Further, since the ICT does not require a lengthy introduction or a randomization device, and thus imposes less cognitive burden on the respondent, researchers see a solution for the problem of social desirability in this technique (Coutts and Jann, 2011; Glynn, 2013). However, looking at our results, we cannot support these hypotheses.

The second central result is that despite these methodological challenges and a low statistical power of the RRT estimator, we are able to investigate individual factors fostering undeclared work using logistic regression analyses for our RRT data. The perceived share of undeclared work in one's own network proved to be one robust explanatory factor in our substantive models explaining undeclared work (norm hypothesis). Other factors contributing to the understanding of engagement in undeclared work that were found to be statistically significant include: receipt of benefit (as a proxy for monetary gains from undeclared work), the number of people potentially helping with the job search, active organizational membership, as well as acceptance and approval of undeclared work. Those are all arguments within the utility and opportunity hypotheses framework. Furthermore, younger respondents (aged 34 and younger) are more likely to engage in undeclared work compared to those aged 35 and older. We cannot replicate differences between East and West Germany (Mummert and Schneider, 2001)[16] or between genders (Boockmann et al., 2010; Enste, 2012).

The main methodological conclusion of this chapter is, thus, that neither the RRT nor the ICT consistently outperform direct questioning. The main substantive conclusion is that we find particular evidence for the norm hypothesis in our models explaining undeclared work. These results are particularly relevant due to the cost implications for future studies: The increased costs in the RRT study and the ICT study—all other things being equal—are due to a larger sample size in the experimental conditions, longer interview times, statistically more complex analyses, and more intensive interviewer training. Given our empirical evidence, additional costs of an RRT or ICT data collection for undeclared work are not justified.

16 Which seems plausible given that the study was conducted in 2010.

3 Item Sum: A Novel Technique for Asking Continuous Sensitive Questions

While the previous chapter focused exclusively on the collection of information relating to binary sensitive characteristics, the following chapter introduces a novel method to collect data on continuous sensitive information based on the item count technique: We call it the 'item sum technique' (IST).[1] Recall from Chapter 1.3.4 that the main idea of ICT is to provide a subsample of respondents with a 'short list' of innocuous items, and another subsample of respondents with a 'long list' of items, containing a number of innocuous questions plus the sensitive question of interest. Respondents are then asked to indicate the number of items that apply to them (i.e., the total number of 'Yes' answers), without answering each question individually.

Compared to direct questioning, the ICT provides a higher degree of privacy protection and is thus assumed to yield more reliable self-reports. This expectation has been confirmed in several experimental studies comparing the ICT to direct self-reports. In the majority of these studies on topics such as employee theft (Wimbush and Dalton, 1997), risky sexual behavior (LaBrie and Earleywine, 2000), hate crime victimization (Rayburn et al., 2003), or shoplifting (Tsuchiya et al., 2007), the ICT yielded higher prevalence estimates for the sensitive behavior than direct questioning did (for an overview see Holbrook and Krosnick, 2010b).

Compared to other dejeopardizing techniques, such as the randomized response technique (RRT Warner, 1965), the ICT has the advantage that it does not require a randomizing device and that the procedure is much easier to administer. Thus, only a moderate cognitive burden is imposed on the respondent, likely increasing the respondent's ability to comply with the interview protocol and to provide more honest self-reports. Recent empirical studies indicate that the ICT outperforms the RRT in reducing social desirability bias in survey measures of sensitive attributes (see Holbrook and Krosnick, 2010a, b). To our knowledge, the ICT has only been applied to dichotomous items so far. We therefore present a generalization of the ICT that can be used to measure continuous sensitive characteristics—the IST—and report the results of an empirical application of the new method.

The remainder of the chapter is organized as follows: In Section 3.1, we describe our new technique. Section 3.2 briefly describes our empirical study in which we applied the new technique. Section 3.3 presents the results of the study and in Section 3.4, we draw conclusions and discuss limitations.

1 This chapter is based on a paper by Trappmann et al. (2014) and reprinted with modifications by permission of American Association of Public Opinion Research (AAPOR), the American Statistical Association (ASA) and Oxford University Press.

3.1 The Item Sum Technique

The item sum technique (IST) works as follows: Analogously to the ICT, two random subsamples are generated, whose respondents either receive a long list of questions (LL) or a short list of questions (SL). The long list contains the sensitive question plus at least one innocuous question, the short list only contains the innocuous question(s). The respondents are then asked to report the *sum* of the answers to the questions in their list. While, in theory, there is no restriction on the number of innocuous questions, it is desirable to keep the lists as short as possible. The variance of the sum of the answers usually increases with the number of individual questions, which reduces the statistical efficiency of the procedure. Furthermore, the cognitive demand of adding the answers together is increased with each additional item. We therefore suggest using just one innocuous question. Other than in the ICT where the non-sensitive items are binary, a single innocuous question with many possible values should be enough to make privacy protection credible in the IST.

Both the sensitive and the innocuous questions should be *continuous*, and preferably (but not necessarily) measured on the same scale (e.g., hours or monetary units). Respondents in the first subsample are asked to report the sum of the answers to both questions; respondents in the second subsample provide a direct answer to the innocuous question. For example, to estimate the extent of undeclared work, the following questions could be used:

Table 3.1: An example: the item sum technique

Item	Short List	Long List
How many hours did you spend watching TV last week?	X	X
How many hours do you usually spend in undeclared work per week?		X

Because respondents in the LL subgroup only report the sum of hours from both items, the extent of undeclared work remains unknown at the individual level. Assuming that respondents appreciate this privacy protection, the procedure can therefore be expected to elicit more honest answers to the sensitive question than direct questioning.

To estimate the amount of undeclared work from the IST data, we can simply compute the mean difference of answers between the two subsamples. Let Y be the observed answer, S be the sensitive variable (e.g., hours of undeclared work), and C be the non-sensitive variable (e.g., hours of watching TV). In the long-list

sample, we observe $Y = S + C$, while in the short-list sample we observe $Y = C$. Hence, as long as the two samples are unbiased, we can estimate the expected value of S, μ, as the mean difference of Y between the long-list sample and the short-list sample, that is

$$\hat{\mu} = \overline{Y}_{LL} - \overline{Y}_{SL},$$

where $\overline{Y}_{LL}$ is the mean for the long-list sample and likewise, $\overline{Y}_{SL}$ the mean for the short-list sample. Furthermore, as long as the samples are independent, the variance of $\hat{\mu}$ can be estimated as the sum of the sampling variances of the two group means, that is

$$\hat{V}(\hat{\mu}) = \hat{V}(\overline{Y}_{LL}) + \hat{V}(\overline{Y}_{SL}),$$

where standard formulas are used for the variances on the right hand side. Methods for estimating regression models for IST data are outlined in Appendix B.1.

3.2 Experimental Design

We implemented the new technique in the nation-wide RRT study on undeclared work in Germany that was briefly outlined in Chapter 2.1. In both samples, the 'employee' and the 'benefit recipient' sample, respondents who were originally assigned to the randomized response condition were automatically assigned to the IST condition, while respondents assigned to direct questioning (DQ) continued in the DQ condition. Within the IST condition, half of the respondents were assigned to the short-list IST group, and the other half to the long-list IST group (i.e., overall all respondents who were in the RRT condition). Some of the respondents originally assigned to the IST groups, however, opted out of being questioned using a special technique and were then given the survey with direct questioning. To be precise, respondents were given the option to switch to direct questioning after refusing to answer an RRT question measuring the prevalence of undeclared work (as a binary variable) earlier in the survey. They were given this option in order to prevent item nonresponse or even interview break-offs. In the subsequent IST experiment measuring the extent of undeclared work (as a metric variable), these 'non-compliers' or 'defiers' were kept in the direct questioning condition so as not to confront them with a second privacy preserving technique.

Table 3.2 gives an overview of the number of respondents in each experimental group and the number of respondents from the IST groups who opted for direct questioning. Chi-squared tests and two-sided t-tests (assuming unequal variances)

indicate that the 'defiers' (those who opted for direct questioning) do not differ significantly from the 'compliers' (those who stayed with the IST) with respect to: gender ($\chi^2 = 0.60$, $p = 0.44$), their attitudes towards undeclared work ($t = -1.32$, $p = 0.19$), the presumed prevalence of undeclared work among friends ($t = 0.44$, $p = 0.65$), and the perceived risk of being caught and sanctioned conducting undeclared work ($t = 1.00$, $p = 0.32$).[2] Defiers, however, are more likely to receive benefits than compliers ($\chi^2 = 33.20$, $p < 0.01$). Because the defiers might also differ from compliers with respect to other, possibly unobserved, characteristics we have to be careful about how to treat them in the data analysis that follows.

Table 3.2: Number of respondents per experimental condition

	Employees	Benefit recipients	Total
Direct questioning (DQ)	565	580	1,145
Short-list IST group			
Remained with IST	496	460	956
Opted for DQ	38	90	128
Long-list IST group			
Remained with IST	459	377	836
Opted for DQ	55	91	146
Total	1,613	1,598	3,211

In the DQ condition, respondents received a filter question asking whether they engaged in undeclared work in the past year (preceded by a confirmation that responses will be handled confidentially). Depending on the answer to the filter question, respondents were then led to questions about the weekly hours of undeclared work and the monthly income from undeclared work. For respondents who answered 'No' to the filter question in DQ mode, the two variables were set to zero.

Within the IST condition, the first group received two long lists (LL). Each of the long lists contained a sensitive question and an innocuous question. In the first list, the question about weekly hours of undeclared work was paired with a question about the number of hours the respondent watched TV last week; in the second list, the question about monthly earnings from undeclared work was paired with a question about the monthly costs for housing (both in Euro). Respondents were then asked to report the sum of the two answers for each list. Prior to the experimental section, respondents in the long list condition received a minimum of one 'training list' example—using information that we were able to validate

2 We focus on those covariates assumed to affect both noncompliance and the amount of undeclared work.

from prior survey information—to get acquainted to the novel method. The second group received two short lists (SL), each containing just the innocuous question about hours of TV or housing costs, respectively. See Table 3.3 for the wording of all questions (translated from German; for the full instructions, see Appendix A.2).

Apart from the experimental manipulation of the sensitive questions, all respondents received the same questionnaire covering items on demographics, employment, social networks, opportunity structures, attitudes and norms. Before asking the questions on undeclared work, a definition of undeclared work based on the German legal context was provided to the respondents. Undeclared work (including moonlighting) is defined as paid labor that is hidden from (tax) authorities for various reasons. Illegal (e.g., drug trafficking) or unpaid activities (e.g., neighborly help) are usually not included in the definition of 'undeclared work' or 'informal employment' (cf. Pedersen, 2003; Schneider and Enste, 2007; Williams, 2009).

Table 3.3: Item sum technique: wording of the items measuring the amount of undeclared work (translated from German)

Direct Questioning (DQ)	
S1:	How many hours do you usually spend in undeclared work per week?
S2:	On average, how much do you earn per month from undeclared work?
Long Lists (LL)	
C1:	How many hours did you spend watching TV last week?
S1:	How many hours do you usually spend in undeclared work per week?
C2:	How high are your monthly costs for your apartment or your house? Monthly costs can include rent, utilities, administrative fees, and mortgage.
S2:	On average, how much do you earn per month from undeclared work?
Short Lists (SL)	
C1:	How many hours did you spend watching TV last week?
C2:	How high are your monthly costs for your apartment or your house? Monthly costs can include rent, utilities, administrative fees, and mortgage.

Note: Monetary units are in Euro.

3.3 Empirical Results

3.3.1 Comparing Item Sum and Direct Questioning

This section provides an analysis of the item sum data from the CATI survey. The goal is to model how the reported amount of undeclared work depends on the data collection method (IST compared to direct questioning) and the labor market status (benefit recipient vs. employee sample).

Table 3.4 reports the estimated mean hours of and earnings from undeclared work for both samples across experimental conditions. The results are based on the realized assignment to either direct questioning (DQ) or the IST and, hence, ignore the fact that some respondents were initially assigned to a different mode.

Table 3.4: Mean estimates of hours of undeclared work per week and monthly earnings from undeclared work depending on questioning mode and sample (standard errors in parentheses)

	Hours of Undeclared Work		Earnings from Undeclared Work	
	Employees	Benefit recipients	Employees	Benefit recipients
Direct Questioning (DQ)	0.07	0.14	1.8	3.4
	(0.03)	(0.06)	(0.7)	(1.2)
Item Sum technique (IST)	0.85	-0.17	113.8	83.4
	(0.70)	(1.06)	(40.1)	(27.4)

Direct questioning leads to an estimate of 0.07 hours of undeclared work per week for employees and 0.14 hours for benefit recipients. Using IST, the estimate for employees rises to 0.85 hours, while a negative estimate of -0.17 hours results for benefit recipients.[3] For mean earnings from undeclared work, we get a DQ estimate of €1.8 per month for employees and €3.4 per month for benefit recipients. If using the IST, the estimates rise substantially to €113.8 and €83.4 per month respectively.

In Table 3.5, the differences between the estimates from direct questioning and the IST are shown (for a regression presentation see Appendix B.2). The first row ('naive estimate') contains the differences between the raw estimates as reported in Table 3.4. For hours of undeclared work, the effect of the questioning method (IST) does not appear to be significant (with a p-value of 0.26 in the employee

3 A negative value for the number of hours of undeclared work does not make sense, of course. However, note that the estimate is not significantly different from zero.

sample and 0.77 in the benefits recipient sample respectively). For earnings from undeclared work, however, the IST yielded significantly higher estimates than direct questioning in both samples ($p < 0.01$).

Table 3.5: Differences between direct questioning and the IST (standard errors in parentheses)

	Hours of Undeclared Work		Earnings from Undeclared Work	
	Employees	Benefit recipients	Employees	Benefit recipients
Naive Estimate	0.78	-0.31	**112.0***	**80.0***
	(0.70)	(1.06)	(40.1)	(27.5)
ITT Estimate	0.70	-0.32	**99.7***	**62.5***
	(0.63)	(0.89)	(35.4)	(22.0)
IV Estimate	0.78	-0.40	**111.9***	**77.9***
	(0.70)	(1.04)	(41.3)	(27.8)

Standard errors for the ITT and IV estimates were obtained by the bootstrap method (1000 replications, stratified by assigned experimental condition). For estimation methods, see Appendix B.1
* $p < 0.01$ (two-sided t-tests)

A comparison of the naive estimates might provide biased estimates of the effects of the questioning method because the group of respondents who opted out of the IST may be selective. One approach to deal with such treatment assignment noncompliance in randomized experiments is to compute the so-called intention-to-treat effect (ITT; see e.g. Hollis and Campbell, 1999; Newell, 1992): Instead of measuring the effect of actually receiving the treatment, the effect of being assigned to the treatment is estimated. The ITT is a conservative estimate for the causal treatment effect. That is, because only a fraction of the assigned treatment group is actually treated, the ITT is a weighted average of the true treatment effect and a zero effect. Hence, other than the naive approach, the ITT protects from possible overestimation of the causal effect of the treatment.

ITT estimates of the effect of the questioning method can be found in the second row of Table 3.5 (although the ITT principle is conceptually simple, its application is somewhat involved in our situation; see Appendix B.1 for details). For hours of undeclared work, the ITT estimates of the effect of the IST are almost identical to the naive estimates, supporting our earlier finding that in both samples the IST did not yield significantly higher estimates than direct questioning. For the reported earnings from undeclared work, however, the ITT estimates are smaller than the naive estimates. Yet, the effects are still significant at the 1% level. Hence, if we also employ the conservative ITT approach, we find that the IST had a substantial effect on the reported earnings in both samples.

As indicated above, treatment effects may be underestimated by the ITT procedure. We can improve on the ITT by using an instrumental variables (IV) approach, in which the realized treatment is instrumented by the assigned treatment. If a treatment effect is homogeneous (i.e., the same for everyone), then such an approach yields a consistent estimate of the causal effect. Alternatively, in the case of a heterogeneous treatment effect, the so-called local average treatment effect (LATE) is estimated (Angrist et al., 1996). This is the average treatment effect for the subpopulation of those who actually received treatment.

The last row of Table 3.5 displays the IV estimates of the effect of the IST on response behavior for our data (for methods, again see Appendix B.1). As can be seen, the results from the IV procedure are almost identical to those from the naive comparison of the direct questioning estimates and the IST estimates. Hence, we conclude that noncompliance with the treatment assignment did not substantially bias the data, so that the findings based on the naive estimates appear valid.

To summarize, irrespective of the employed estimation method, we find that in three out of four independent comparisons, the IST yielded larger estimates than direct questioning. In two out of these three cases, the difference is statistically significant (see Appendix B.2 for an equivalent representation in regression form).

3.3.2 Does it Work for Everybody? Differential Item Sum Effects

The differences reported in Table 3.5 are average effects over the respondents in our experimental groups and assume homogeneous treatment effects. The performance of the IST, however, might depend on cognitive skills of the respondents. Therefore, using the regression technique outlined in Appendix B.1, we evaluated whether the treatment effects vary not only by sample status, but also by age and education. Due to a lack of better data, we use age and education as proxies for cognitive skills: The models include binary indicator variables for respondents aged '58 or older' (vs. 57 and younger) and for respondents with 'no or the lowest secondary school leaving certificate' (vs. intermediate or upper secondary school leaving certificate). Including an interaction of each indicator variable with the IST treatment allows us to estimate differential IST effects for these groups. The following section will focus exclusively on these differential effects and only to a lesser extent on the main effects.

In Table 3.6, we estimate two regression models predicting weekly hours of undeclared work and monthly earnings from undeclared work as a function of these covariates. Due to the small differences between estimation techniques, we only present 'Naive' and 'IV' estimates for these regressions.

Table 3.6: Regression estimates for hours of undeclared work per week and monthly earnings from undeclared work

	Hours of Undeclared Work		Earnings from Undeclared Work	
	Naive Estimation	IV Estimation	Naive Estimation	IV Estimation
S				
IST	0.61	0.61	**102.28***	**102.11***
	(0.78)	(0.79)	(42.40)	(40.02)
Benefit Recipient	0.07	0.11	1.57	2.52
	(0.07)	(0.08)	(1.30)	(1.70)
IST x Benefit Recipient	-1.13	-1.22	-42.96	-45.03
	(1.25)	(1.33)	(49.18)	(46.47)
Age (≥ 58)	**-0.09***	**-0.09**$^{+}$	-1.66	-1.73
	(0.04)	(0.05)	(1.04)	(1.32)
IST x Age (≥ 58)	1.28	1.29	-34.67	-34.73
	(2.00)	(2.05)	(68.76)	(69.43)
Lower Formal Education	0.01	0.01	-0.55	-0.69
	(0.08)	(0.10)	(1.59)	(2.04)
IST x Lower Formal Education	-0.14	-0.14	67.01	67.48
	(1.61)	(1.60)	(59.68)	(62.20)
Constant	**0.08**$^{+}$	**0.08**$^{+}$	**2.16***	**2.30***
	(0.04)	(0.05)	(0.88)	(1.03)
C *(estimates are not displayed in the output)*				
N	3,199	3,211	3,130	3,211
N_SL		954		912
N_Total		3,072		3,003

Robust standard errors in parentheses (IV: bootstrap standard errors); IV: Instrumental variables estimates
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

All models contain an indicator for the questioning mode ('IST'), the sample ('Benefit recipients'), respondent age (≥ 58), level of education (no or lower secondary school leaving certificate) and their interactions with the treatment indicator.

The main effect of 'Benefit recipients' estimates the difference in mean reported hours (or earnings) between benefit recipients and employees in the DQ mode, while the main effect of 'Age' estimates the same difference for those respondents aged 58 and older compared to those respondents aged 57 or younger. The main effect of 'Education' assesses the difference between those respondents with a lower educational background and those with a higher school leaving certificate.

Turning to the central indicators of interest, the main effect of 'IST' is an estimate of the difference in mean reported hours (or earnings, respectively) between the IST mode and the DQ mode in the employee sample. The 'constant' term provides an estimate of those respondents in the direct questioning condition, who are in the employee sample, below the age of 58 and with a high educational degree. The interaction terms indicate whether the IST effect in the benefit recipient sample (as well as age ≥ 58 or lower educational background, respectively) is different from the IST effect in the employee sample (as well as ≤ 57 or higher educational background, respectively).

We are only interested in the functioning of the IST. Thus, we do not discuss any substantive findings, such as the main effects or the overall levels of 'hours of undeclared work' or 'earnings from undeclared work' in each subgroup. Our empirical results with respect to the differential IST effects in the models displayed in Table 3.6, show that:

- The main IST effect ('IST')—i.e., in the employee sample, below the age of 58, and with a high level of formal education—is positive for both the reported number of weekly hours of undeclared work and the monthly earnings from undeclared work. That is, employees below the age of 58, and with a higher educational background, reported more undeclared working hours (an additional 0.61 hours, SE = 0.78, p = 0.43) and more undeclared income (an additional €102.28), if interviewed by the IST compared to direct questioning. However, the effect is significant only for earnings (SE = 42.40, $p \leq 0.02$).
  The results obtained using IV estimation yield similar results (hours: 0.61, SE = 0.79, p = 0.44; income: 102.11, SE = 40.02, p = 0.01).
- The IST effects tend to be smaller in the benefit recipients sample compared to the employee sample (interaction effect). These differences in the IST effects between employees and benefit recipients are not significant, however. Turning to the overall effect, for hours of undeclared work, the point estimate for the overall IST effect is even negative among benefit recipients (0.61 – 1.13 = -0.53, SE = 1.04, p = 0.62). In the benefit recipient sample, the IST effect for monthly earnings is positive and significant (102.28 – 42.96 = 59.33, SE = 31.86, p = 0.06).
  The results obtained using IV estimation yield similar results (hours: 0.61–1.22 = -0.62, SE = 1.08, p = 0.57; income: 102.11 – 45.03 = 57.09, SE = 31.67, p = 0.07).
- The IST effect between age groups is not significantly different (interaction effect). However, the overall IST effects tend to be larger for respondents aged 58 and older compared to those aged 57 and younger for hours of undeclared work (0.61 + 1.28 = 1.89, SE = 1.90, p = 0.32), while they tend to be smaller for

income from undeclared work (102.28 – 34.67 = 67.61, SE = 68.14, p = 0.32). However, while the overall point estimate of the IST effect is positive for both items for respondents aged 58 and older, it is not statistically significant from zero, due to the comparatively low number of respondents in the upper age group (and thus lower statistical power to detect group differences).

The results obtained using IV estimation yield similar results (hours: 0.61 + 1.29 = 1.89, SE = 1.95, p = 0.33; income: 102.11 – 34.73 = 67.38, SE = 68.33, p = 0.32).

- Regarding education, we observe a similar pattern to that of benefit receipt: The interaction effect itself is nonsignificant, indicating that there is no differential IST effect. The overall point estimate of the IST effect tends to be smaller for respondents with a lower educational background for hours of undeclared work (0.61 – 0.14 = 0.47, SE = 1.55, p = 0.76). Among respondents with a lower educational background, the IST effect for monthly earnings is positive and significant (102.28 + 67.01 = 169.29, SE = 66.41, p = 0.01).

  The results obtained using IV estimation yield similar results (hours: 0.61 – 0.14 = 0.47, SE = 1.56, p = 0.76; income: 102.11 + 67.48 = 169.59, SE = 69.07, p = 0.01).

To sum up, when looking at the results for the regression estimates, we do not find systematic evidence for a differential effect of IST between groups varying in cognitive skills.

3.4 Discussion and Conclusion

In this chapter, we presented a new method, the item sum technique (IST), for the measurement of continuous sensitive characteristics. Compared to alternative methods, such as continuous RRT schemes (cf. Himmelfarb and Edgell, 1980; Eichhorn and Hayre, 1983; Gjestvang and Singh, 2007), the IST has several advantages: (1) a randomizing device is not required; (2) the cognitive effort demanded from respondents is relatively low; (3) implementation is easily possible in both interviewer- and self-administered interviews. The experimental evidence of our empirical study suggests that the IST is a promising data collection technique. It yielded significantly higher estimates of earnings from undeclared work than direct questioning in both the employee sample and the benefit recipient sample. For hours of undeclared work, estimates from the IST were also higher than from direct questioning in one of the two samples, although not significantly so. Furthermore, there was no evidence of significant interaction effects, i.e., differential item sum effects. Survey researchers aiming at measuring

sensitive behaviors on an ordered or continuous scale could therefore benefit from using the IST. Nonetheless, our study can only be regarded as a first step in the development and evaluation of the new technique.

One issue of our study is that a considerable share of respondents did not remain in the treatment condition they were initially assigned to, thus compromising the randomization of experimental groups. Using intention-to-treat and instrumental variables strategies, however, we believe that we convincingly demonstrated that our findings are robust regardless of this problem. A more serious concern is that, in the direct questioning condition, the continuous sensitive questions were preceded by a filter question on whether any undeclared work had been carried out at all, while no such filter question was present in the item sum condition.[4] Evidence suggests that filtering a continuous question may lead to underestimation of the quantity of interest. For example, in a study on crime victimization, Knäuper (1998) found that direct questioning estimates were twice as high compared to estimates obtained from a filtered question. Such biases are most likely due to differential interpretations of a construct depending on question format in cases where the construct is not clearly defined. In our IST study, however, an explicit definition of undeclared work was given directly before the relevant questions were asked (see Chapter 2.1.4). Furthermore, for earnings from undeclared work, our IST estimates are much higher than the direct questioning estimates (factor 63 in the employee sample; factor 25 in the benefit recipients sample; see Table 3.4). Thus, we do not believe that the filtering could explain the differences observed in our study.

An interesting finding is that the IST was only successful for one of the two questions. For earnings from undeclared work, the IST impressively outperformed direct questioning, but no significant differences were found for the question on the number of hours of undeclared work. We do not believe that this null-result is due to a lack of statistical power. First, relative differences as observed for earnings (which one would expect if working hours and earnings are proportional) would have been easily detected given the power of the study.[5] Second, the effect is absent in both samples, the employee sample and the benefit recipient sample, which makes us believe that the pattern is systematic. A candidate explanation may lie in the choice of innocuous items. While the item on housing costs appears unproblematic, the question on the number of hours a respondent watched TV, which was paired with

4 The IST experiment described in our article was preceded by an experiment using direct questioning versus RRT to estimate the prevalence of undeclared work. Respondents from the direct questioning group skipped the subsequent continuous questions if they answered that they did not carry out any undeclared work. In contrast, all respondents from the RRT group were directed to the continuous questions (and randomized into either the short list or long list of the IST), since filtering based on RRT answers is not possible.

5 Results from a simulation based on the characteristics of our data indicate that the power to detect such a difference at the 5% level would have been about 97% (not shown).

the question on hours of undeclared work, might not have been an optimal choice. On the one hand, there is evidence that answers to this question strongly depend on question format (Schwarz et al., 1985). On the other hand, the question on watching TV might be considered sensitive by some respondents. In this case, respondents would tend to underreport their TV consumption if asked directly, which would positively bias our IST estimate of hours of undeclared work. Yet, the opposite is what we observe empirically. Perhaps a better explanation might be that there were learning effects. The item sum technique may not have worked well for the first item (hours of undeclared work), because respondents were not used to the unusual questioning format. However, after getting acquainted with it, the new technique may have worked better for the second item (earnings from undeclared work). We can only speculate whether this explains the differing findings. In any case, however, we can suggest including suitable training questions when employing novel questioning techniques such as the IST in a survey.

As illustrated by the above qualifications, further experimental research is needed to fully understand the mechanisms at work when respondents are confronted with sensitive questions in the item sum format. Obtaining unbiased estimates for sensitive variables by the IST rests on a number of assumptions. In the following, we outline how an implementation of the IST that maximizes the credibility of these assumptions might look like and develop ideas about how the validity of the assumptions could be evaluated (see Blair and Imai, 2012 or Glynn, 2013 for similar discussions in the context of the item count technique).

First, it is necessary to assume that the respondents comply with the design. Therefore, careful cognitive pretests are necessary to make sure that respondents fully understand the procedure. In addition, measures could be implemented to minimize nonresponse or noncompliance with the IST (for an overview see De Leeuw et al., 2003, p. 154 f.). These could, for example, include interviewer probing after an initial 'do not know' response.

Second, it is crucial that the answers to the non-sensitive item are independent of the question format, that is, that the answers do not depend on whether they are given directly or serve as a summand in the item sum format ('no-design-effect,' see Blair and Imai (2012)). It is important, therefore, that the innocuous item is truly non-sensitive and not affected by social desirability bias itself. Furthermore, the summation task should be made as easy and convenient as possible. Although adding two numbers might appear simple, the ability to perform such a task error-free is likely to depend on cognitive capabilities.[6] Even if most respondents

6 We assume that the task is easier for respondents if the items are on the same scale. Also, the items should be disjoint to prevent confusion over whether the overlap must be subtracted from the sum or not.

exhibit sufficient cognitive skills to correctly add the items, they might engage in satisficing (approximating the sum), most likely leading to rounding errors (see Tourangeau et al., 2000 for an overview on rounding). Such 'heaping' might bias the results if the net effect of rounding is different between the short list and the long list, for example due to differing distributions of the true values around focal rounding points. In order to reduce summation as well as rounding errors in our telephone survey, we asked respondents to write down the answers to the individual items before adding them up. This is obviously not to be recommended in face-to-face settings.

The 'no-design-effect' assumption can be evaluated, for instance, by administering a survey in which one experimental group is asked to answer two separate non-sensitive questions and the other group is asked to report the sum of the two questions. If the 'no-design-effect' assumption holds, the results in the two groups should be the same (for the ICT see Tsuchiya et al., 2007; Tsuchiya and Hirai, 2010).

Third, careful power analyses are necessary to determine sufficient sample sizes. For this purpose, it is helpful if the distribution of the innocuous item is known in the survey population and at least crude ideas about the distribution of the sensitive item and its covariance with the innocuous item exist. The variance of the innocuous item plays a crucial role in the trade-off between privacy protection and statistical efficiency.[7] If the variance is too small, then it does not sufficiently protect the privacy of respondents; if it is too large, then the estimation becomes inefficient.[8] In addition, the covariance between the items also matters: A negative covariance between the non-sensitive item and the sensitive item reduces the total variance and therefore increases efficiency (cf. advice by Glynn, 2013 for ICT designs).[9] More research that experiments with

7 Furthermore, for ease of statistical modeling, it is convenient if the innocuous item has a distribution that is approximately normal.

8 Another prerequisite for sufficient privacy protection is that the sum of the two items does not exceed the possible maximum value of the innocuous item. The innocuous items used in our study seem unproblematic in this regard. Monthly housing costs have no upper limit and follow a skewed distribution so that extreme outliers are rare but credible. Watching TV has a theoretical upper limit of 168 hours per week, but the sum of undeclared work and watching TV will not reach this limit (at least if they are disjoint) because both have to fit into the overall time budget of a person.

9 We conducted some preliminary simulations to evaluate the bias-variance trade-off for the IST compared to direct questioning (not shown). Assuming that the standard deviation of the innocuous variable is about five times the standard deviation of the sensitive variable (which roughly corresponds to our data for the second item) and given a fixed sample size of 500 observations for both questioning techniques, the IST has a lower mean squared error (MSE) than direct questioning, if the direct questioning results are biased by half a standard deviation of the sensitive variable or more. If the sample size for the IST is doubled (as in our case), then the MSE of the IST is smaller than the MSE of direct questioning, if the bias is about 30 percent of a standard deviation or more. Of course, the variance of the innocuous variable has a strong effect on these results. For example, if both variables have the same variance, then the bias can be as low as 10 percent of a standard deviation before the MSE of the IST exceeds the MSE of direct questioning (given a fixed sample size of 500 for both questioning techniques). The correlation between the sensitive variable and the innocuous variable has substantial effects only if the variances of the two variables are similar. We conclude that, as long as the relative variance of the innocuous variable is not too large, the IST provides estimates that are superior to direct questioning even if social desirability bias is only moderate.

these issues to find an optimal trade-off between perceived privacy protection and statistical efficiency is needed. For example, a promising approach might be to use non-sensitive items whose variance (or covariance with the sensitive item) is subjectively overestimated by survey respondents (cf. Diekmann, 2012, for RRT). A different strategy for a more efficient estimation might be to select non-sensitive items that can be predicted with high accuracy from other variables collected in the same survey. In order to increase statistical power, future studies could also consider the use of a double-list design (similar to the double-list variant of the ICT where both groups receive a long list and a short list with varying non-sensitive items; see Droitcour et al. (1991); Biemer et al. (2005) requiring a smaller sample size to achieve a given level of statistical power.

Finally, an ideal study to evaluate the IST would not rely on the 'more-is-better' assumption but instead would use validation data with known true scores for the sensitive variable. Opportunities for validation studies are notoriously difficult to find, and data protection issues arise if individual-level data from different sources are linked without the informed consent of the respondents. In the past, validation studies for dejeopardizing techniques have successfully made use of register samples that were constant with respect to the dependent variable, avoiding individual linkage of survey and register data (compare the validation studies for RRT reviewed in Lensvelt-Mulders et al. (2005a)). This strategy might be permissible from a data protection perspective in the case of continuous sensitive items as well. However, samples created in this way would be very specific as everyone in the study would have the exact same true value for the continuous item of interest. This might raise questions about the generalizability of such results. Another approach might be to conduct validation studies in which the distribution of the sensitive item among respondents is known, but the data are not linked at the individual level (for similar approaches in a different context see Kreuter et al. (2010); Sakshaug and Kreuter (2012)). In such studies, at least an overall evaluation of the validity of the IST would be possible.

4 Validating Sensitive Questions: A Comparison of Survey and Register Data

Usually, the RRT and other dejeopardizing techniques are considered successful if they produce higher prevalence estimates than traditional direct questioning in surveys for items that are expected to be subject to underreporting ('more-is-better' assumption e.g., Weissman et al., 1986; Lara et al., 2004, 2006 for an overview of studies relying on this assumption, see Umesh and Peterson, 1991; Lensvelt-Mulders et al., 2005a).[1] The previous chapters evaluated different dejeopardizing techniques relying on this 'more-is-better' assumption comparing aggregate estimates. Whether the 'more-is-better' assumption is in fact correct can only be evaluated against additional information regarding the actual status of the respondent. Validation studies serve this purpose comparing (aggregated) survey reports with (aggregate) auxiliary data. Though the most powerful validation can be achieved if the 'true' value of a respondent is known at the individual level and can be compared to the survey report. Using individual-level validation data provides another possibility, namely to analyze motivations that contribute to misreporting. This would not be possible using aggregate data and relying on the 'more-is-better' assumption.

Due to the particular design of the RRT study, we can empirically evaluate the performance of the RRT using validation data, however, due to the challenges outlined previously, not for our main item of interest (undeclared work). Another item in the context of labor market surveys that qualifies for this purpose is the receipt of welfare benefits: We can plausibly assume this to be sensitive information that is underreported in surveys, while at the same time, validation data is available at the individual level. In general, surveys which collect data on welfare and unemployment receipt often find that these variables regarding sensitive labor market information are underreported. The known extent of underreporting of receipt of basic income support, a form of social security payment, in German surveys ranges between 9 and 17 percentage points depending on the exact population under study (Kreuter et al., 2010, 2013).

Accurate information regarding the types and extent of receipt of these payments is essential for policy decisions. If the failure to report welfare receipt is systematically different for certain social groups of respondents, resulting statistics such as regression coefficients likely suffer from considerable bias (Hausman, 2001). Hence, any substantive findings with respect to the dynamics of receipt, such as individual characteristics leading into benefit receipt, are likely to

1 This chapter is based on a manuscript by Kirchner (2013).

be invalid. Insights into the mechanisms of misreporting and possible remedies are therefore of great importance for researchers using such data, since it will prevent them from drawing wrong inferences, and ultimately for policy makers, as they will be able to make better informed decisions.

While unintentional misreporting, e.g., due to recall error, is certainly an issue in the reporting of social security receipt (Manzoni et al., 2010; Kreuter et al., 2013), particular attention should be devoted to intentional misreporting. Respondents are likely to conceal sensitive information due to fear of legal and/or extralegal sanctioning, which in turn, has a negative effect on the validity of the data and prevalence estimates (Lee, 1993). Further, estimates of parameters such as proportions, averages as relationships between variables will be biased (Hausman, 2001).

In order to explore whether the RRT method is a successful means of improving the quality of data regarding the receipt of basic income support, we rely on data from the RRT study outlined in Chapter 2.1. Due to the specific study design (using administrative record data), we know the true percentage of respondents who have received transfer payments for basic income support and thus the percent who should have reported receipt. Thus, we can validate the reported percentage against the known true rate for the responding cases hence assessing the bias of the estimates. Such administrative record data is quite rare in the literature on sensitive questions (Lensvelt-Mulders et al., 2005a; Wolter, 2012), and provides a unique opportunity to evaluate the RRT compared to traditional direct questioning without having to rely on the 'more-is-better' assumption.

The study contributes in several ways to the existing research on the RRT and response bias: To the best of our knowledge, the performance of the RRT in a telephone survey has never been validated (especially not with respect to the receipt of basic income support). All existing RRT validation studies were implemented in a face-to-face, but never in a telephone setting (cf. also Lensvelt-Mulders et al., 2005a; Wolter, 2012, p. 108). The choice of a telephone mode, however, might be perceived as more private by respondents, thus leading to more honest answers. While collecting data by means of the RRT has many advantages, RRT procedures also suffer from considerable disadvantages compared to direct questioning: For one, a larger sample size is needed to achieve the same statistical power (Warner 1965); second, interview duration increases due to an explanation of the application of the procedure, while third, the cognitive burden placed on respondents is higher. Examining the functioning of a telephone implementation of the RRT might prove useful, given that it is more cost efficient compared to face-to-face surveys. We thus follow the recommendation by Lamb and Stem (1978, p. 617) that "each time

the [RRT] method is changed or used in a different setting, further evaluation is appropriate." Furthermore, we contribute by investigating which individual-level factors influence accurate reporting and whether these mechanisms differ across experimental conditions.

To summarize, this chapter addresses two research questions. First, whether item specific response bias in surveys can be reduced when the randomized response technique is applied with respect to a) the true value in the administrative data and b) direct questioning (DQ) in the survey data. And second, which subgroups are especially affected by response error.

The remainder of this chapter is organized as follows: In Section 4.1 we discuss the challenge of asking sensitive questions. Section 4.2 describes the empirical study and the available data. Section 4.3 lays out the method of analysis, while Section 4.4 presents the results of the experiment. In Section 4.5, we draw some final conclusions and discuss the limitations of the study.

4.1 Background

Remember that the level of 'threat' or 'sensitivity' of a question as perceived by the respondent can be established by three theoretical criteria (Tourangeau and Yan, 2007): intrusiveness, risk of disclosure as well as social desirability.

Several of these theoretical dimensions apply to the receipt of basic income support (UB II)[2]: People can apply for welfare benefits in Germany either if they have been unemployed long-term or if they cannot make a living from their current job (if the resulting income is below a certain threshold). Respondents receiving basic income support may not wish to report this information in a survey: Admitting to the interviewer that they either have not been able to find a job over a longer period, that they live in poverty or that they do not earn enough to support their families can be quite embarrassing. The concept of 'injunctive social norms' (Cialdini, 2007), i.e., one's perception or expectation of what most others approve or disapprove of, plays a vital role in this context. Negative beliefs and prejudice about welfare recipients in the United States and Great Britain comprise anything from not being motivated enough to find a job, uninterested in self-improvement, dishonesty, to laziness and dependence (Bullock, 2006, p. 2060). The receipt of basic income support in Germany is associated with similar prejudice and can thus be defined as socially undesirable,

2 Since the so called Hartz Reforms in the German social security system in 2005, people are entitled to welfare benefits called 'Unemployment Benefit II' if they are between 15 and 64 years of age, capable of working, and if the household they live in—or more precisely, benefit community—does not have sufficient income to secure a livelihood. Thus, the receipt of UB II does not depend on the current employment status, but depends on the criterion of sufficient income. More detailed information on the receipt of basic income support can be found in Appendix C.2.

in terms of the commonly perceived norm, and perceived as potentially negatively stigmatizing, causing embarrassment when admitting to such. Also, Rasinski and colleagues' (1999, p. 479) study shows that normative role expectations significantly contribute to underreporting (of abortions). More precisely, the greatest associated risk that respondents identified were interviewer disapproval and embarrassment when answering 'threatening' questions.

To avoid errors from (item) nonresponse and misreporting ('under-' as well as 'overreporting') due to the sensitive nature of a question, survey methodologists have suggested a range of guidelines with respect to the design of a questionnaire (for an extensive overview see Lee, 1993; Bradburn et al., 2004; Tourangeau and Yan, 2007). One of these strategies is the RRT (Warner, 1965), as described in Chapter 1.3.3 to elicit information on sensitive questions. Remember that the main idea of all RRT variants is to establish a probabilistic relationship between the survey response of an individual and the sensitive trait, e.g., by means of a randomization device with a known probability distribution (e.g., coins, cards, dice, spinner) (Fox and Tracy, 1986).

In the most recent meta-analysis (Lensvelt-Mulders et al., 2005a), a total of six individual-level RRT validation studies and 32 comparative RRT studies without validation data were investigated (for an overview of validation studies, see Appendix C.1). The RRT produced some response error, however, lower than a comparable standard face-to-face questioning: For the validation studies under investigation, in the RRT, the mean response was underreported by 38 percent, while in the face-to-face condition, mean underreporting was 42 percent. One of these validation studies, conducted by van der Heijden and colleagues (2000, see also Lensvelt-Mulders et al. (2006)), tested two different implementations of the RRT against standard face-to-face questioning. Results suggest that both RRT versions yield significantly lower response error with respect to social security fraud. Other experimental studies without validation data (based on the 'more-is-better' assumption) also showed that the RRT increased validity of the estimates by eliciting more truthful responses (to name a few: Weissman et al., 1986; Lara et al., 2004, 2006).

In general, the RRT seems to elicit more honest answers and reduce social desirability bias, especially when dealing with more sensitive questions (Fidler and Kleinknecht, 1977; Landsheer et al., 1999, p. 2; Lensvelt-Mulders et al., 2005a). The pioneering study by Locander and colleagues (1976) using individual-level validation data show for example, that the response error for RRT is (significantly) lower compared to that of direct questioning in three out of five instances. While the trend—i.e., the RRT eliciting higher prevalence estimates—is as expected in most validation studies, some studies also find no effects (Lamb and Stem, 1978;

Wolter, 2012). Furthermore, some RRT validation studies provide mixed or contrary evidence (Locander et al., 1976; Tracy and Fox, 1981) as do other experimental studies that do not use validation data (Umesh and Peterson, 1991; Holbrook and Krosnick, 2010a; Coutts and Jann, 2011; Coutts et al., 2011).

Given the largely positive empirical evidence, the following chapter addresses the question, of whether accurate data on basic income support can be collected using the RRT in the context of large-scale labor market surveys and how much bias, if any, still remains with this method of data collection.

4.2 Data and Methods

In order to assess the amount of misreporting of the receipt of transfer payments in survey reports, we draw on data from the RRT survey (outlined in Chapter 2.1). Due to the particular sampling design, we can—in addition to these survey data—draw upon supplementary information available on the sampling frame (administrative data). The combination of both data sources allows us to address the research questions stated above. The next section provides an overview of the survey data, the administrative data and the combined data.

4.2.1 The Survey Data

For the purpose of this chapter, we focus exclusively on the data collected on welfare benefit receipt, which will be used to form the dependent variable.

4.2.1.1 Sampling and Data Collection

The survey is a dual-frame survey that permits validation of the survey reports as well as a comparison of benefit recipients and employed persons. Remember that both samples were drawn from the registers of the Federal Employment Agency according to the following criteria: The first sample, the benefit recipients sample (UB II), consists of persons who were known to have received basic income support in June 2010—information that we will later on use to validate the survey reports. In order to be part of the population for the second sample, the employee sample, the requirement was employment in December 2009.

4.2.1.2 Measurement of the Dependent Variable

Several measures were taken in order to reduce potential error sources in the survey measure of welfare receipt, such as recall error, and to decrease sensitivity (Tourangeau et al., 2000; Groves et al., 2009; Manzoni et al., 2010). We decided

to implement two different operationalizations: For the UB II sample—known to have received benefits in June 2010—we asked participants to report 'benefit receipt ever'[3]. In the employed sample, we asked participants to report receipt in 'September 2010'[4].

While these different operationalizations ensure that we can validate (aggregate) responses, another criterion was to keep the questions as simple as possible in order to ensure understanding and correct recall. To ease recall in the employed sample (and allow validation), the question relates to a defined period of receipt just prior to data collection. Further, all question formats were kept as similar as possible to commonly used questions in labor market surveys (cf. the PASS study as described by Trappmann et al., 2010).

4.2.1.3 Independent Variables and Operationalizations

A range of indicators explaining underreporting of UB II will be analyzed in the scope of the second research question. Existing empirical evidence shows that underreporting of UB II is more frequent among males, among people aged 25 and younger as well as those between 40 and 57 (Kreuter et al., 2013). The authors also found a significant effect of household size and years of schooling. Those respondents with a higher education and those living in a larger household underreported more frequently. Household size in that particular instance is not to be taken literally: It is rather an indicator capturing a higher propensity to conduct the interview with someone less knowledgeable about the receipt of UB II and thus response error should be larger. Kreuter et al. (2010, 2013) also show that respondents who are more reluctant to participate in a survey are slightly more likely to underreport benefit receipt. The authors attribute this effect to a lower motivation of these respondents, while controlling for sample composition and recall error due to a longer recall period. Both studies mentioned above only applied direct questioning techniques.

Drawing on main insights from these studies, as well as on behavioral theories and the response process (Tourangeau, 1984; Tourangeau and Rasinski, 1988), we will consider variables that capture subjective costs, risks and utilities that are associated with accurate reporting of UB II. Further, those measures that capture understanding and application of the RRT as well as paradata will be assessed, containing relevant information regarding the survey process.[5] In theory, we assume that we will see the following significant (negative) effects

3 "Did you ever receive unemployment benefits II?" – Y/N. (note: translated from German).

4 "Did you receive unemployment benefits II in September 2010?" – Y/N. (note: translated from German).

5 Collected as a byproduct of the survey data collection (Couper, 1998; Kreuter, 2013).

in the model of the direct questioning split, namely, for characteristics that are associated with higher subjective reporting costs. These costs are subjectively higher, if receipt of UB II is perceived as particularly sensitive, e.g., when a respondent has a higher education. If the RRT reduces perceived reporting costs associated with item sensitivity in a particular subgroup, we would expect this effect to vanish in the RRT model. According to the work of Böckenholt and van der Heijden (2007) the RRT works especially well if the RRT instructions are clearly understood and the cognitive burden is kept as low as possible. The authors also stress the importance of personal benefits and social influences, e.g., a person's expected benefits of noncompliance, which would be captured by the first set of indicators. Therefore, additional factors regarding the RRT and the survey process will be analyzed.

Table 4.1 presents an overview of all independent variables. Factors contributing to perceived item sensitivity, and hence associated reporting costs, comprise: employment status, occupational status, and a respondent's willingness to provide socially undesirable answers. Further, the reluctance of the respondents to answer sensitive questions is operationalized with an indicator variable, measuring item nonresponse for the item household income. Equally important is a measure of how common the receipt of UB II is in a respondent's environment. Admitting to receiving UB II would be perceived as being less of a norm violation and reported more accurately. Ideally, this indicator would be measured at the neighborhood level, which is not possible in this particular case due to data privacy issues. Thus, an indicator recipient rate at the more aggregate municipal level is included in the models.

A second set of factors relates to the survey process and to the application of the RRT by the respondents. The first indicator captures whether a respondent refused to cooperate in the RRT condition (DQ_RRT) and was then surveyed in the direct questioning mode. In order to capture understanding of the RRT, two proxy indicators are used (Landsheer et al. 1999): First, interviewers were asked to rate the language skills (German) of a respondent immediately following the telephone interview. A second indicator pertaining to the understanding of the RRT instructions is educational attainment (formal training). Response latency, i.e., the speed at which a respondent answers, is used as a measure for response quality.

All models control for gender (0 male, 1 female), age (below 25, 25-40, 41-57, 58 and above), and if a respondent lives in East Germany (0 West Germany, 1 East Germany) or in a single person household (0 multi-member household, 1 single person household). Including these controls seems appropriate given noncompliance to the randomization as well as the assumed differential underlying mechanisms in both experimental groups.

Table 4.1: Description of variables used in the multivariate analyses

Indicator	Description
Factors Contributing to Perceived Reporting Costs and Item Sensitivity	
Employment Status	At the time of survey 0 Not employed (unemployed, parental leave, student etc.) 1 Marginally employed with income up to 400€ 2 Employed with labor income > 400€
Occupational Status	International socio-economic index of occupational status (ISEI) (Ganzeboom et al., 1992). Coded based on ISCO88 of present or last job[6] 0 No ISEI available, i.e., never held a job before (score = .) 1 Low or medium ISEI of present or last job (score 16–43) 2 High ISEI of present or last job (score > 43)
Previous Undesirable Response	Previous socially undesirable response regarding tax honesty. Tax honesty is 0 Worthwhile, absolutely worthwhile 1 Not worthwhile, absolutely not worthwhile
Reluctance	Item nonresponse for household income 0 Valid response 1 Missing data
Recipient Rate	Share of UB II in Municipality
Survey Process and Application of RRT	
RRT Refusal (DQ_RRT)	0 RRT condition 1 DQ_RRT condition
Language Skills	Scale from 1 very good to 6 non-existent (recoded 0,1) 0 Good (< 3) 1 Poor (≥ 3)
Formal Training	Tertiary degree
Response Latency	Standardized response time in experimental section (recoded according to quartiles) 0 Slow response ($< Q_{25}$) 1 Mean response (Q_{25}–Q_{75}) 2 Fast response ($> Q_{75}$)
Controls	
Gender	Female
Age	0 < 25 1 25 to 40 2 41 to 57 3 > 57
Region of Residence	East Germany
Single Person Household	0 Multi-member household 1 Single person household

6 To code the ISCO88 data the Stata ado iskoisei was used (Hendrickx, 2002).

4.2.2 Register Data

The analysis to address both research questions uses register data based on social security reports and reports from the FEA itself as gold standard. This supplementary administrative data from the sampling frame contains various receipt related information (e.g., type of benefit), and employment related information (e.g., type of employment and income for all sample units). Data on employment is reported by the employers on a yearly basis for all employees subject to social security contributions (compulsory notification scheme). Civil servants or self-employed persons are thus excluded, while marginally employed and part-time workers are included. Aside from information on the employment status, FEA obtains records of received unemployment benefits, job search and participation in active labor market programs. Information regarding basic income receipt is a by-product of the FEA activities, i.e., process data based on the information provided by the applicants themselves. Both the employee records, as well as the basic income receipt records, are collected on an individual-, spell-based level and are kept in the registers at the Institute for Employment Research.

For the analyses, only one indicator in the administrative data is of relevance: whether an individual received UB II. As a general rule, all data relevant to payments and claims (taxes, pensions, unemployment benefits etc.), i.e., the primary use of the social security system, are known to be of very good data quality (Jacobebbinghaus and Seth, 2007). The analyses thus rest on the crucial assumption that we can capture the true value of our respondents with these data. The UB II receipt indicator is known to be both accurate and complete and can serve as gold standard.

4.2.3 The Linked Data

There are some limitations to this validation study, since we did not ask respondents for consent to link their survey data to the administrative data and cannot link the two data sources.

However, due to the sampling plan, we do know that each individual in the UB II sample should by default respond with 'Yes' to the 'benefit receipt ever' question. Overreporting is not possible by definition. With the true aggregate prevalence being 100 percent, we can create an indicator variable on the individual person level that captures whether an individual reported accurately even without linkage of the two data sources. This measure of reporting accuracy is a simple dummy variable that takes on the value 1 if the survey report matches

the true value in the administrative records, and 0 if the survey report is 'No,' i.e., a mismatch between the survey data and the administrative records.[7]

For the employed sample, the missing linkage consent question slightly complicates the analyses. In theory, the individual-level data on UB II receipt in September 2010 is available for all respondents on the frame. However, again, we are not allowed to link the survey data to the respective administrative records for reasons of data privacy. Thus, nothing is known at the individual level for these respondents with respect to their 'true' UB II status. It is therefore not possible to construct a variable indicating reporting accuracy at an individual level for this sample. However, we are allowed to link paradata regarding survey participation to the administrative data on the sampling frame. Linking this survey response indicator (0 nonrespondent, 1 respondent) to the administrative data, it is possible to derive and compare aggregate measures for respondents. According to the administrative data, the true aggregate prevalence of 'benefit receipt in September 2010' for all respondents of the employed sample is 3.8%. In this subsample of the survey, overreporting could theoretically be an issue. However, it seems unreasonable to assume that respondents, aside from overreporting due to satisficing or acquiescence, would (consciously) overreport UB II receipt for the reasons mentioned above.

4.3 Statistical Analyses

To assess the impact of measurement error from the two alternative techniques of data collection, we calculate the response bias. The bias of a statistic is simply the difference between the statistic's expectation and the true population value. The estimator of the response bias (B_j) in the respective experimental condition j is thus (adapted from Biemer, 2010, p. 49):

$$B_j = \bar{y}_{j,svy} - \bar{y}_{j,adm}, \tag{4.1}$$

which is the difference of the means of accurate reporting in the sample survey measurements ($\bar{y}_{j,svy}$) and the gold standard measurements ($\bar{y}_{j,adm}$). This approach will then allow for a comparison of the overall response bias of the RRT and the DQ in both subsamples.

7 We observe item nonresponse for four respondents (out of 3,211) across all experimental conditions. Those cases with no survey information on receipt of benefits are excluded from the analyses.

The following assumptions are critical for the subsequent analyses of response error and interpretations. *First*, for the purpose of this chapter, we assume that there is no measurement error in the administrative data. *Second*, with respect to the idea of resampling and thus sampling variance of the administrative data: If we randomly sample from the distribution of UB II recipients, we would by definition always observe a true prevalence of 100 percent. While this assumption of zero sampling variance in the administrative data is reasonable for the UB II sample, it is not entirely true for the employed sample. Nonetheless, for reasons discussed below, we will make this assumption. Based on this assumption, an unpaired t-test (unequal variances) to test whether the mean survey reports differ from the mean administrative values will yield correct standard errors of estimates of differences, since the covariance term is by assumption zero. Also, since both experimental splits are randomly assigned and independent, the means are assumed to be uncorrelated, yielding correct t-statistics for the significance tests of estimates of differences between experimental groups in an unpaired t-test. *Third*, we assume that there is no overreporting of welfare benefit receipt. If, again, we were to resample from the respective population, the estimate of the mean prevalence in the survey data then actually captures the propensity to report UB II receipt accurately [0, 1] while the absolute response bias is a measure of the propensity to misreport taking on values between -1 and 0. The magnitude of the response bias is also dependent on the absolute prevalence of receipt of UB II. However, calculating the relative response bias i.e., depending on the occurrence of the prevalence in the administrative data, is not meaningful since 'absolute' occurrence of misreporting is the measure of interest. This issue is only relevant for the employed sample, since in the UB II sample, relative and absolute bias are identical. The mechanisms that cause underreporting also differ in the experimental groups: While in the DQ condition we have 'pure' underreporting, in RRT we also have 'noncompliance.' Thus, calculating a relative bias in the employed sample is not sensible. Second, if the questions were the same across subsamples, conditioning on the true value ($\bar{y}_{j,adm}$), should yield the same point estimate of the relative response bias in both samples. The estimated confidence intervals in the employed sample should then be larger in comparison to those of the UB II sample due to the lower precision. In analyses not presented here, we can show that the relative response bias is larger in the employed sample, as are the confidence intervals.

After analyzing the overall bias (research question 1), in a next step we estimate logistic regression models of accurate reporting as a function of individual characteristics (research question 2). In both experimental conditions, the dependent variable Y_{ji} represents an individual's response behavior (0 – underreporting; 1 –

accurate reporting). If the assumptions of privacy protection in the RRT condition hold, significant predictors of underreporting related to perceived item sensitivity in the direct model should then become nonsignificant in the RRT model or, put differently, should be (significantly) positively related to accurate reporting. While for the direct condition a logistic model of compliance is appropriate, for the RRT split we will again use the rrlogit routine in Stata (version 12.1) with an adapted likelihood function that accounts for the additional noise introduced by the RRT procedure (Jann, 2011, see also Chapter 1.3.3).

4.4 Empirical Results

Before we address the initial research questions, we need to establish whether the receipt of basic income support has been significantly underreported in our experiment, and if we can replicate the findings of other studies (Kreuter et al., 2010, 2013). Since the exact questions asked in the survey differ across the two subsamples, response bias estimates are not comparable across subsamples and should only be interpreted separately.

Table 4.2 shows the estimated prevalence in percent for direct questioning across both subsamples, as well as the resulting estimate of the absolute magnitude of the response bias (%pts). The estimated response bias pointing in the expected direction is boldfaced, indicating a statistically significant amount of underreporting.

Table 4.2: Estimated proportions in percent and absolute response bias in percentage points for DQ

Sample Type	j	$\bar{y}_{j,svy}$ [95% C.I.]	$\bar{y}_{j,adm}$	B_j [95% C.I.]	n
UB II	DQ	0.870 [0.843;0.898]	1.000	**-0.130** [-0.157;-0.102]	579
Employed	DQ	0.021 [0.009;0.033]	0.030	-0.009 [-0.021;0.003]	564

As expected, we find underreporting of receipt of benefit in the DQ condition: For the UB II sample, we do observe a significantly lower report of UB II receipt by 13.0 percentage points, indicating substantive misreporting. While receipt of benefits is also underreported by 0.9 percentage points in the employed sample, it is not statistically significant. In absolute terms, the bias is larger in the UB II sample, however, in relative terms, i.e., standardized on the value of true prevalence, it is much larger in the employed sample (29.3% compared to 13.0%). However, as discussed in the previous section, differences are expected due to the less difficult question of replying to receipt 'ever' in the UB II sample, and 'September' in the employed sample.

4.4.1 Reduction of Response Bias by Means of RRT?

We expect to see a statistically nonsignificant divergence of the survey data estimates from the gold standard if the RRT alleviates bias due to item sensitivity (assuming the RRT is understood and trusted). Table 4.3 is equivalent to Table 4.2, however, additionally reports the estimates for all RRT experimental conditions.

Table 4.3: Estimated proportions in percent and absolute response bias in percentage points for DQ and RRT

Sample Type	j	$\bar{y}_{j,svy}$ [95% C.I.]	$\bar{y}_{j,adm}$	B_j [95% C.I.]	n
UB II	DQ	0.870 [0.843;0.898]	1.000	**-0.130** [-0.157;-0.102]	579
	RRT	0.854 [0.816;0.892]	1.000	**-0.146** [-0.184;-0.108]	836
	DQ_RRT	0.906 [0.862;0.949]	1.000	**-0.094** [-0.138;-0.051]	180
Employed	DQ	0.021 [0.009;0.033]	0.030	-0.009 [-0.021;0.003]	564
	RRT	0.004 [-0.025;0.032]	0.042	**-0.038** [-0.067;-0.010]	955
	DQ_RRT	0.043 [0.001;0.085]	0.042	-0.001 [-0.041;0.043]	93

Contrary to the initial expectations, the response bias in the RRT condition differs significantly from zero: In the UB II sample receipt of welfare benefits is underreported by 14.6 percentage points and, in the employed sample, by 3.8 percentage points. As for the DQ condition, the relative bias is larger in the employee sample (91.3%) compared to the UB II sample (14.6%). Respondents who refused to apply the RRT are the ones who show the lowest levels of underreporting in both subsamples and thus seem to be the more accurate respondents (9.4%pts and 0.1%pts; also in relative terms: 9.4% and 2.4%).

We also expected to see that the RRT estimates are less biased compared to those in the DQ condition (though not necessarily significantly so). What can easily be seen in Table 4.3 is that the estimates of response bias in both RRT conditions are even larger compared to those in the direct questioning conditions.

Results of the one-sided t-test of the differences of the mean estimates of response biases across experimental groups ($B_{DQ} - B_{RRT}$, by subsample) are displayed in Table 4.4. If the response bias in the RRT condition were to be smaller compared to that in the DQ condition, we would expect to see a negative difference.

Table 4.4: Differential response bias of RRT compared to DQ

Sample Type	B_{DQ}	B_{RRT}	$B_{DQ} - B_{RRT}$ [95% C.I.]	t-statistic
UB II	-0.130	-0.146	0.016 [-0.031;0.064]	0.68
Employed	-0.009	-0.038	0.030 [-0.001;0.060]	1.89

While the difference of both response bias estimates in the UB II sample is statistically nonsignificant (p = 0.25), the response bias is 1.13 times higher in the RRT condition compared to the DQ condition. In the employed sample, the RRT performs even worse: The response bias is 4.35 times higher in the RRT condition compared to direct questioning and this result is statistically significant (p = 0.03).

To summarize some of the results for our initial research question: 1) Our particular implementation of the RRT cannot reduce bias in the estimated prevalence of basic income support in Germany, while 2) RRT performs significantly worse if the item under investigation is of a low prevalence rate, as in the case of our employed sample.

We can only speculate about the reasons for the poor performance of the RRT in our particular study. One reason might be that the initial assumption that unemployment benefit receipt is sensitive is not true. In that case, we would not expect to see the RRT producing estimates closer to the truth compared to direct questioning. The second argument might be that respondents do not apply the randomization procedure correctly, i.e., that either they do not flip coins at all or they do not adhere to the RRT instructions. In the first instance, this could mean that a face-to-face implementation, with an interviewer supervising the randomization procedure, could perform better (Holbrook and Krosnick, 2010a). The second issue is trust in the method: Despite understanding the method, it is also crucial that respondents trust the privacy protection provided by the RRT (Holbrook and Krosnick, 2010a; Coutts and Jann, 2011). While we can assume that unintentional noncompliance to the rules should not occur if the method is understood, i.e., respondents accidentally providing a wrong answer, trust is essential. 'Innocent' respondents might consciously decide to edit their answer and ignore the researchers' instructions: I.e., they might say 'No,' even if they should have said 'Yes' according to the randomization device,[8] if they lack trust in the method. Or, if prompted to answer truthfully, respondents might edit their answer and report a 'No' (even if the truth is 'Yes'), if they do not trust the method. These so-called 'cheaters' or 'non-compliers' lead to the fact that there

8 They refuse to provide a false positive answer when prompted to answer 'Yes.'

is still underreporting in the RRT (see also Clark and Desharnais, 1998; Boeije and Lensvelt-Mulders, 2002). A third reason could be the mode of data collection via telephone itself. Respondents might find it easier to 'cheat' on the phone than in a face-to-face mode.

This result is particularly relevant for future studies due to the cost implications: The increased costs in the RRT condition are due to—all other things being equal—a larger sample size, longer interview times, statistically more complex analyses, more intensive interviewer training and, most important, a higher respondent burden. Given our empirical evidence, additional costs of an RRT data collection for welfare receipt are not justified. Thus, in terms of bias versus efficiency, our results clearly favor direct questioning to collect data on welfare benefit receipt in Germany.

4.4.2 Is Response Bias Subgroup Specific?

The results so far indicate a tremendous amount of misreporting, contrary to the expectations in both experimental conditions. So far, this study has only considered bivariate comparisons between experimental groups.

The following section will now take a closer look at how misreporting differs between subgroups while controlling for a differential sample composition across both experimental conditions. Since individual-level data is available only for the UB II sample, further analyses are limited to this sample and inferences can only be drawn with respect to this specific population.[9] The following analysis investigates the mechanisms leading to response error across both experimental conditions in order to find out if the RRT performs differently in special subgroups. The dependent variable, 'accurate reporting,' will be modeled as a function of several individual characteristics for respondents in the UB II sample separately for each experimental split. In order to account for potential nonlinear relationships, all variables enter the regression equation categorically.

Table 4.5 displays the average marginal effects (AME) from logistic regression models (Stata version 12.1, rrlogit, Jann, 2011), modeling accurate reporting as a function of the indicators mentioned above. The AME is the average of discrete or partial changes over all observations and yields a straightforward interpretation of estimation results and allows comparison between models (Bartus, 2005; Mood, 2010). Subsequently, we will only report the AMEs providing an immediate interpretation of effect size.

9 Even if individual-level data were available for the employed sample, due to the low prevalence (close to zero) of receipt of UB II in that split, the statistical power to meaningfully detect group differences would be too low.

Table 4.5: Logistic regression models analyzing accurate reporting of receipt of UB II (average marginal effects and 95% confidence intervals)

Y: Accurate Reporting	Model 1: DQ	Model 2: RRT
Factors Contributing to Perceived Reporting Costs and Item Sensitivity		
No Employment (ref. Employed > 400€)	**0.118*** ** [0.055,0.180]	**0.095** ** [0.025,0.165]
Marginally Employed	0.017 [-0.050,0.083]	0.013 [-0.056,0.082]
Low/Med. ISEI (ref. N/A (Never Employed))	**0.069**+** [-0.000,0.137]	**0.063**+** [-0.010,0.136]
High ISEI	0.037 [-0.046,0.120]	-0.003 [-0.098,0.092]
Previous Socially Undesirable Response (Tax Honesty)	**-0.045**+** [-0.097,0.007]	**0.158** ** [0.039,0.276]
Reluctance (Item NR)	**-0.113** ** [-0.186,-0.039]	**-0.148*** ** [-0.227,-0.069]
Recipient Rate	0.050 [-0.221,0.322]	0.069 [-0.241,0.379]
Survey Process and Application of RRT		
DQ_RRT		0.049 [-0.020,0.117]
Language Skills (Poor)	**-0.057**+** [-0.120,0.005]	-0.058 [-0.128,0.013]
Tertiary Degree	0.028 [-0.072,0.128]	0.075 [-0.042,0.192]
Fast Response (ref. Mean Response)	0.038 [-0.030,0.106]	-0.004 [-0.070,0.061]
Slow Response	0.011 [-0.049,0.071]	0.042 [-0.030,0.114]
Controls		
Female	-0.001 [-0.051,0.049]	**0.049**+** [-0.005,0.103]
Age < 25 (ref. 25 to 40)	**-0.137*** ** [-0.199,-0.076]	**-0.168*** ** [-0.240,-0.097]
Age 41 to 57	0.007 [-0.055,0.069]	-0.028 [-0.109,0.053]
Age > 57	-0.008 [-0.122,0.105]	-0.063 [-0.164,0.039]
East Germany	0.009 [-0.050,0.069]	0.032 [-0.037,0.102]
Single Person Household	**0.092*** [0.020,0.163]	**0.071*** [0.001,0.140]
Model Fit		
N	579	1016
LR Chi^2 (df)	119.975 (17)	95.894 (18)
Pseudo R^2	0.27	0.09
AIC	360.433	966.079
BIC	434.576	1054.704

95% confidence intervals in brackets; + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Two models are estimated: Model 1 analyzes accurate reporting in the direct questioning condition and serves as a baseline for examining reporting accuracy. Model 2 replicates the same model in the RRT condition. We expect to see more accurate reporting in the RRT condition, especially for those variables related to perceived item sensitivity. Thus, all (negative) effects that we find in the direct split related to item sensitivity should become more positively related (or at least nonsignificant) to accurate reporting. This second model also presents insights regarding the question of which variables related to the survey process provide more accurate responses. The subsequent analyses will be organized along these expectations.

Focusing on the results of Model 1 in Table 4.5, those variables related to perceived item sensitivity are of particular interest. Unconditional on other covariates, as expected, respondents with no current employment are on average 11.8%pts more likely than respondents with an income of 400 Euro and above to report receipt of UB II. Marginally employed respondents do not differ systematically from the reference category. Regarding occupational status, we expect those respondents with a high (present or past) status to report receipt of UB II less often than the other categories. Contrary to our initial expectations, respondents with a high ISEI have a slight tendency to report more accurately compared to the reference category (no job), while those respondents with a low or medium status report receipt significantly more frequently (6.9%pts) than those who have never held a job before. Regarding the difference between respondents with a high ISEI and those with a low or medium ISEI, we observe no significant difference. The item 'previous socially undesirable answer' regarding tax honesty significantly explains accurate reporting, but in a surprising way: Respondents with an honest, i.e., more lenient, attitude towards tax dishonesty are on average 4.5%pts more likely to underreport the receipt of basic income support than those respondents displaying a less lenient attitude towards tax honesty. At first, this finding seems much at odds with what we would expect to see, i.e., behaving socially undesirably in one instance would result in a higher propensity to admit another undesirable characteristic. One potential explanation could be that given that 'tax dishonesty' is acceptable, misreporting on other characteristics is considered acceptable as well. Reluctance, i.e., item nonresponse with respect to household income contributes significantly to the explanation of underreporting of UB II (11.3%pts). Regarding the share of UB II recipients at the municipality level, we do not see any significant effect, supporting our hypothesis regarding the wrong level of measurement.

Those characteristics relating to the survey process contribute less to the explanation of accurate reporting. Poor language skills are the only significant influence and contribute to underreporting of UB II (5.7%pts). With respect to

the controls, we observe that young respondents, aged 24 and below significantly underreport receipt (13.7%pts). In line with our expectations, the indicator 'single person household' significantly improves reporting accuracy (9.2%pts). Both results support the argument that proxy reports with less knowledgeable persons on receipt of UB II are less accurate, since younger respondents are more likely to still live with their parents who apply for UB II for the entire household.

Turning our attention to Model 2—the RRT model—the results are strikingly similar, both in direction and magnitude. Contrary to what we would expect, those variables relating to perceived item sensitivity exert approximately the same influence as in the first model with one exception: previous socially undesirable response. Those respondents in the RRT condition now report on average 15.8% more accurately. This difference between both models is statistically significant, indicating that the RRT reduces social desirability concerns for those respondents. Given this evidence, the above explanation for this finding seems implausible. A different explanation might solve the puzzle: In Germany, tax dishonesty is largely associated with undeclared work/income. Given that UB II is based on accurate reporting of all forms of income and misreporting of income to the authorities is heavily pursued, these results are internally consistent.

To summarize, contrary to our expectation, the RRT does not elicit more accurate reports for those respondents for whom reports of UB II are especially sensitive, with the exception of one indicator. This indicates that the same misreporting mechanisms are at work in both experimental conditions.

Similar to Model 1, those characteristics relating to the survey process and the application of the RRT overall contribute less to the explanation of accurate reporting. Those respondents who refused the application of the RRT report more accurately than those respondents in the RRT condition (4.9%pts). We have anecdotal evidence from interviewer observations that those respondents either distrust the RRT or claim that they 'have nothing to hide' and want to be questioned directly. This effect is thus as we expect: These respondents should express lower levels of misreporting compared to the RRT condition. The effect size of lack of language skills is negative and roughly the same as in model 1, however, just fails to be statistically significant ($p = 0.101$). We assume that respondents who do not accurately understand what is asked of them in either condition (particularly so in the RRT) will not trust the method and therefore report (a 'self-protective') 'No' (Coutts and Jann, 2011). Thus, the result is as expected for both models. Remember that while a tertiary degree contributes to accurate reporting (2.8%pts) in Model 1, in Model 2 this effect is larger in comparison to Model 1, but not compared to the reference category (7.5%pts). Due to the small number of people holding a tertiary degree, confidence intervals are rather large for this estimate.

Further regression analyses were conducted but are not presented here: They account for the fact that if language skills are poor, neither educational degree will make a difference in the reporting accuracy. Assuming good language skills (essentially an interaction), our results show an even larger effect of university degree in Model 2. This suggests that the RRT reduces underreporting for these respondents: However, it remains unclear whether this effect is due to a better understanding of the RRT compared to the reference category (Poor German Skills and No Tertiary Degree) or the RRT guaranteeing anonymity and reducing item sensitivity for the more highly educated group. Response latency, i.e., the speed at which a respondent answers, is used as a measure for response quality. Surveying in the RRT condition by definition takes longer than a comparable direct question, since respondents have to follow the RRT protocol. In theory, irrespective of the experimental condition, a longer answering process could indicate more editing of the true response and thus a poorer data quality (Holtgraves, 2004). On the other hand, it could also be associated with higher quality information and processing in the RRT condition (Wolter 2012). Results for response latency exhibit no clear pattern across models and are nonsignificant: In Model 2, a slower response indicates on average more reporting accuracy (4.2%pts, 0.4%pts underreporting for fast reporters; this difference is statistically nonsignificant), while in Model 1, we observe both fast and slow responders to be more accurate compared to the reference category (3.8%pts and 1.1%pts).

With respect to the controls, we observe effects similar to those of Model 1, with the exception of women reporting on average more accurately in Model 2 (4.9%pts).

To summarize our results, we can replicate results from previous studies in Model 1, i.e., especially for characteristics relating to item sensitivity (employment status, occupational status, socially undesirable response, reluctance) and structural characteristics (age, single person household) (Kreuter et al., 2013). Contrary to our expectations, the RRT cannot resolve social desirability concerns for these items; as expected, structural influences persist. Our expectations regarding the survey process and the application of the RRT cannot be confirmed with our results. Overall, the poorer model fit in the second model indicates that some other mechanisms are at work that cannot be accounted for by our data.

4.5 Discussion and Conclusion

Our initial research question addressed the performance of the RRT for the estimation of welfare receipt compared to direct questioning. Empirical evidence suggests that the RRT does not reduce underreporting in the data collection on

welfare benefit receipt in a telephone survey. Our results show that it performs even worse in the employee sample, where the overall prevalence is close to zero. Thus, we conclude that the additional burden imposed on the respondent and the additional costs emerging from the data collection via RRT are not justified.

Insights into who underreports receipt of UB II were the main focus of the second research question. Inferences are limited to the population of UB II recipients in Germany only. We find significantly more accurate reporting across both methods for respondents who perceive reporting of UB II as less of a norm violation, i.e., respondents who are not employed (compared to those employed) or have a low or medium occupational status (compared to those who have never held a job before and those with a high status (RRT model only)). Respondents who admit to tax dishonesty report more accurately in the RRT model, however, less accurately in the DQ model, as do respondents who are unwilling to provide information on other items such as income. Thus, we can conclude that there is a tendency for underreporting whenever receipt of welfare benefits is perceived as more sensitive in both models.

If the RRT were to resolve the concerns of social desirability, we would expect to see differential effects or different mechanisms at work, across both methods for those items capturing sensitivity. Our results do not support this argument, since differences between models are statistically significant in one instance only: for those respondents previously having given a socially undesirable response. Furthermore, we expect that those items fostering understanding of the RRT increase reporting accuracy. While most effects point in the expected direction, they are statistically nonsignificant.

Overall, our findings regarding the poor performance of the RRT seem to be at odds with prior studies reporting a success of the RRT compared to direct questioning (Lensvelt-Mulders et al., 2005a). We can only speculate about the potential reasons for the failure of the RRT in our study. One argument discussed above relates to the potential lack of sensitivity of the item under study. If underreporting were not be caused by perceived sensitivity, then the RRT would not be expected to decrease bias. However, studies regarding the perception of welfare receipt would not support this argument (Bullock, 2006). Other arguments explaining the poor performance of the RRT, relate to 'cheating' and 'noncompliance.' For one, it remains unclear while on the telephone whether respondents really implement the randomization procedure (Holbrook and Krosnick, 2010a). In that instance a face-to-face mode would be more appropriate. A second concern is that respondents 'forced' by the randomization device to provide a false positive answer might decide not to comply with the RRT rules and reply 'No' instead of 'Yes' (Coutts and Jann, 2011). This concern cannot be

ruled out even in the f2f mode. This can be particularly problematic if the overall prevalence is close to zero, as is the case for one of our samples. This result is not surprising if we assume a fixed amount of 'non-compliers' in both samples: In relation to those complying to the RRT rules, the share of non-compliers would be much smaller in a high prevalence sample admitting to the sensitive item. Thus, in that instance, the 'true' effect of the RRT, i.e., had everybody complied, would only be slightly downward biased. If the same amount of noncompliance occurred in a low prevalence sample, this ratio would change drastically, introducing a larger distortion, potentially causing even negative prevalence estimates. This is exactly the result that we observe in our study.

The question remains, why we observe this amount of underreporting in absolute terms. To explain the overall deviance from the true values, it helps to have more background information on the nature of welfare support itself: Some forms of support subsumed under UB II are never directly received by the respondent and might not be experienced as such. To give an example, respondents sometimes only receive UB II in the form of housing assistance. This transfer payment is sometimes remitted directly to the landlord by the federal employment agency. Furthermore, while young adults (24 and younger) living in a benefit community officially receive UB II, they might not perceive themselves as recipients, since officially the money could be transferred to their parents. For these respondents we would not expect to see a difference between direct questioning and RRT, since this error is mostly due to encoding errors, as well as errors of question interpretation, and not due to item sensitivity. Thus, a certain amount of underreporting of welfare benefit receipt could always be expected.

To conclude, the evidence in our study supports the notion that the RRT performs particularly well in certain populations: those respondents with good language skills, those more highly educated, and those who take enough time to respond in the RRT condition, i.e., the correct application of the randomization process being observed in some way (note that many of the most successful RRT applications in the literature work with student populations). With populations with a lower educational background, the results can be very different. When the population of interest covers many individuals with poorer language skills, or a lower educational background, other techniques, such as the crosswise or triangular technique (Yu et al., 2008), might be a preferable method. These methods do not require a randomization device, are less of a cognitive burden for respondents, are easier to implement on the telephone and might thus perform better.

5 Discussion and Conclusion

The main question of this dissertation was whether we can elicit more truthful reports to sensitive labor market questions using so-called 'dejeopardizing' data collection techniques. We implemented the most prominent examples in two experimental surveys conducted in 2010, namely, the randomized response technique and the item count technique and compared each with standard direct questioning. Furthermore, to facilitate data collection for continuous sensitive information, we developed and applied the item sum technique.

The findings with respect to the RRT and the ICT are rather enlightening: Our empirical evidence is at best mixed. While the magnitude of estimates of undeclared work and unemployment benefits in our studies is comparable to those obtained in other surveys, neither the RRT nor the ICT consistently outperforms standard methods such as direct questioning and consistently reduces response bias ('more-is-better' assumption as well as using validation data). To state it even more clearly, in some instances both techniques even perform (significantly) worse compared to direct questioning.

These results, i.e., the lack of improvement in data quality for data collected on undeclared work and receipt of welfare benefits, are relevant for future studies due to the involved survey costs when implementing these special techniques. On the one hand, increased monetary costs in the RRT study and the ICT study are mostly due to—all other things being equal—the necessity of larger sample sizes to achieve the same level of statistical precision in the experimental conditions. Furthermore, longer interview times, more complex interviewer training and, last but not least, statistically more complex analyses contribute to higher overall costs. Not forgetting the additional cognitive burden imposed on the respondents and potential bias due to respondents refusing the application of these techniques (see the non-compliers in our RRT study). On the other hand, anecdotal evidence suggests that the RRT was very well received by many respondents and interviewers in terms of novelty or innovation and motivation. Taking all evidence into account, the additional respondent burden and the additional monetary costs of an RRT or ICT data collection do not seem justified in the context of undeclared work and welfare benefit receipt.

Evidence regarding the newly developed item sum experiments, however, shows that the IST is a rather promising data collection technique for continuous sensitive variables. For both items under study, the IST yielded higher, and thus presumably more valid, estimates in three out of four instances. However, this effect was significant only for earnings from undeclared work. Survey researchers aiming at measuring sensitive behaviors at an ordinal or continuous scale could

benefit from using the IST. In contrast to other techniques, such as continuous RRT schemes, no randomizing device is necessary, respondent burden can be assumed to be lower, the IST can be easily conducted over the phone and by implementing a double-list design increased survey costs could be kept in check.

5.1 Contribution

Reverting to the initially presented research questions, the main findings of this dissertation comprise several aspects:

For the *first research question*—i.e., prevalence of undeclared work comparing DQ, RRT, ICT as well as substantive analyses—presented in Chapter 2 we could demonstrate that in comparison to direct questioning techniques concerning undeclared work ('more-is-better' assumption), the RRT provided slightly lower prevalence estimates in our employee sample, while RRT prevalence estimates for undeclared work in our benefit recipients sample were higher (though nonsignificant in one out of two instances). While this result is not in favor of the RRT in general, it also does not suggest that the RRT performs worse in our 'more difficult' target population of benefit recipients. On the contrary, we have empirical evidence that the RRT (as well as the ICT) elicits more accurate reports of undeclared work ('more-is-better' assumption), the more sensitive the item is perceived to be, i.e., when higher penalties can be expected. This would explain the findings for 'undeclared work for a company' or for benefit recipients who could be charged with social security fraud when admitting to conduct undeclared work. These results are in accordance with the literature suggesting a better performance with increasing sensitivity of the topic (Lensvelt-Mulders et al., 2005a; Tourangeau and Yan, 2007).

Similar to the findings for the RRT, the evidence regarding the ICT is mixed: The ICT only elicits more accurate responses for the item 'undeclared work for a company.' This leads us to the overall conclusion that neither the RRT nor the ICT consistently outperforms direct questioning.

Recent studies suggest that the ICT outperforms the RRT in bias reduction, however, at the same time show that the ICT is somewhat less efficient statistically (Corstange, 2009; Holbrook and Krosnick, 2010a, b; Coutts and Jann, 2011): While respondents in the RRT sometimes believe in some kind of trick, do not understand the rules or consciously 'cheat,' the ICT seems easier to administer, more convincing and more applicable in large scale surveys. Further, since the ICT does not require a lengthy introduction or a randomization device—and thus imposes less cognitive burden on the respondent—researchers see a solution for the problem of social desirability in this technique (Coutts and Jann, 2011; Glynn, 2013). Due to the

use of different samples (and data collection agencies) in both studies, however, the results of the RRT and the ICT studies are, strictly speaking, not comparable. Daring a very careful comparison—by assuming the employee sample and the general population sample to be sufficiently similar—our results do not support these hypotheses: Compared to the RRT, the ICT prevalence estimates are lower for the item 'undeclared work for a private person' and higher for the item 'undeclared work for a company.'

Our substantive contributions concern insights into who engages in undeclared work and who underreports welfare benefit receipt using logistic regression models. Regarding engagement into undeclared work, we pooled both RRT samples. In the analyses of 'undeclared work for a private person' or 'for a company,' the perceived share of undeclared work in one's own network proved to be a robust explanatory factor in our substantive models (norm hypothesis). Other factors contributing to the explanation for engaging in undeclared work that were found to be statistically significant were: receipt of benefit (as a proxy for monetary gains from undeclared work), the number of people potentially helping with the job search, active organizational membership, as well as acceptance and approval of undeclared work. Those are all arguments within the utility and opportunity hypotheses framework. Furthermore, respondents aged 34 and younger are more likely to engage in undeclared work compared to those aged 35 and older. The main substantive conclusion is that we find particular evidence for the norm hypothesis in our models explaining undeclared work as well as a significant RRT effect (though mostly driven by the UB II sample).

Our *second research question* addressed in Chapter 3 concerned the development and presentation of a new data collection method to improve data collection for continuous sensitive information—the item sum technique (IST). This new technique yielded significantly higher estimates of earnings from undeclared work than direct questioning in both the employee sample and the benefit recipients sample. Relying on the 'more-is-better' assumption, this leads us to conclude that we obtain more accurate reports from respondents engaged in undeclared work. While the IST effect was significant for earnings from undeclared work for both samples, IST estimates regarding hours of undeclared work were more ambiguous: Though higher in the employee sample, estimates in the benefit recipient sample were lower compared to direct questioning. Furthermore, there was no evidence of significant interaction effects, i.e., differential item sum effects, demonstrating that the technique seems to work equally well for respondents with differing cognitive abilities.

The *third research question* embedded in the RRT study was devoted to the opportunity to assess the RRT without having to rely on the 'more-is-better'

assumption. Chapter 4 analyzed the performance of the RRT for the estimation of welfare receipt compared to direct questioning. Again, our empirical evidence illustrates that the RRT does not reduce underreporting. Our results show that the RRT performs significantly worse than direct questioning in the employed sample, where the overall prevalence is close to zero. Only the direct questioning technique contained the true value obtained from administrative data. For the benefit recipients, the poorer performance of the RRT compared to direct questioning was nonsignificant. However, neither the RRT nor direct questioning contained the true value for this sample.

In order to assess who underreports the receipt of welfare benefits, we ran two separate logistic regressions (direct questioning and RRT), modeling accuracy of the reports of welfare benefit receipt as a function of different covariates. Inferences for this item are limited to the population of UB II recipients in Germany. In the models, we focused particularly on explanatory factors related to the perceived item sensitivity, as well as those factors related to the survey process and the application of the RRT itself. Reporting accuracy is significantly higher in both models for respondents for whom it can be reasonably argued that they perceive reporting of UB II as less of a norm violation. Remembering the model of the response process, we would expect to see differential effects across both methods for those items capturing sensitivity if the RRT were to resolve these concerns of social desirability. Our results do not support this argument. Further, we expect that those items fostering understanding of the RRT would contribute to an increased reporting accuracy. While most effects point in the expected direction, they are statistically nonsignificant.

5.2 Limitations

Overall, our findings confirm the results of more recent studies reporting mixed results of the RRT and ICT compared to direct questioning (Biemer et al., 2005; Holbrook and Krosnick, 2010a; Coutts and Jann, 2011; Wolter, 2012); the prior success of the RRT and positive evidence for the ICT cannot be replicated (Lensvelt-Mulders et al., 2005a; Tourangeau and Yan, 2007). Our findings are, however, subject to certain limitations.

Remember that one limitation that affects all surveys on undeclared work equally is that these mostly capture undeclared work conducted for and by private individuals. Undeclared work conducted by a company for a company is seldom captured. Thus, even if all bias due to social desirability were to be removed, these estimates of the magnitude of undeclared work obtained in surveys can still be considered a lower bound.

First, due to the particular design and data availability (administrative records), in-depth and substantive analyses were only conducted within the RRT study and are limited to the benefit recipient subsample. Generalizations to the general population can only be made with respect to the overall performance of the RRT, assuming that the employee sample is sufficiently similar. No such limitations exist for the ICT study. Furthermore, we cannot disentangle whether the 'failure' of the RRT or the ICT is due our specific design or whether the RRT and the ICT do not work in general for the collection of data on undeclared work and welfare benefit receipt. The underlying argument being that it could be perceived as easier to cheat on the phone. However, since the main goal was to evaluate the use of the RRT in the large scale telephone surveys with this particular sampling, this leaves the overall conclusions unaltered.

Second, an ideal study for evaluating the RRT, ICT or IST would not rely on the 'more-is-better' assumption but would instead use validation data with known true scores for the sensitive variables. Opportunities for validation studies are very difficult to find, and data protection issues arise if individual-level data from different sources are linked without the informed consent of the respondents. Due to the nature of undeclared work, we were not able to validate these individual survey reports directly, but had to rely on the 'more-is-better' assumption for evaluating the RRT, the ICT and the IST. While this is certainly a major limitation, this assumption is often used in the literature.

We also have evidence, that there is noncompliance in the RRT condition despite a careful introduction and explanation to the respondents. Our negative prevalence estimate in the RRT employee sample for undeclared work suggests that not all respondents complied with the instructions. Due to lack of additional data, we can only speculate about the potential reasons for noncompliance. For one, it remains unclear on the telephone whether respondents really implement the randomization procedure (Holbrook and Krosnick, 2010a). In this instance, a face-to-face mode would seem more sensible. A second concern, however, could also not be entirely ruled out in the face-to-face mode: Respondents 'forced' by the randomization device to provide a false positive answer might decide not to comply with the RRT rules and reply 'No' instead of 'Yes' (Coutts and Jann, 2011). This can be particularly problematic if the overall prevalence is close to zero, as is the case for undeclared work or welfare benefit receipt in the employed sample. If we assume a fixed total amount of noncompliance in both samples, the share of non-compliers would be comparatively small in a high prevalence sample in relation to those complying to the RRT rules. Thus, the 'true' effect of the RRT had everybody complied, would be only marginally attenuated. If the same total amount of noncompliance occurred in a low prevalence sample, this ratio would

change drastically, introducing a larger bias, potentially causing even negative prevalence estimates. This is exactly the result we observe in our study.

Taking all evidence into account and analyzing complete and partial interviews, a total of 369 respondents refused the application of the RRT (15.8% of complete and partial interviews). This considerable share of respondents who did not remain in initially assigned condition thus compromised the randomization of experimental groups. All of the RRT analyses control for this fact. Using intention-to-treat and instrumental variables strategies in the IST-study, for example, we believe that we convincingly demonstrated that our findings are robust despite this problem.

We did not encounter similar problems in the ICT study, nonetheless, a closer look at the data of the ICT study also reveals problems, such as 'ceiling effects' (Glynn, 2013). Ceiling effects occur whenever all items, the sensitive and the nonsensitive items, apply to a respondent. In this instance, anonymity is not granted any longer. Furthermore, for both, the ICT as well as the IST, it is crucial that the answers to the non-sensitive items are independent of the question format, that is, that the answers do not depend on whether they are given directly or serve as a 'filler' item in the item count/sum format (assumption of 'no-design-effect,' see Tsuchiya and Hirai (2010); Blair and Imai (2012)). Thus, careful designing of the survey instrument (i.e., the innocuous questions) and even more extensive pretesting seem advisable.

Third, regarding the results of consistently higher response bias for reporting of welfare benefits in the RRT condition, another potential concern—aside from noncompliance to the instructions—is related to the perceived item sensitivity. While undeclared work is sensitive in all theoretical dimensions we discussed, welfare benefit might be considered as less sensitive than we assumed. If underreporting is not caused by perceived sensitivity but is due to other mechanisms, then the RRT would not decrease bias. The alternative hypothesis regarding the amount of underreporting in absolute terms involves structural arguments relating to the understanding and the interpretation of the question itself. Some forms of transfer payments subsumed under UB II are never actually received by the respondent. To give an example, respondents sometimes only receive UB II in the form of housing assistance. This transfer payment can then be remitted directly to the landlord by the federal employment agency. Further, while young adults living in a benefit community with their parents officially receive UB II, they might not view themselves as recipients, but only their parents. While interviewers were alerted to probe if there were indications for this problem, we find evidence for this argument in our analyses. Since this error is mostly due to errors of encoding, as well as errors of question interpretation and not due to item sensitivity, we would not expect to see a difference between direct questioning

and RRT for these respondents. Thus, a certain total amount of underreporting of welfare benefit receipt could always be expected.

Fourth, while survey researchers aiming at measuring sensitive behaviors on a continuous scale could benefit from using the new IST technique, this study is only a first step in the development and evaluation of the IST. Considering 'significance' as the main criterion, our analyses show that the IST was particularly successful for only one of the two items, i.e., the second item 'earnings,' while the effect is absent in both samples for the other item, i.e., the first item 'hours.' We believe that this effect is systematic, and that potential learning effects might explain the difference in the findings. Despite one 'training' example prior to the IST, we therefore suggest including several suitable training questions when employing novel questioning techniques such as the IST in a survey.

Another major limitation of the IST is that the continuous sensitive questions were preceded by a filter question on whether any undeclared work was carried out at all in the direct questioning condition, while due to the nature of the IST no such filter question is necessary in the item sum condition. We know from prior studies that 'filtering' might downwardly bias the estimates in the direct questioning condition. Given our design and the effect sizes, however, it seems unreasonable to assume that these differences could be solely a result due to filtering.

5.3 Implications for Future Research

As illustrated by the above qualifications, further experimental research is needed to fully understand the mechanisms at work when respondents are confronted with sensitive questions, especially in the newly developed IST. Furthermore, future research on sensitive questions could evaluate other dejeopardizing techniques that can be implemented on the phone. This is highly relevant when dealing with large population surveys or special populations like ours, especially in the future when even more restrictive budgets to conduct surveys can be expected. When the population of interest covers many individuals with poorer language skills or a lower educational background, other, more recently developed techniques to reduce response bias, such as the crosswise or triangular technique (Yu et al., 2008), might be a preferable method. These methods do not require a randomization device, are less of a cognitive burden for respondents, easier to implement on the telephone and could thus perform better than the RRT or the ICT (Coutts et al., 2011).

Furthermore, if (individual-level) auxiliary data are available, future research on sensitive questions could aim at making estimates obtained with these special techniques more efficient. Irrespective of which method is used, larger sample sizes are typically needed to achieve the same level of precision as comparable

direct questioning. In order to be more efficient or to even allow estimates that are aggregated at a lower level such as by industry or state for undeclared work or welfare benefit receipt—thus, yielding even smaller sample sizes—the ideas of small area estimation could prove fruitful (Fay and Herriot, 1979; Battese et al., 1988; for an overview see Ghosh and Rao, 1994). Borrowing strength from larger areas using auxiliary data could then lead to a reduction in variance.

So far, we are also not aware of any studies using data generated by means of special techniques—RRT, ICT or in that vein IST or crosswise data—as explanatory variables. Existing studies typically explore these variables as dependent variables. However, if—unlike in our study—these techniques successfully reduce bias, accounting for variables, such as undeclared work, to potentially correct predictions of income and entire distributions seems advisable. Programs, such as the R-package SIMEX are useful in this case (Carroll et al., 1996; Lederer and Küchenhoff, 2006), though one would have to be careful with respect to the endogenous relationship of income and undeclared work.

Last but not least, substantively, more research is needed to assess what respondents understand by undeclared work. While this affects the differential effect of direct questioning technique and our special techniques to a lesser extent, it affects the total amount. Despite using 'standard' definitions and questions, we also conducted cognitive pretests. Due to these, respondents were asked to classify three examples of different 'tasks' as declared or undeclared work at the end of the main survey. The examples for (un)declared work were to a large extent misclassified (as declared work; false negative), suggesting that the respondents had either meanwhile forgotten the definition or had decided to ignore it. In our data this share of false negatives is estimated at 11.8% in the DQ condition, at 11.9% in the RRT condition and at 17.2% in the RRT non-complier condition, while the share of false positives is estimated at 39.0% DQ, at 42.6% RRT and at 46.0% DQ_RRT. Supervising many of the survey interviews, anecdotal evidence supports the notion that respondents differentiate as to what exactly constitutes undeclared work according to frequency and regularity, amount of income and purpose, i.e., what the money was needed for. It was clearly not considered as a binary choice (Yes/No), and therefore should be measured on a more appropriate scale. More qualitative research combined with factorial survey design could help to explore these notions in depth and help to improve the questionnaire items. Further research is also needed with respect to welfare benefit receipt since a similar criticism holds for this item. We have evidence for structural deficiencies with the survey question itself that lead to a substantive amount of underreporting. Tackling these issues for both items could help to reduce measurement error that is not due to concerns of social desirability.

A Appendix to Chapter 2

A.1 RRT Instructions

"I will now introduce you to a technique, that will allow you to keep your personal experiences anonymous by means of a coin flip. Even if this might sound strange to you, I kindly ask you to help us try this new method. This method is scientifically approved and is fun. Would you please get a paper, a pencil and three coins?

You will be able to answer all of the following questions either with 'Yes' or 'No.' Before answering each question, I would kindly ask you to flip the three coins. Please do not tell me the outcome of this coin flip. According to the outcome, please answer as follows:

- *3 tails; please always respond with 'Yes'*
- *3 heads; please always respond with 'No'*
- *a mixture; i.e., a combination of heads and tails, such as 2 heads and 1 tail, please respond truthfully.*

As you can see, chance decides whether you actually respond to the question or provide a surrogate answer. Thus, your privacy is always protected. I, as the interviewer, will never know the result of your coin toss. Thus, I can never know, why you respond with 'Yes' or 'No.' Do you have any further questions regarding the technique?

Let us walk through one example together.

If you flip 3 heads, and I ask you if you are 18 years or older, what would you reply? (Int: Pause; let the respondent reply first. 'No,' according to the rule)

If you flip 3 tails, and I ask you if you are 18 years or older, what would you reply? (Int: Pause; let the respondent reply first. 'Yes,' according to the rule)

If you have a mixed result, e.g., flip 2 heads and 1 tail, and I ask you if you are 18 years or older, what would you reply? (Int: Pause; let the respondent reply first. The response has to be 'Yes'[1])

Do you have any further questions?"

(Note to the reader: If there were further questions, the rules were repeated and a new example provided.) (Translated from German)

1 This is a requirement of our sampling design.

A.2 IST Long-List Instructions

"Thank you very much. We are now done with the coin flip method. Please, as of now, always respond to each question truthfully. However, please keep the paper and the pencil.

I will now read two blocks of two questions each to you. The response to each of these questions is numerical. However, it is possible that you will respond with 'zero' to either one or both questions.

I would ask you to write down your individual responses to each question. Afterwards, I would ask you to add both responses and only tell me the result. It is important that you do NOT tell me your answer to the single questions. As you see, I can never know why you give a certain response.

Let us walk through one example together. I will now read two questions to you:

1.: How many persons live in your household, including yourself ?
Please do NOT tell me the answer to that question, but write it down.

2.: How many persons living in your household are aged 18 or older, including yourself ?

Please do NOT tell me the answer to that question, but write it down.

(Note to the reader: Interviewers were trained to leave respondents enough time at each stage)

Thank you very much. What is your result?

(Note to the reader: These particular questions were chosen for demonstration purposes, only because we knew the responses to the individual questions from previous survey information. Thus, we were able to 'validate' the response and check if the respondent understood the method.)

Do you have any further questions regarding the procedure?" (Translated from German)

A.3 ICT Instructions and ICT Lists

"I will now read blocks of 3 to 4 questions each to you. Please indicate how many of these questions apply to you. Please do NOT to tell me how you respond to individual questions within the list!

Do you have any further questions regarding the procedure?" (Translated from German)

Table A.1 provides an overview of the exact items used in our study.

Table A.1: Lists: item count technique (translated from German)

List	Operationalization	Short List	Long List
1a	Did you ever keep a diary capturing monthly household expenditures?	X	X
	Does your household own more TV's than there are members of the household?	X	X
	Do you own a mobile phone?	X	X
	Have you engaged in any undeclared work for a private person this year?		X
1b	Do you use public transportation on more than 5 days per week?	X	X
	Are you covered by liability insurance?	X	X
	Did you grow up in the countryside?	X	X
	Have you engaged in any undeclared work for a private person this year?		X
2a	Have you been invited to a job interview this year?	X	X
	Do you pay taxes for your dog?	X	X
	Do you have health insurance?	X	X
	Have you engaged in any undeclared work this year for a company, which paid you without reporting your income to the authorities?		X
2b	Have you ever used public transportation without a valid ticket?	X	X
	Are you allowed to carry a gun?	X	X
	Do you have a driver's license?	X	X
	Have you engaged in any undeclared work this year for a company, which paid you without reporting your income to the authorities?		X

A.4 Prevalence Estimates Undeclared Work

Table A.2: Prevalence estimates undeclared work

	RRT Study						ICT Study	
	Employees			Benefit Recipients			General Population	
	DQ	RRT	DQ_RRT	DQ	RRT	DQ_RRT	DQ	ICT
Private Person								
Est.	1.77	1.22	1.09	3.29	4.07	2.22	3.14	0.26
% C.I.	[0.68,2.86]	[-1.66,4.11]	[-1.04,3.22]	[1.83,4.74]	[0.79,7.34]	[0.06,4.38]	[1.81,5.01]	[-5.33,5.86]
Company								
Est.	0.53	-0.86	1.09	1.56	5.98	0.55	1.21	6.41
% C.I.	[-0.07,1.13]	[-3.60,1.88]	[-1.04,3.22]	[0.55,2.57]	[2.58,9.38]	[-0.53,1.64]	[0.24,2.17]	[1.21,11.62]

A.5 Overview of Items and Operationalizations (RRT study)

Table A.3: Operationalizations

Variable	Definition/Question	Label	Frequencies
Sample	Sample Status	Employees	1,613 (50.2%)
		Benefit Recipients	1,598 (49.8%)
Undeclared Work for a Private Person	Have you engaged in any undeclared work for a private person this year?	Yes	292 (9.1%)
		No	2,912 (90.9%)
Undeclared Work for a Company	Have you engaged in any undeclared work this year for a company, which paid you without reporting your income to the authorities?	Yes	269 (8.4%)
		No	2,936 (91.6%)
Utility from Undeclared Work	Sample indicator combined with income. "How much did you earn last month? If you held multiple jobs, please report the income from all of them together. Please indicate your gross income, that is your income before the deduction of tax and social security contributions. <Self-employed> For self-employed jobs, please instead indicate your monthly profit before tax. <Employees> If you received special payments last month, such as a Christmas bonus or back payments, do not include these. However, do include any pay for overtime." (Open-numerical response, categorical response scale in case of item nonresponse)	Employee w. Income ≤€800	415 (12.9%)
		Employee w. Income >€800	1,198 (37.3%)
		Benefit Recipient w. Income ≤€800	1,259 (39.2%)
		Benefit Recipient w. Income >€800	339 (10.6%)
Preferred Working Hours	"Please assume that you could choose your own number of working hours. Please also consider the necessary income: How many hours would you yourself like to work currently?" Categorized in reference to the actual current working hours (including overtime).	Adequate (≤2 hours)	1,686 (52.5%)
		Inadequate (≥3 hours)	1,525 (47.5%)
Risk Perception	Product of the following two questions, divided by 100. Categorization into empirical quartiles: 1) "Please imagine 100 persons who are engaged in undeclared work. How many of these 100 individuals will, in your opinion, be caught by the authorities?" 2) "Now please now imagine someone who has been engaged in undeclared work for 6 months, earning €12,000 for full-time work. How much will the monetary penalty in your opinion be, in the event that this individual is discovered by the authorities? Please do not include back duties."	Hardly Any Risk (€0 to €120)	841 (26.2%)
		Low Risk (€121 to €600)	830 (25.8%)
		Medium Risk (€601 to €2,500)	751 (23.4%)
		High Risk (€2,501 and more)	789 (24.6%)

Variable	Definition/Question	Label	Frequencies
Occupational Status	Generated (Hendrickx, 2002, iskoegp) for current or last job. Coding of ISCO 88 according to Erikson Goldthorpe Portocarero class scheme (Erikson et al., 1979). Aggregation to the following	N/A (Never Employed)	399 (12.4%)
		Low-High Controllers	692 (21.6%)
		Routine Non-Manual	789 (24.6%)
		Self-Employed	95 (3.0%)
		Manual Supervisor & Skilled Manual	435 (13.5%)
		Semi-Skilled/Unskilled Manual	801 (24.9%)
Networks (Job Search)	"How many persons do you know, outside of your household, who could support you finding a job and would really do so?" Collapsing one category 'Nobody' and the remaining cases into terciles.	Nobody	1,233 (38.4%)
		1 to 4 Persons	753 (23.5%)
		5 to 10 Persons	819 (25.5%)
		11 Persons and more	406 (12.6%)
Organizational Membership	"Now we would like to know if you are actively engaged in an organization or a club? Are you an active member of ... a labor union, 2) a political party, 3) a church or religious organization, 4) a club for music, sports or culture, 5) another organization that I have not mentioned?"	None	1,859 (57.9%)
		At Least One	1,352 (42.1%)
Undeclared Work in Network	"In your opinion, how large is the proportion of persons within your entire network who are engaged in undeclared work?" Collapsing one category 'Nobody' and the remaining cases into terciles.	0%	1,456 (45.3%)
		1 to 3%	603 (18.8%)
		4 to 10%	684 (21.3%)
		11% and more	468 (14.6%)
Approval of Undeclared Work	Sum score of seven items (1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree: "My personal tax burden due to taxes and social security contributions is too high" (recoded). "Tax honesty pays off for me personally." "Undeclared work, causes honest people to have personal disadvantages." "Given the high tax and social security burden, it's the state's own fault that so many people engage in undeclared work" (recoded). "People engaging in undeclared work should be reported." "Given all the bureaucracy and the formalities, it is not surprising that so many people engage in undeclared work" (recoded). "Please imagine, that your neighbor would engage undeclared workers on a regular basis. Would you ... (1 = strongly approve, 2 = approve, 3 = disapprove, 4 = strongly disapprove)" (recoded).	7 to 14 (Disapproval)	945 (29.4%)
		15 to 16	818 (25.5%)
		17 to 19	895 (27.9%)
		20 to 28 (Approval)	553 (17.2%)
Gender	Coded by interviewer, derived from name and voice.	Female	1,709 (53.2%)
		Male	1,502 (46.8%)
Age	"Would you please tell me how old you are?" (open-numerical scale in years)	18 to 34	1,065 (33.2%)
		35 to 49	1,176 (36.6%)
		50 to 70	970 (30.2%)

Variable	Definition/Question	Label	Frequencies
Formal Training	"What is the highest formal training you obtained? Or are you currently in training (vocational training, pupil)? Do you have... no vocational qualification and are currently not in training? a vocational qualification? a college or university degree?"	Pupil / No Degree Vocational Training Tertiary Degree	656 (20.4%) 2,110 (65.7%) 445 (13.9%)
Migrant Background	Indicator if respondent or respondent's parents have migrant background: "Let us now move on to the topic of country of birth, nationality and languages. Were you born in Germany?" "Was either of your parents born outside of Germany?"	Yes No	934 (29.1%) 2,277 (70.9%)
Residence	Aggregated using state.	East Germany West Germany	761 (23.7%) 2,450 (76.3%)

A.6 Logistic Regression Models Analyzing Undeclared Work Based on Listwise Deletion

Table A.4.: Logistic regression models analyzing undeclared work based on listwise deletion

Y: Undeclared Work for a ...		Model 1: Private Person	Model 2: Company
		AME [95% C.I.]	AME [95% C.I.]
Methods Effect			
Experimental Condition (ref. DQ)	RRT	**0.030*** [0.003,0.056]	**0.059***** [0.029,0.089]
	DQ_RRT	-0.006 [-0.049,0.037]	-0.017 [-0.079,0.046]
Utility Hypothesis			
Utility (ref. UB II w. Income ≤ €800)	Employee w. Income ≤€800	-0.032 [-0.079,0.015]	-0.024 [-0.059,0.011]
	Employee w. Income > €800	-0.024 [-0.062,0.013]	**-0.062*** [-0.110,-0.015]
	UB II w. Income > €800	0.008 [-0.027,0.043]	0.000 [-0.036,0.037]
Pref. Working Hours (ref. Adequate (≤ 2))	Inadequate (≥ 3)	0.011 [-0.019,0.042]	-0.000 [-0.030,0.030]
Cost Hypothesis			
Risk Perception (ref. €0 to €120)	Low Risk (€121 to €600)	0.001 [-0.027,0.028]	-0.005 [-0.037,0.027]
	Medium Risk (€601 to €2,500)	0.004 [-0.024,0.032]	-0.001 [-0.029,0.028]
	High Risk (> €2,500)	-0.034 [-0.076,0.009]	-0.009 [-0.040,0.022]
Opportunity Hypothesis			
Occupational Status (ref. (Semi)Unskilled Manual)	N/A (Never Employed)	-0.020 [-0.061,0.021]	-0.021 [-0.056,0.014]
	Low-High Controllers	-0.007 [-0.046,0.032]	-0.005 [-0.043,0.034]
	Routine Non-Manual	-0.010 [-0.044,0.024]	-0.008 [-0.038,0.022]
	Self-Employed	0.025 [-0.032,0.082]	-0.015 [-0.077,0.046]
	Manual Supervisor & Skilled Manual	0.024 [-0.008,0.055]	-0.018 [-0.061,0.026]
Networks Job Search (ref. Nobody)	1 to 4 Persons	**0.037*** [0.003,0.072]	0.009 [-0.024,0.043]
	5 to 10 Persons	**0.042*** [0.009,0.075]	0.023 [-0.005,0.051]
	≥ 11 Persons	**0.041*** [0.002,0.081]	0.023 [-0.010,0.056]

Continued on Next Page

Y: Undeclared Work for a ...		Model 1: Private Person	Model 2: Company
Membership in an Organization (ref. None)	At Least 1	0.019	0.010
		[-0.004,0.043]	[-0.012,0.032]
Norm Hypothesis			
Undeclared Work in Network (ref. Nobody)	1 to 3%	**0.097***	0.017
		[0.022,0.173]	[-0.021,0.054]
	4 to 10%	**0.102****	0.011
		[0.027,0.177]	[-0.032,0.054]
	≥ 11%	**0.126*****	**0.049****
		[0.052,0.201]	[0.013,0.086]
Approval of Undeclared Work (ref. Disapproval 7 to 14)	Some Disappr. (15 to 16)	-0.001	0.019
		[-0.048,0.046]	[-0.015,0.054]
	Some Appr. (17 to 19)	0.020	0.006
		[-0.021,0.061]	[-0.032,0.043]
	Approval (20 to 28)	**0.042***	0.028
		[0.004,0.079]	[-0.006,0.062]
Controls			
Gender (ref. Male)	Female	-0.008	-0.017
		[-0.034,0.017]	[-0.043,0.008]
Age (ref. 35 to 49)	Age ≤ 34	0.015	**0.039****
		[-0.011,0.041]	[0.012,0.067]
	Age ≥ 50	0.017	-0.031
		[-0.016,0.051]	[-0.105,0.044]
Formal Training (ref. Vocational Training)	Pupil/No Degree	0.003	-0.001
		[-0.028,0.034]	[-0.028,0.026]
	Tertiary Degree	**-0.069**$^{+}$	-0.008
		[-0.144,0.007]	[-0.057,0.042]
Migr. Background (ref. None)	Yes	0.002	-0.013
		[-0.023,0.027]	[-0.038,0.013]
Residence (ref. West Germany)	East Germany	0.003	-0.001
		[-0.023,0.030]	[-0.028,0.027]
Model Fit			
N		2,515	2,517
LR Chi^2 (df)		113.231 (31)	81.928 (31)
AIC		1487.477	1366.113
BIC		1668.208	1546.869

95% confidence intervals in brackets; $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.7 Logistic Regression Models Analyzing Undeclared Work by Experimental Conditions

Table A.5: Logistic regression models analyzing undeclared work by experimental conditions

Y: Undeclared Work for a ...		Model 1: DQ Private Person	Model 1: RRT Private Person	Model 2: DQ Company	Model 2: RRT Company
		AME [95% C.I.]	AME [95% C.I.]	AME [95% C.I.]	AME [95% C.I.]
Methods Effect					
Experimental Condition (ref. RRT)	DQ_RRT		**-0.033**$^{+}$ [-0.071,0.006]		**-0.042*** [-0.081,-0.003]
Utility Hypothesis					
Utility (ref. UB II w. Income ≤ €800)	Employee w. Income ≤ €800	-0.018 [-0.055,0.019]	-0.039 [-0.087,0.008]	-0.002 [-0.028,0.024]	**-0.044*** [-0.083,-0.005]
	Employee w. Income > €800	-0.008 [-0.040,0.024]	**-0.041**$^{+}$ [-0.085,0.003]	-0.021 [-0.059,0.016]	**-0.102***** [-0.159,-0.045]
	UB II w. Income > €800	0.022 [-0.008,0.052]	-0.028 [-0.076,0.020]	0.006 [-0.022,0.033]	-0.020 [-0.060,0.019]
Pref. Working Hours (ref. Adequate (≤ 2))	Inadequate (≥ 3)	-0.015 [-0.041,0.012]	0.026 [-0.009,0.061]	-0.017 [-0.042,0.008]	-0.001 [-0.030,0.027]
Cost Hypothesis					
Risk Perception (ref. €0 to €120)	Low Risk (€121 to €600)	-0.009 [-0.032,0.015]	0.013 [-0.018,0.044]	-0.013 [-0.040,0.014]	0.018 [-0.013,0.049]
	Medium Risk (€601 to €2,500)	0.003 [-0.018,0.025]	-0.017 [-0.053,0.020]	0.000 [-0.022,0.023]	0.010 [-0.020,0.039]
	High Risk (> €2,500)	-0.029 [-0.065,0.007]	-0.037 [-0.096,0.022]	-0.003 [-0.028,0.022]	-0.003 [-0.033,0.027]
Opportunity Hypothesis					
Occupational Status (ref. (Semi)Unskilled Manual)	N/A (Never Employed)	0.022 [-0.012,0.056]	**-0.061*** [-0.117,-0.005]	**0.030**$^{+}$ [-0.001,0.061]	**-0.038**$^{+}$ [-0.076,0.000]
	Low-High Controllers	0.022 [-0.006,0.051]	-0.014 [-0.070,0.042]	0.009 [-0.026,0.044]	0.020 [-0.014,0.054]
	Routine Non-Manual	0.013 [-0.016,0.041]	-0.021 [-0.060,0.017]	0.020 [-0.008,0.047]	-0.316 [-20.351,19.718]
	Self-Employed	0.005 [-0.045,0.055]	0.021 [-0.031,0.072]	omitted	-0.011 [-0.059,0.037]
	Manual Supervisor & Skilled Manual	0.018 [-0.010,0.047]	0.030 [-0.010,0.070]	omitted	-0.009 [-0.039,0.021]
Networks Job Search (ref. Nobody)	1 to 4 Persons	0.011 [-0.015,0.036]	**0.054*** [0.004,0.104]	-0.011 [-0.037,0.014]	0.013 [-0.024,0.049]
	5 to 10 Persons	0.003 [-0.023,0.028]	**0.077**** [0.029,0.126]	-0.005 [-0.027,0.017]	**0.035*** [0.005,0.065]
	≥ 11 Persons	0.010 [-0.018,0.038]	**0.076*** [0.005,0.147]	-0.008 [-0.034,0.018]	0.016 [-0.023,0.056]
Membership in an Organization (ref. None)	At Least 1	0.005 [-0.013,0.024]	**0.032**$^{+}$ [-0.000,0.064]	0.014 [-0.005,0.033]	0.016 [-0.006,0.038]

Continued on Next Page

Y: Undeclared Work for a ...		Model 1: DQ Private Person	Model 1: RRT Private Person	Model 2: DQ Company	Model 2: RRT Company
Norm Hypothesis					
Undeclared Work in Network (ref. Nobody)	1 to 3%	0.026 [-0.008,0.059]	**0.077**$^{+}$ [-0.007,0.161]	0.003 [-0.024,0.030]	0.025 [-0.013,0.064]
	4 to 10%	0.022 [-0.012,0.055]	**0.091*** [0.019,0.163]	0.004 [-0.023,0.031]	0.022 [-0.021,0.064]
	≥ 11%	**0.060***** [0.026,0.093]	**0.079*** [0.008,0.151]	**0.022**$^{+}$ [-0.004,0.048]	**0.060***** [0.025,0.094]
Approval of Undeclared Work (ref. Disapproval 7 to 14)	Some Disappr. (15 to 16)	0.023 [-0.014,0.060]	-0.012 [-0.084,0.059]	0.005 [-0.019,0.030]	**0.029**$^{+}$ [-0.005,0.064]
	Some Appr. (17 to 19)	0.029 [-0.007,0.064]	0.031 [-0.022,0.083]	-0.007 [-0.033,0.019]	0.015 [-0.028,0.057]
	Approval (20 to 28)	**0.033**$^{+}$ [-0.001,0.068]	**0.066**** [0.021,0.110]	0.001 [-0.024,0.026]	**0.033**$^{+}$ [-0.003,0.068]
Controls					
Gender (ref. Male)	Female	-0.014 [-0.035,0.006]	**0.027**$^{+}$ [-0.003,0.056]	-0.005 [-0.024,0.014]	**-0.020**$^{+}$ [-0.043,0.003]
Age (ref. 35 to 49)	Age ≤ 34	**0.031*** [0.004,0.059]	-0.004 [-0.033,0.025]	**0.032**$^{+}$ [-0.002,0.065]	**0.034**** [0.008,0.059]
	Age ≥ 50	**0.034*** [0.004,0.065]	-0.004 [-0.046,0.037]	0.005 [-0.034,0.045]	-0.321 [-38.021,37.379]
Formal Training (ref. Vocational Training)	Pupil/No Degree	-0.004 [-0.030,0.022]	0.019 [-0.012,0.050]	-0.006 [-0.028,0.015]	-0.012 [-0.040,0.017]
	Tertiary Degree	**-0.051*** [-0.099,-0.002]	-0.389 [-23.462,22.685]	omitted	-0.023 [-0.071,0.025]
Migr. Background (ref. None)	Yes	-0.014 [-0.036,0.008]	-0.004 [-0.034,0.027]	-0.001 [-0.021,0.019]	0.001 [-0.025,0.026]
Residence (ref. West Germany)	East Germany	-0.006 [-0.029,0.017]	-0.011 [-0.047,0.025]	0.007 [-0.014,0.028]	-0.012 [-0.040,0.015]
Model Fit					
N		1,142	2,062	810	2,062
LR Chi² (df)		77.626 (29)	101.912 (30)	31.315 (29)	88.206 (30)
AIC		250.679	1484.605	145.597	1462.477
BIC		396.855	1653.548	267.720	1631.42

95% confidence intervals in brackets; $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B Appendix to Chapter 3

B.1 Regression Estimates for IST

Let L_i be an indicator for the long-list sample (i.e., L_i equals 1 if observation i belongs to the long-list sample and else is 0). In the long-list sample, the respondents are asked about the sum of a sensitive item S (e.g., hours of undeclared work) and a non-sensitive item C (e.g., hours of watching TV). In the short-list sample, the respondents are only asked about the non-sensitive item C. The answers of the respondents can therefore be written as

$$Y_i = \begin{cases} S_i + C_i & \text{if } L_i = 1 \text{ (long-list sample)} \\ C_i & \text{else (short-list sample)} \end{cases} \tag{B.1}$$

Furthermore, let $X_i = (X_{1i}, ..., X_{ki}, 1)'$ and $Z_i = (Z_{1i}, ..., Z_{mi}, 1)'$ be two vectors of covariates (each including a constant; typically, $Z = X$) and assume that S and C can be modeled as

$$S_i = X_i'\beta + \upsilon_i, E(\upsilon_i) = 0 \text{ and } C_i = Z_j'\gamma + \nu_i, E(\nu_i) = 0 \tag{B.2}$$

where β and γ are vectors of coefficients and υ and ν are random errors. It follows that

$$\begin{aligned} Y_i &= L_i S_i + C_i \\ &= L_i(X_i'\beta + \upsilon_i) + (Z_i'\gamma + \nu_i) \\ &= L_i X_i'\beta + Z_i'\gamma + \varepsilon_i \end{aligned} \tag{B.3}$$

with $\varepsilon_i = L_i\upsilon_i + \nu_i$ and, hence, $E(\varepsilon_i) = 0$. For example, if $X = Z = 1$, that is, if there are no covariates, then an estimate for the mean of S can be gained by regressing Y on L using the least squares method (assuming heteroscedastic errors to account for the fact that the error variance depends on L). The slope coefficient of the fitted model, $\hat{\beta}$, is an estimate of $E(S)$, the intercept, $\hat{\gamma}$, is an estimate of $E(C)$. More generally, least-squares regression of Y on $L * X$ and Z (again assuming heteroscedastic errors) provides an estimate of effects of covariates X on S; the coefficients we are interested in are the ones attached to the interaction term $L * X$. Alternatively, assuming bivariate normality of υ and ν (with variances σ_υ^2 and σ_ν^2 and correlation ρ), coefficients can also be estimated by the maximum-likelihood method, based on the log-likelihood

$$ln\mathcal{L} = \sum_{i=1}^{n} \left\{ L_i ln \left[\frac{1}{\sigma_\varepsilon} \Phi \left(\frac{Y_i - X_i'\beta - Z_i'\gamma}{\sigma_\varepsilon} \right) \right] + (1 - L_i) ln \left[\frac{1}{\sigma_v} \Phi \left(\frac{Y_i - Z_i'\gamma}{\sigma_v} \right) \right] \right\} \quad \text{(B.4)}$$

where $\Phi(.)$ is the density function of the standard normal distribution and n is the sample size. Formally, $\sigma_\varepsilon^2 = \sigma_\upsilon^2 + \sigma_v^2 + 2\rho\sigma_\upsilon\sigma_v$, but σ_v *and* ρ cannot be identified separately in this model so that σ_ε is estimated directly. The maximum-likelihood approach will yield identical point estimates for β as the least-squares method, although standard errors may differ. The results in this chapter were computed using the least-squares procedure.

The methods presented can easily be extended to include a third sample of respondents for which the sensitive item S was measured via direct questioning. Let D_i be an indicator for the direct-questioning sample (i.e., D_i equals 1 if observation i belongs to the direct questioning sample and else is 0). Because S is sensitive, S^* is measured in the direct-questioning sample, which is equal to S plus social-desirability bias. In the direct-questioning sample, let $Y_i = S^* = X_i'\beta^* + \upsilon_i$, where β^* is a coefficient vector that includes social-desirability bias, then we can write the regression model across the three subsamples as

$$Y_i = L_i X_i'\beta + (1 - D_i) Z_i'\gamma + D_i X_i'\beta^* + \varepsilon_i \quad \text{(B.5)}$$

or, equivalently, as

$$Y_i = L_i X_i'(\beta - \beta^*) + (1 - D_i) Z_i'\gamma + (D_i + L_i) X_i'\beta^* + \varepsilon_i \quad \text{(B.6)}$$

In the latter form, the coefficients attached to the interaction term L^*X provide an estimate of the (negative of the) bias in β^*. For example, if regressing Y on L, $(1 - D)$, and $(D + L)$ (i.e., if no covariates are taken into account), then the coefficient attached to L provides an estimate of the effect of the item sum technique (i.e., the degree to which the IST leads to a higher average value of the sensitive variable than direct questioning), the coefficient attached to $(1 - D)$ reflects the mean of the non-sensitive item C, and the coefficient attached to $(D + L)$ is an estimate of the mean of the sensitive variable based on direct questioning.

In our study, some of the respondents initially assigned to the IST mode opted for direct questioning. That is, there is noncompliance with the treatment assignment. To gain an intention-to-treat estimate (ITT) of the effect of the IST, we employ a two-step procedure, where in the first step we fit model $C_i = Z_i'\gamma + v_i$ using the observations from the (realized) short-list group. In the second step, we residualize the (realized) long-list observations using the least-squares estimate $\hat{\gamma}$ from the first step and then fit model

$$\tilde{Y}_i' = L_i X_i'(\beta - \beta^*) + X_i'\beta^* + \varepsilon_i \text{ with } \tilde{Y}_i \begin{cases} Y_i - Z_i\hat{\gamma} \text{ if } L_i = 1 \\ Y_i \qquad \text{else} \end{cases} \tag{B.7}$$

based on the respondents who were initially assigned to the long-list sample or to direct questioning, where $\tilde{L}$ is an indicator for initial assignment to the long-list sample. The least-squares estimate of $(\beta - \beta^*)$ is a consistent estimate of the intention-to-treat effect of the IST,[1] but note that standard errors are biased because the additional uncertainty introduced by the variance of $\hat{\gamma}$ is not taken into account. We therefore apply the non-parametric bootstrap procedure across the two steps to compute the standard errors (Davison and Hinkley, 1997).

The ITT is a conservative estimate of the causal treatment effect. We can improve on the ITT by fitting

$$\tilde{Y}_i' = L_i X_i'(\beta - \beta^*) + X_i'\beta^* + \varepsilon_i \tag{B.8}$$

in the second step above (i.e., using L_i instead of $\tilde{L}_i$ in the equation), while instrumenting L_i with $\tilde{L}_i$ based on a two-stage least squares procedure. Since $\tilde{L}_i$ is randomized, it is a valid instrument. Such an instrumental variables (IV) estimate of $(\beta - \beta^*)$ provides a consistent estimate of the local average treatment effect (LATE) of the IST.

1 A necessary assumption for the ITT estimate to be consistent is that the respondents opting out of the short-list sample (who are completely discarded in our ITT estimation procedure) are not systematically different from the respondents opting out of the long-list group. This assumption appears plausible in our case because respondents did not know to which of the two IST groups they had been assigned to when deciding to opt for direct questioning (we also tested the assumption by comparing the outcomes for the two subgroups; no significant differences were found).

B.2 IST Results Displayed as Regression Output

Table B.1: Estimates for hours of undeclared work per week and monthly earnings from undeclared work by questioning mode and sample

	Hours of Undeclared Work			Earnings from Undeclared Work		
	Naive Estimation	ITT Estimation	IV Estimation	Naive Estimation	ITT Estimation	IV Estimation
S model						
IST	0.78	0.70	0.78	**112.03****	**99.66****	**111.90****
	(0.70)	(0.63)	(0.70)	(40.08)	(35.36)	(41.32)
IST x Benefit	-1.09	-1.02	-1.18	-32.02	-37.17	-34.03
	(1.27)	(1.09)	(1.25)	(48.57)	(41.04)	(51.21)
Benefit Recipient	0.07	0.12	0.12	1.55	2.49	2.49
	(0.07)	(0.09)	(0.08)	(1.36)	(1.72)	(1.69)
Constant	**0.07***	**0.07**+**	**0.07**+**	**1.82****	**1.93***	**1.93***
	(0.03)	(0.04)	(0.04)	(0.70)	(0.84)	(0.75)
C model						
Benefit Recipient	**3.83*****			**-176.11*****		
	(0.83)			(22.33)		
Constant	**11.46*****			**688.76*****		
	(0.44)			(19.43)		
N	3,199	3,211	3,211	3,130	3,211	3,211
N_SL		954	954		912	912
N_Total		3,072	3,072		3,003	3,003

Robust standard errors in parentheses (ITT/IV: bootstrap standard errors); ITT: Intention-to-treat estimates; IV: Instrumental variables estimates

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

C Appendix to Chapter 4

C.1 Validation Studies

Table C.1: Overview of validation studies (adapted from Lensvelt-Mulders et al. (2005a); Wolter (2012))

Study not exhaustive	Item(s)	Response Error	RRT-Method	Mode (RRT)	N (DQ/RRT)
Locander/ Sudman/ Bradburn 1976	Voter registration+ Library card+ Bankruptcy involvement Vote primary+ Drunken driving	DQ > RRT DQ < RRT **DQ > RRT** DQ < RRT DQ > RRT	Unrelated question technique	f2f	92, 61* 93, 61* 38, 26* 80, 50* 30, 23*
Lamb/Stem 1978	Failing course grades	DQ > RRT	Unrelated question technique	f2f	63, 121
Tracy/Fox 1981	Arrests per person = 1 > 1	**DQ > RRT** DQ < RRT DQ > RRT	Liu and Chow 1976	f2f	120, 410 80, 326 40, 84
van der Heijden et al. 2000	Social security fraud	**DQ > (FRT = Kuk)**	Forced response technique, Kuk	f2f	99, 96, 105
Wolter 2012	Criminal conviction	DQ > RRT	Forced response technique	f2f	208, 309

+ randomsample, * *effective N; weighted.* Bold face typing indicates signicant difference in response error $p < 0.1$. Similar to Wolter (2012) the studies by Folsom (1974) and Kulka et al. (1981) could not be retrieved and are thus not included.

C.2 Social Assistance and Entitlements in Germany

Unemployment Benefits are a benefit of the unemployment insurance that is granted when a person becomes unemployed, has completed the qualification period and has registered as unemployed. This benefit is a remuneration substitute payment. When the eligibility for unemployment benefits has expired or the payment does not suffice to secure a livelihood, persons can apply for Unemployment Benefit II (UB II) according to the second book of the Social Code.

Since the so called Hartz Reforms in the German social assistance system in 2005, people are entitled to Unemployment Benefit II (UB II) if they are between 15 and 64 years of age, capable of working, and if the household they live in—or more precisely, benefit community[1] — does not have sufficient income to secure a livelihood. According to legislation, the 'standard' requirements to be able to secure a living and the amounts to be paid to each recipient cover, for example, the costs of food, clothing and energy in order to satisfy the needs of everyday life. In this welfare benefit scheme at the household level, it is usually the head of the household who applies for the entire benefit community. In order to establish benefit entitlement and receive benefits from the FEA, the household income has to be declared along with a series of other relevant information on composition of the household. This information is needed to establish eligibility and the amount, which is then paid to each benefit community. In certain situations, e.g., in a low income situation, it is possible that a benefit community might not be eligible for the full amount but that only a share is covered, such as assistance for housing. These transfer payments are then sometimes remitted to the landlord directly by the FEA, without ever being transferred to the applicant/the benefit community.

1 While a household includes all its members, the definition of a benefit community differs: it "consists of a least one person capable of work eligible for benefits, his/her partner, and the unmarried children under 25 years living in the household" (FEA – Federal Employment Agency, 2012). For example a three generation household could by definition never be a single benefit community but two separate ones.

Bibliography

AAPOR – The American Association for Public Opinion Research (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (7th Edition ed.). Lanexa: AAPOR.

Ahart, A. M. and P. R. Sackett (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods 7*, 101–114.

Aldashev, A. and B. Fitzenberger (2009). Der Zugang von Arbeitnehmern in den Bezug von Arbeitslosengeld II. *ZEW Discussion Paper 09-063*, 1–42.

Allingham, M. G. and A. Sandmo (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics 1*, 323–338.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods and Research 28*, 186–208.

Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Journal of Economic Literature 36*, 818–860.

Anglewicz, P., D. Gourvenec, I. Halldorsdottir, C. O'Kane, O. Koketso, M. Gorgens, and T. Kasper (2013). The effect of interview method on self-reported sexual behavior and perceptions of community norms in Botswana. *AIDS and Behavior 17* (2), 674–687.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*, 444–455.

Aquilino, W. S., D. L. Wright, and A. J. Supple (2000). Response effects due to bystander presence in CASI and paper-and-pencil surveys of drug use and alcohol use. *Substance Use and Misuse 35(6-8)*, 845–867.

Barton, A. H. (1958). Asking the embarrassing question. *Public Opinion Quarterly 22*, 67–68.

Bartus, T. (2005). Estimation of marginal effects using margeff. *The Stata Journal 5(3)*, 309–329.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 81(401)*, 28–36.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*, 169–217.

Benjamini, Y. and S. Maital (1985). Optimal tax evasion and optimal tax evasion policy: Behavioral aspects. In W. Gaertner and A. Wenig (Eds.), *The Economics of the Shadow Economy*, pp. 245–264. New York: Springer-Verlag.

Best, H. and C. Wolf (2012). Modellvergleich und Ergebnisinterpretation in Logit- und Probit-Regressionen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie 64*, 377–395.

Biemer, P. P. (2010). Overview of design issues: Total survey error. In P. P. Biemer, P. V. Marsden, and J. D. Wright (Eds.), *Handbook of Survey Research*, pp. 27–57. Bingley: Emerald Publishing Group Limited.

Biemer, P. P., B. K. Jordan, M. L. Hubbard, and D. Wright (2005). A test of the item count methodology for estimating cocaine use prevalence. In J. Kennet and J. Gfroerer (Eds.), *Evaluating and Improving Methods Used in the National Survey on Drug Use and Health*, pp. 149–174. Rockville, MD: Substance Abuse and Mental Health Service Administration, Office of Applied Studies.

Biemer, P. P. and L. E. Lyberg (2003). *Introduction to survey quality*. New York: Hoboken: Wiley & Sons.

Blair, G. and K. Imai (2012). Statistical analysis of list experiments. *Political Analysis 20,* 47–77.

Böckenholt, U. and P. G. M. van der Heijden (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika 72(2)*, 245–262.

Boeije, H. and G. J. L. M. Lensvelt-Mulders (2002). Honest by chance: A qualitative interview study to clarify respondents' (non-)compliance with computer-assisted-randomized response. *Bulletin de Methodologie Sociologique 75*, 24–39.

Boockmann, B., R. Döhrn, M. Groeneck, and H. Verbeek (2010). Abschätzung des Ausmaßes der Schwarzarbeit. Technical report, Tübingen/Essen: Institut für Angewandte Wirtschaftsforschung e.V.

Bordignon, M. (1993). A fairness approach to income tax evasion. *Journal of Public Economics 52*, 345–62.

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist 6*, 308–311.

Bradburn, N., S. Sudman, and B. Wansink (2004). *Asking Questions. Revised Edition*. San Francisco: Jossey-Bass.

Breusch, T. (2005). Estimating the underground economy, using mimic models. Technical report, Working Paper, National University of Australia, Canberra, Australia.

Buehn, A. (2012). The shadow economy in German regions: An empirical assessment. *German Economic Review 13(3)*, 275–290.

Buehn, A., A. Karmann, and F. Schneider (2009). Shadow economy and do-it-yourself activities: The German case. *Journal of Institutional and Theoretical Economics 165(4)*, 701–722.

Bullock, H. E. (2006). Attributions for poverty: A comparison of middle-class and welfare recipient attitudes. *Journal of Applied Social Psychology 29(10)*, 2059–2082.

Carroll, R. J., H. Küchenhoff, F. Lombard, and L. A. Stefanskie (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association 91(433)*, 242–250.

Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika 72(2)*, 263–268.

Clark, S. J. and R. A. Desharnais (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods 3(2)*, 160–168.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Volume 2. Edition. Hillsdale, NJ: Erlbaum.

Corstange, D. (2009). Sensitive questions, truthful answers? Modeling the list experiment with listit. *Political Analysis 17(1)*, 45–63.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 41–49.

Coutts, E. and B. Jann (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods and Research 40(1)*, 169–193.

Coutts, E., B. Jann, I. Krumpal, and A.-F. Näher (2011). Plagiarism in student papers: Prevalence estimates using special techniques for sensitive questions. *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik) 231(5+6)*, 749–760.

Cowell, F. A. (1990). *Cheating the Government: the Economics of Evasion*. London: MIT Press Books.

Cross, P., G. Edwards-Jones, H. Omed, and A. P. Williams (2010). Use of a randomized response technique to obtain sensitive information on animal disease prevalence. *Preventive Veterinary Medicine 96*(3-4), 252–262.

Cullis, J. G. and A. Lewis (1997). Why people pay taxes. From a conventional economic model to a model of social convention. *Journal of Economic Psychology 18*, 305–321.

Dalton, D. R., J. C. Wimbush, and C. M. Daily (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology 47*, 817–828.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

De Leeuw, E. D., J. J. Hox, and M. Huisman (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics 19*, 153–176.

De Schrijver, A. (2012). Sample survey on sensitive topics: Investigating respondents' understanding and trust in alternative versions of the randomized response technique. *Journal of Research Practice 8(1)*, M1.

Dell'Anno, R. (2009). Tax evasion, tax morale and policy maker's effectiveness. *The Journal of Socio-Economics 38(6)*, 988–997.

Diekmann, A. (2012). Making use of "Benford's Law" for the randomized response technique. *Sociological Methods and Research 41*, 325–334.

Dietz, P., H. Striegel, A. G. Franke, K. Lieb, P. Simon, and R. Ulrich (2013). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy 33*(1), 44–50.

Droitcour, J., R. A. Caspar, M. L. Hubbard, T. L. Parsley, W. Visscher, and T. M. Ezzati (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*, pp. 185–210. New York: Wiley & Sons.

EC – European Commission (2007). Special Eurobarometer 284/Wave 67.3: Undeclared work in the European Union. Report, European Commission, Brussels. Accessed: 17.06.2012.

Eichhorn, B. H. and L. S. Hayre (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference 7*, 307–316.

Eifler, S. (2009). *Kriminalität im Alltag: Eine handlungstheoretische Analyse von Gelegenheiten*. VS Verlag.

Enste, D. H. (2012). Schwarzarbeit und Schattenwirtschaft. *Wirtschaftsdienst 92(2)*, 136–138.

Enste, D. H. and F. Schneider (2006a). Schattenwirtschaft und irreguläre Beschäftigung: Irrtümer, Zusammenhänge und Lösungen. In J. Alt and M. Bommes (Eds.), *Illegalität. Grenzen und Möglichkeiten der Migrationspolitik*, pp. 35–59. Springer VS.

Enste, D. H. and F. Schneider (2006b). Welchen Umfang haben Schattenwirtschaft und Schwarzarbeit? Ein Versuch zur Lösung des Rätsels. *Wirtschaftsdienst 86(3)*, 85–198.

Enste, D. H. and F. Schneider (2008). Much ado about not(h)ing? – Eine Replik zu Koch (2008) und Graf (2008). *List Forum für Wirtschafts- und Finanzpolitik 34(2)*, 112–122.

Erikson, R., J. H. Goldthorpe, and L. Portocarero (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *British Journal of Sociology 30*, 341–415.

Evers, H. D. (1987). Subsistenzproduktion. Markt und Staat. Der sog. Bielefelder Verflechtungsansatz. *Geographische Rundschau 39*, 136–140.

Falk, A. (2003). Homo oeconomicus versus Homo reciprocans: Ansätze für ein neues Wirtschaftspolitisches Leitbild? *Perspektiven der Wirtschaftspolitik 4(1)*, 141–172.

Fay, R. E. I. and R. A. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association 74(366)*, 269–277.

FEA – Federal Employment Agency (2012). Homepage. accessed 12.12.12.

Feige, E. L. (1979). How big is the irregular economy. *Challenge 22*, 5–13.

Feige, E. L. (1990). Defining and estimating underground and informal economies: The new institutional economics approach. *World Development 18(7)*, 989–1002.

Feld, L., A. Schmidt, and F. Schneider (2007). Tax evasion, black activities and deterrence in Germany: An institutional and empirical perspective. *Discussion Paper, presented at the Public Choice Meeting in Amsterdam, 29. March to 1. April 2007*.

Feld, L. P. and C. Larsen (2005). *Black Activities in Germany in 2001 and in 2004. A comparison based on survey data*. Copenhagen: The Rockwool Foundation Research Unit.

Feld, L. P. and C. Larsen (2008). *"Black" Activities Low in Germany in 2006*. Copenhagen: The Rockwool Foundation Research Unit.

Feld, L. P. and F. Schneider (2010). Survey on the shadow economy and undeclared earnings in OECD countries. *German Economic Review 11(2)*, 109–149.

Fidler, D. S. and R. E. Kleinknecht (1977). Randomized response versus direct questioning: Two data-collection methods for sensitive information. *Psychological Bulletin 84(5)*, 1045–1049.

Fisher, I. (1922). *The Purchasing Power of Money. Its Determination and Relation to Credit, Interest and Crises* (2 ed.). New York: The Macmillian Co. Assisted by Harry G. Brown.

Folsom, R. E. (1974). A randomized response validation study: Comparison of direct and randomized reporting of DUI arrests (final report 2550-807). Technical report, Chapel Hill: Research Triangle Institute.

Fox, J. A. and P. E. Tracy (1986). *Randomized Response: A Method for Sensitive Surveys*. Berverly Hills: Sage Publications.

Franke, A. G., C. Bagusat, P. Dietz, I. Hoffmann, P. Simon, R. Ulrich, and K. Lieb (2013). Use of illicit and prescription drugs for cognitive or mood enhancement among surgeons. *BMC Medicine 11*, 102.

Frey, B. S. and H. Weck-Hannemann (1984). The hidden economy as an "unobserved" variable. *European Economic Review 26(1)*, 33–53.

Ganzeboom, H. B. G., P. M. De Graaf, and D. J. Treiman (1992). A standard international socio-economic index of occupational status. *Social Science Research 21*, 1–56.

Ghosh, M. and J. N. K. Rao (1994). Small area estimation: An appraisal. *Statistical Science 9(1)*, 55–93.

Gjestvang, C. R. and S. Singh (2007). Forced quantitative randomized response model: A new device. *Metrika 66*, 243–257.

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly 77*, 159–172.

Gordon, J. (1989). Individual morality and reputation costs as deterrents to tax evasion. *European Economic Review 33(4)*, 797–805.

Granovetter, M. (1995 [1974]). *Getting A job. A Study of Contacts and Careers* (2nd edition ed.). Chicago: Chicago Press.

Greenberg, B. G., A. L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association 64(326)*, 520–539.

Greenberg, B. G., R. R. Kuebler Jr., J. R. Abernathy, and D. G. G. Horvitz (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association 66(334)*, 243–250.

Groves, R. M. (1991). Measurement error across the disciplines. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*, pp. 1–25. New York: Wiley & Sons.

Groves, R. M. (2004[1989]). *Survey Error and Survey Costs*. Hoboken: Wiley & Sons.

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

Groves, R. M., F. J. Fowler, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology*. Hoboken: Wiley & Sons.

Gutmann, P. M. (1977). The subterranean economy. *Financial Analysts Journal 34*, 24–27.

Häder, S. and S. Gabler (1998). Ein neues Stichprobendesign für telefonische Umfragen in Deutschland. In S. Gabler, S. Häder, and J. Hoffmeyer-Zlotnik (Eds.), *Telefonstichproben in Deutschland*, pp. 69–88. Opladen: Westdeutscher Verlag.

Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *The Journal of Economic Perspectives 15(4)*, 57–67.

Hendrickx, J. (2002). ISKO: Stata module to recode 4 digit ISCO-88 occupational codes, statistical software components s425802. Boston College Department of Economics. revised 20 Oct 2004.

Hessing, D. J., H. Elffers, H. S. J. Robben, and P. Webley (1993). Needy or greedy? The social psychology of individuals who fraudulently claim unemployment benefits. *Journal of Applied Social Psychology 23(3)*, 226–243.

Himmelfarb, S. and S. E. Edgell (1980). Additive constants model: A randomized response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin 87*, 525–530.

Holbrook, A. L. and J. A. Krosnick (2010a). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly 74(2)*, 328–343.

Holbrook, A. L. and J. A. Krosnick (2010b). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly 74(1)*, 37–67.

Hollis, S. and F. Campbell (1999). What is meant by intention to treat analysis? Survey of published randomizsed concontrol trials. *British Medical Journal 319*, 670–674.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin 30(2)*, 161–172.

Horvitz, D. G., B. V. Shah, and W. R. Simmons (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section. American Statistical Association*, 65–72.

IAW – Institut für Angewandte Wirtschaftsforschung e.V. (2013). Schattenwirtschaftsprognose 2013: Relativ günstige Wirtschaftsentwicklung und Entlastungen bei der Rentenversicherung führen zu weniger Schattenwirtschaft. Pressemitteilung. Accessed: 23.3.2013.

Jacobebbinghaus, P. and S. Seth (2007). The German Integrated Employment Biographies Sample IEBS. *Schmollers Jahrbuch 127*, 335–342.

Janisch, U. and D. Brümmerhoff (2004). Möglichkeiten und Grenzen der Schätzung der Schattenwirtschaft – Eine kritische Auseinandersetzung mit den Schätzergebnissen der Bargeldmethode nach Schneider. Thuenen-Series of Applied Economic Theory. Working Paper Nr. 43. Universität Rostock, 21-43. Rostock 43-2, Working Paper, University of Rostock, Institute of Economics, Germany.

Jann, B. (2011). Rrlogit: Stata module to estimate logistic regression for randomized response data. Statistical Software Components, Boston College Department of Economics.

Jessen, J., W. Siebel, C. Siebel-Rebell, U. J. Walther, and I. Weyrather (1988). *Arbeit nach der Arbeit. Schattenwirtschaft. Wertewandel und Industriearbeit*. Opladen: Westdeutscher Verlag.

Karmann, A. J. (1990). Schattenwirtschaft und ihre Ursachen: Eine empirische Analyse zur Schwarzwirtschaft und Selbstversorgung in der Bundesrepublik Deutschland. *Zeitschrift für Wirtschafts- und Sozialwissenschaften (ZWS) 110(3)*, 185–20.

Kaufmann, D. and A. Kaliberda (1996). Integrating the unofficial economy into the dynamics of post-socialist economies: A framework of analysis and evidence. *Policy Research Working Paper Series 1691*, 1–44.

Kirchgässner, G. (1983). Size and development of the West German shadow economy, 1955-1980. *Zeitschrift für die gesamte Staatswissenschaft 139(2)*, 197–214.

Kirchner, A. (2013). Validating sensitive questions in labor market surveys: A comparison of survey and register data. Paper presented at the 68th American Association for Public Opinion Research Annual Conference, May 18th.

Kirchner, A., M. Trappmann, I. Krumpal, and B. Jann (2012). Item sum: A new technique for asking quantitative sensitive questions. Paper presented at the 67th American Association for Public Opinion Research Annual Conference, May 5th.

Kirchner, A., M. Trappmann, I. Krumpal, and H. v. Hermanni (2011). Eliciting illicit work: Item count and randomized response technique put to the test. Paper presented at the Joint Statistical Meeting, Aug. 3rd.

Kirchner, A., M. Trappmann, I. Krumpal, and H. v. Hermanni (2013). Messung und Erklärung von Schwarzarbeit in Deutschland – Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit. *Zeitschrift für Soziologie 42(4)*, 291–314.

Kish, L. (1965). *Survey Sampling*. New York: Wiley & Sons.

Knäuper, B. (1998). Filter questions and question interpretation: Presuppositions at work. *Public Opinion Quarterly 62*, 70–78.

Koch, W. A. S. (2005). Das Schwarzarbeit-Änigma. *Wirtschaftsdienst 85(11)*, 715–723.

Koch, W. A. S. (2008). Sisyphusarbeiten – Untersuchungen zu Schattenwirtschaft und Schwarzarbeit. *List Forum für Wirtschafts- und Finanzpolitik 34(2)*, 81–101.

Kreuter, F. (2013). Paradata in the total survey error framework. In F. Kreuter (Ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information*, pp. 1–10. New York: Wiley & Sons.

Kreuter, F., G. Müller, and M. Trappmann (2010). Nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly 74(5)*, 880–906.

Kreuter, F., G. Müller, and M. Trappmann (2013). Mechanisms leading to lower data quality of late or reluctant respondents. Unpublished Manuscript.

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.

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research 41(6)*, 1387–1403.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity 47(4)*, 2025–2047.

Krumpal, I. and A.-F. Näher (2012). Entstehungsbedingungen sozial erwünschten Antwortverhaltens: Eine experimentelle Studie zum Einfluss des Wordings und des Kontexts bei unangenehmen Fragen. *Soziale Welt 63(1)*, 65–89.

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika 77(2)*, 436–438.

Kuklinski, J. H., P. M. Sniderman, K. Knight, T. Piazza, P. E. Tetlock, G. R. Lawrence, and B. Mellers (1997). Racial prejudice and attitudes toward affirmative action. *American Journal of Political Science 41*, 402–419.

Kulka, R. A., M. F. Weeks, and R. E. Folsom (1981). A comparison of the randomized approach and direct questioning approach to asking sensitive survey questions (working paper). Technical report, Research Triangle Institute.

LaBrie, J. W. and M. Earleywine (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. *Journal of Sex Research 37*, 321–326.

Lackó, M. (1998). The hidden economies of Visegrad countries in international comparison: A household electricity approach. In L. Halpern and C. Wyplosz (Eds.), *Hungary: Towards a Market Economy*, pp. 128–152. Cambridge: Cambridge University Press.

Lackó, M. (1999). Do power consumption data tell the story – electricity intensity and hidden economy in post-socialist coutries. *Budapest Working Papers on the Labour Market 1999(2)*, 1–31.

Lago-Peñas, I. and S. Lago-Peñas (2010). The determinants of tax morale in comparative perspective: Evidence from European countries. *European Journal of Political Economy 26*, 441–453.

Lamb, C. W. and D. E. Stem (1978). An empirical validation of the randomized response technique. *Journal of Marketing Research 15*, 616–621.

Lamnek, S., G. Olbrich, and W. J. Schäfer (2000). *Tatort Sozialstaat. Schwarzarbeit, Leistungsmissbrauch, Steuerhinterziehung und ihre (Hinter-)Gründe*. Opladen: Leske+Budrich.

Landsheer, J. A., P. G. M. van der Heijden, and G. van Gils (1999). Trust and understanding. two psychological aspects of randomized response. A study of a method for improving the estimate of social security fraud. *Quality & Quantity 33*, 1–12.

Langfeldt, E. (1984). The unobserved economy in the Federal Republic of Germany. In E. L. Feige (Ed.), *The Unobserved Economy*, pp. 236–260. Cambridge: Cambridge University Press.

Lara, D., S. G. García, C. Ellertson, C. Camlin, and J. Suárez (2006). The measure of induced abortion levels in Mexico using random response technique. *Sociological Methods and Research 35*, 279–301.

Lara, D., J. Strickler, C. D. Olavarrieta, and C. Ellertson (2004). Measuring induced abortion in Mexico. *Sociological Methods and Research 32(4)*, 529–558.

Lavender, J. M. and D. A. Anderson (2009). Effect of perceived anonymity in assessments of eating disordered behaviors and attitudes. *International Journal of Eating Disorders 42*(6), 546–551.

Lederer, W. and H. Küchenhoff (2006). A short introduction to the SIMEX and MCSIMEX. *R News 6(4)*, 26–31.

Lee, R. M. (1993). *Doing Research on Sensitive Topics*. London: Sage.

Lelkes, Y., J. A. Krosnick, D. M. Marx, C. M. Judd, and B. Park (2012). Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology 48*(6), 1291–1299.

Lensvelt-Mulders, G. J. L. M., J. J. Hox, P. G. M. van der Heijden, and C. J. M. Maas (2005a). Meta-analyss of randomized response research: Thirty-five years of validation. *Sociological Methods and Research 33(3)*, 319–348.

Lensvelt-Mulders, G. J. L. M., J. J. Hox, and P. G. M. Van der Hejden (2005b). How to improve the efficiency of randomised response designs. *Quality & Quantity 39*, 253–265.

Lensvelt-Mulders, G. J. L. M., P. G. M. Van der Heijden, O. Laudy, and G. van Gils (2006). A validation of computer-assisted randomized response survey to estimate the prevalence of undeclared work in social security. *Journal of the Royal Statistical Society (Series A) 169(2)*, 305–318.

Little, R. J. and S. Vartivarian (2005). Does weighting for nonresponse increase the variance of the survey means. *Survey Methodology 31(2)*, 161–168.

Locander, W., S. Sudman, and N. Bradburn (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association 71*, 269–275.

Long, S. J. (1997). *Regression Models for Categorical and Limited Dependent Variables*. London: Sage.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

Mander, A. and D. Clayton (1999). Hotdeck: Stata module to impute missing values using the hotdeck method. Technical report, Statistical Software Components S366901, Boston College Department of Economics. Revised 02 Sep 2007.

Mangat, N. S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society (Series B) 56*, 93–95.

Mangat, N. S. and R. Singh (1990). An alternative randomized response procedure. *Biometrika 77*, 439–442.

Manzoni, A., J. K. Vermunt, R. Luijkx, and R. Muffels (2010). Memory bias in retrospectively collected employment careers: A model-based approach to correct for measurement error. *Sociological Methodology 40(1)*, 39–73.

Mehlkop, G. (2011). *Kriminalität als rationale Wahlhandlung: Eine Erweiterung des Modells der subjektiven Werterwartung und dessen empirische Überprüfung*. VS Verlag.

Mehlkop, G. and R. Becker (2004). Soziale Schichtung und Delinquenz. Eine empirische Anwendung einse Rational-Choice-Ansatzes mit Hilfe von Querschnittsdaten des ALLBUS 1990 und 2000. *Kölner Zeitschrift für Soziologie und Sozialpsychologie 56*, 95–126.

Merz, J. and K. G. Wolff (1993). The shadow economy: Illicit work and household production: A microanalysis of West Germany. *Review of Income and Wealth 39(2)*, 177–194.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review 26(1)*, 67–82.

Moors, J. J. A. (1971). Optimization of the unrelated randomized response model. *Journal of the American Statistical Association 66*, 627–629.

Moshagen, M., J. Musch, M. Ostapczuk, and Z. Zhao (2010). Reducing socially desirable responses in epidemiologic surveys an extension of the randomized-response technique. *Epidemiology 21*(3), 379–382.

Mummert, A. and F. Schneider (2001). The German shadow economy: Parted in a united Germany? *FinanzArchiv 58*, 286–316.

Myles, G. D. and R. A. Naylor (1996). A model of tax evasion with group conformity and social customs. *European Journal of Political Economy 12*, 49–66.

Newell, D. J. (1992). Intention-to-treat analysis: Implications for quantitative and qualitative research. *International Journal of Epidemiology 21*, 837–841.

OECD – Organisation for Economic Co-operation and Development (2002). *Measuring the Non-Observed Economy. A Handbook*. Paris: OECD Publications Service.

Oehlert, G. W. (1992). A note on the delta method. *American Statistician 46*, 27–29.

Pedersen, S. (1998). *The Shadow Economy in Western Europe. Measurement and Results for Selected Countries*. Copenhagen: The Rockwool Foundation Research Unit.

Pedersen, S. (2003). *The Shadow Economy in Germany, Great Britain and Scandinavia: A Measurement Based on Questionnaire Surveys*. Copenhagen: The Rockwool Foundation Research Unit.

Peeters, C. F. W., G. J. L. M. Lensvelt-Mulders, and K. Lasthuizen (2010). A note on a simple and practical randomized response framework for eliciting sensitive dichotomous and quantitative information. *Sociological Methods and Research 39*(2), 283–296.

Pfau-Effinger, B. (2009). Varieties of undeclared work in European societies. *British Journal of Industrial Relations 47(1)*, 79–99.

Pickard, M. and J. Sardà (2011). The size of the underground economy in Germany: A correction of the record and new evidence from the modified-cash-deposit-ratio approach. *European Journal of Law and Economics 32*, 143–163.

Raghavarao, D. and W. T. Federer (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 41*, 40–45.

Rasinski, K. A., G. B. Willis, A. K. Baldwin, W. Yeh, and L. Lee (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology 13*, 465–484.

Rayburn, N. R., M. Earleywine, and G. C. Davison (2003). Base rates of hate crime victimization among college students. *Journal of Interpersonal Violence 18*, 1209–1221.

Renooy, P. (1990). *The Informal Economy: Meaning, Measurement and Social Significance. Netherlands Geographical Studies No. 115*. Amsterdam: National Geographical Studies Association.

Renooy, P., S. Ivarsson, O. van der Wusten-Gritsai, and R. Meijer (2004). *Undeclared work in an enlarged Union: An analysis of shadow work. An in-depth study of specific items*. Brussels: TNS Opinion & Social.

Sakshaug, J. W. and F. Kreuter (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods 6*, 113–122.

Scheers, N. J. and C. M. Dayton (1988). Covariate randomized response models. *Journal of the American Statistical Association 83(404)*, 969–974.

Schneider, F. (2003). The shadow economy. In C. K. Rowley and F. Schneider (Eds.), *Encyclopedia of Public Choice*, pp. 286–296. Dordrecht: Kluwer Academic Publishers.

Schneider, F. (2007). Shadow economies and corruption all over the world: New estimates for 145 countries. *Economics 9*, 1–47.

Schneider, F. (2008). The shadow economy in Germany: A blessing or a curse for the official economy? *Economic Analysis & Policy 38(1)*, 89–111.

Schneider, F. (2009). Size and development of the shadow economy in Germany, Austria and other OECD countries. Some preliminary findings. *Revue économique 60(5)*, 1079–1116.

Schneider, F. and D. H. Enste (2000). Shadow economies: Size, causes and consequences. *Journal of Economic Literature 38*, 77–114.

Schneider, F. and D. H. Enste (2007). *The Shadow Economy. An International Survey*. New York: Cambridge University Press.

Schneider, F., J. Volkert, and S. Caspar (2002). *Schattenwirtschaft und Schwarzarbeit: Beliebt bei Vielen – Problem für Alle*. Baden-Baden: Nomos.

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

Schwarz, N., H. J. Hippler, B. Deutsch, and F. Strack (1985). Response categories: Effects on behavioral reports and comparative judgments. *Public Opinion Quarterly 49*, 388–395.

SchwarzArbG (2004). Gesetz zur Bekämpfung der Schwarzarbeit und illegalen Beschäftigung (SchwarzArbG) (Version 01.08.2004).

Singer, E., D. R. Von Thurn, and E. R. Miller (1995). Confidentiality assurances and response: A quantitative review of the experimental literature. *Public Opinion Quarterly 59(1)*, 66–77.

Smith, L. L., W. T. Federer, and D. Raghavarao (1975). A comparison of three techniques for eliciting truthful answers to sensitive questions. *Proceedings of the Social Statistics Section. American Statistical Association*, 447–452.

Spicer, M. and S. Lundstedt (1976). Understanding tax evasion. *Public Finance 31(2)*, 295–305.

Tanzi, V. (1983). The underground economy in the United States: Annual estimates,1930-80. *International Monetary Fund Staff Papers 30(2)*, 283–305.

Tanzi, V. and P. Shome (1993). A primer on tax evasion. *International Monetary Fund, Staff Papers 40(4)*, 807–828.

Thießen, U. (2011). Schattenwirtschaft: Vorsicht vor hohen Makroschätzungen. *Wirtschaftsdienst 3*, 194–201.

Thomas, J. J. (1999). Quantifying the black economy: "Measurement without theory" yet again? *The Economic Journal 109(456)*, F381–F389.

Tourangeau, R. (1984). Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines*. Washington: National Academy Press.

Tourangeau, R. and K. A. Rasinski (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin 103(3)*, 299–314.

Tourangeau, R., L. J. Rips, and Rasinski (2000). *The Psychology of Survey Response*. New York: 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*, 275–304.

Tourangeau, R. and T. Yan (2007). Sensitive questions in surveys. *Psychological Bulletin 133*, 859–883.

Tracy, P. E. and J. A. Fox (1981). The validity of randomized response for sensitive measurements. *American Sociological Review 46(2)*, 187–200.

Trappmann, M., S. Gundert, C. Wenzig, and D. Gebhardt (2010). PASS: A household panel survey for research on unemployment and poverty. *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 130(4)*, 609–622.

Trappmann, M., I. Krumpal, A. Kirchner, and B. Jann (2014). Item sum: A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology 2(1)*, 58–77.

Tsuchiya, T. and Y. Hirai (2010). Elaborate item count questioning: Why do people underreport in item count responses? *Survey Research Methods 4(3)*, 139–149.

Tsuchiya, T., Y. Hirai, and S. Ono (2007). A study of the properties of the item count technique. *Public Opinion Quarterly 71(2)*, 253–272.

Umesh, U. N. and R. A. Peterson (1991). A critical evaluation of the randomized response method: Applications, validation, and research agenda. *Sociological Methods and Research 20*, 104–138.

Van den Hout, A. and P. G. M. van der Heijden (2002). Randomized response, statistical disclosure control and misclassification: a review. *International Statistical Review 70(2)*, 269–288.

Van der Heijden, P. G. M., G. van Gils, J. Bouts, and J. J. Hox (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods and Research 28(4)*, 505–537.

Voss, T. and M. Abraham (2000). Rational choice theory in sociology: A survey. In S. R. Quah and A. Sales (Eds.), *The International Handbook of Sociology*, pp. 50–82. London: Sage Publications.

Warner, S. L. (1965). Randomized-response: A survey technique for eliminating evasive answer-bias. *Journal of the American Statistical Association 60*, 63–69.

Weissman, A. N., R. A. Steer, and D. S. Lipton (1986). Estimating illicit drug use through telephone interviews and the randomized response technique. *Drug and Alcohol Dependence 18(3)*, 225–233.

Weiß, C. (2008). *Auf der Suche nach Schwarzarbeit: explorative Verfahren zur Erfassung devianten Verhaltens am Arbeitsmarkt*. Baden-Baden: Nomos Verlag.

Wenzel, M. (2004). The social side of sanctions: Personal and social norms as moderators of deterrence. *Law and Human Behavior 28(5)*, 547–567.

Williams, C. C. (2009). Formal and informal employment in Europe: Beyond dualistic representations. *European Urban and Regional Studies 16*, 147–159.

Williams, C. C. (2010). Retheorizing participation in the underground economy. *Labor Studies Journal 35*, 246–267.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal 12*, 308–331.

Wimbush, J. C. and D. R. Dalton (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology 82*, 756–763.

Wolff, K. G. (1991). *Schwarzarbeit in der Bundesrepublik Deutschland. Eine mikroanalytische Untersuchung*. Frankfurt: Campus.

Wolter, F. (2012). *Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik*. Springer VS.

Yitzhaki, S. (1974). A note on income tax evasion: A theoretical analysis. *Journal of Public Economics 3*, 201–202.

Yu, J.-W., G. L. Tian, and M. L. Tang (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika 67*, 251–263.

Zoll (2011). Ahndung von Ordnungswidrigkeiten und Verfolgung von Straftaten.

Zoll (2012). Personal Correspondence 02.02.2012. Informations- und Wissensmanagement Zoll. Zentrale Auskunft.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis 13*, 157–170.

Abstract

This dissertation explores methods to improve the quality of data about sensitive labor market topics, such as undeclared work and receipt of basic income support in Germany, using surveys of the general population. Due to the sensitive nature of both topics, respondents may choose to misreport and adjust their answers in accordance with social norms.

Over the past decades, special strategies—particularly targeted to reduce misreporting on sensitive topics—have been developed. One such class of data collection strategies are so-called 'dejeopardizing' techniques, out of which the randomized response technique (RRT) and the item count technique (ICT) are the most popular and best investigated ones. The goal is to elicit more honest answers from respondents by increasing the anonymity of the question-and-answer process. These techniques provide prevalence estimates as well as estimates of regression coefficients, regressing dependent variables generated by means of RRT or ICT on a set of covariates of interest.

However, these dejeopardizing techniques have not been applied to collect data on undeclared work or receipt of welfare benefits in German surveys. This dissertation aims at closing this gap using an experimental design that allows us to compare the performance of these dejeopardizing techniques to direct questioning. In 2010 we conducted two telephone surveys on undeclared work and welfare benefit receipt. We experimentally tested whether the RRT, the ICT, or the newly developed item sum technique (IST) reduce bias due to social desirability compared to direct questioning (under the 'more-is-better' assumption and using validation data in one study).

Our results suggest that neither the RRT nor the ICT provide unambiguous results with respect to improving the accuracy of reports of the socially undesirable behavior, while the IST results were more promising. This dissertation provides insights into a variety of practical and theoretical factors contributing to a successful implementation of the RRT, the ICT and the IST in labor market surveys.

Zusammenfassung

Die vorliegende Dissertation geht der Frage nach, wie das Ausmaß von Schwarzarbeit und Arbeitslosengeld-II-Bezug in Deutschland im Rahmen von Befragungen der allgemeinen Bevölkerung möglichst valide geschätzt werden kann. Aufgrund des heiklen Charakters beider Themen ist davon auszugehen, dass Selbstauskünfte häufig nicht der Wahrheit entsprechen und stattdessen in vielen Fällen sozial erwünschte Antworten gegeben werden und das Verhalten systematisch unterberichtet wird.

Um diesen Antwortverzerrungen entgegen zu wirken, wurden in den letzten Jahrzehnten in der empirischen Sozialforschung alternative Befragungstechniken entwickelt. So basieren beispielsweise die Randomized Response Technique (RRT) und die Item Count Technik (ICT) auf dem Prinzip der verschlüsselten Antworten und sollen durch eine Erhöhung der Anonymität in der Interviewsituation sozial erwünschtes Antwortverhalten reduzieren. Der Vorteil dieser Erhebungsverfahren liegt darin, dass zum einen weniger Annahmen hinsichtlich der Schätzungen getroffen werden und zum anderen mittels statistischer Auswertungen zielgerichtet multivariate Zusammenhänge zwischen einer mit ICT oder RRT generierten abhängigen Variablen und Kovariaten auf individueller Ebene untersucht werden können.

Bislang wurden diese Techniken allerdings noch nicht zur Erhebung von Schwarzarbeit oder des Bezugs von Arbeitslosengeld-II in Deutschland eingesetzt. Die Dissertation schließt diese Lücke und beschäftigt sich mit einem experimentellen Vergleich – sowie einer Weiterentwicklung – von Erhebungstechniken speziell für heikle Fragen mit einer direkten Befragung im Kontext von Arbeitsmarktsurveys. Mittels Fragen zum Thema Schwarzarbeit und zum Arbeitslosengeld-II-Bezug, wird im Rahmen zweier Bevölkerungsbefragungen aus dem Jahre 2010 empirisch untersucht ob die RRT, die ICT bzw. die eigens entwickelte Item Sum Technik (IST) den Befragten tatsächlich ein höheres Ausmaß sozial unerwünschter Antworten entlocken als die direkte Befragung (unter der bekannten 'more-is-better' Annahme sowie mittels einer Validierungsstudie).

Die Befunde zeigen, dass die häufig angenommene Wirkung der RRT oder der ICT auf die Bereitschaft der Befragten, sozial unerwünschtes Verhalten zu berichten, nicht eindeutig ausfällt. Die Ergebnisse der IST fallen hingegen positiver aus. Die vorliegende Dissertation liefert somit Hinweise hinsichtlich verschiedener praktischer als auch theoretischer Faktoren, die zu einer erfolgreichen Implementation der RRT, der ICT und der IST in Arbeitsmarktsurveys beitragen können.

Grundsicherung

Ergebnisse aus der SGB-II-Forschung des IAB

Acht Jahre nach der Einführung der Grundsicherung für Arbeitsuchende im Jahr 2005 zieht das IAB erneut Bilanz. Der Bericht fasst die Ergebnisse aus der SGB-II-Forschung des IAB in den Jahren 2009 bis 2012 zusammen und stellt die Befunde in einen größeren Zusammenhang. Der Stand des Wissens zur Struktur und Dynamik im Leistungsbezug wird ebenso dargelegt wie die Erkenntnisse zum Prozess der Aktivierung und der Betreuung. Zudem präsentiert der Band Forschungsbefunde zu den Wirkungen der arbeitsmarktpolitischen Instrumente sowie zu den gesamtwirtschaftlichen Effekten der Reformen.

Die Autoren zeigen auf, wo die Grundsicherung heute steht und wo – aus Sicht der Forschung und der Praxis – die künftigen Herausforderungen liegen.

Martin Dietz, Peter Kupka, Philipp Ramos Lobato
Acht Jahre Grundsicherung für Arbeitsuchende
Strukturen - Prozesse - Wirkungen
IAB-Bibliothek, 347

2013, 379 S., 42,90 € (D)
ISBN 978-3-7639-4081-3
Auch als E-Book

W. Bertelsmann Verlag 0521 91101-0 wbv.de

Standard surveying techniques are usually not suited to collect valid information on the prevalence of undeclared work or receipt of basic income support. Respondents often misreport their behavior and adjust their answer in accordance with the social norm.

In the social sciences alternative strategies have been developed, particularly targeted to increase respondent anonymity in the interview situation and thus reduce misreporting on sensitive topics.

Antje Kirchner investigates whether these special techniques lead to higher reports of undeclared work and receipt of basic income support. Furthermore, this work presents the Item Sum Technique, a novel questioning technique that shows more promising results compared to direct questioning.

W. Bertelsmann Verlag

ISBN 978-3-7639-4083-7

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