UNIVERSITÄT MANNHEIM

Detecting Interviewer Falsification in Surveys – A Guide for Improving Quality Controls

Inauguraldissertation
zur Erlangung des akademischen Grades
einer Doktorin der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von

Silvia Schwanhäuser, M.Sc.

Mannheim

Dekan der Fakultät für Sozialwissenschaften:	Prof. Dr. Michael Diehl
Erstbetreuer:	Prof. Joseph W. Sakshaug, PhD
Zweitbetreuer:	Prof. Dr. Mark Trappmann
Erstgutachter:	Prof. Dr. Florian Keusch
Zweitgutachter:	Prof. Dr. Mark Trappmann
Tag der Disputation:	14. Oktober 2024

List of Contributions

I. How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification

II. How Falsifiers Make a Long Story Short: Identifying Partial Interviewer Falsification in Panel Surveys

III. Leaving No Data Unturned: Evaluating Machine Learning Algorithms to Detect Interviewer Falsification

Acknowledgments

I would like to express my heartfelt gratitude to all the people whose support made this dissertation possible. I am deeply grateful that they accompanied me on my journey. This is especially true for all my co-authors, to whom I owe thanks for the consistently good, productive, and collegial collaboration. I would also like to explicitly thank the GradAB program for their support. I am sincerely grateful for the feedback I received during the scholarship from both senior researchers and GradAB fellows.

A very special thanks goes to my supervisor Joe Sakshaug. He took so much time for our meetings, gave me so many helpful tips and advice, constantly encouraged me, and supported me in every way possible. I thank him from the bottom of my heart. I would also like to sincerely thank my second supervisor, Mark Trappmann, who has supported me since my university days. He sparked my passion for survey methodology, provided me with my first insights into the world of research, and has always supported my career.

I am firmly convinced that women need other strong women as role models. Therefore, I want to thank three very special women whom I greatly admire. All three have supported me and stood by me with advice. First, I would like to mention Frauke Kreuter, who convinced me to pursue a PhD in the first place and brought me into her network from the very beginning. Second, I want to thank Yuliya Kosyakova, who never tired of encouraging me, showing me new carrier paths, and always believing in me more than I believed in myself. And lastly, my thanks go to Dana Müller, who, as my mentor, has always been there for me, helping me step out of my comfort zone and see the value in my own work.

I was also very fortunate to be part of a Survey Research PhD group at IAB. The time spent together was always a great enrichment for me. Whether it was traveling together, giving practice talks, proofreading, or just having a coffee in between—they all became dear friends to me. For this, I want to thank Sophie Hensgen, Sebastian Hülle, Corinna König, Benjamin Küfner, Jan Mackeben, and Lukas Olbrich very much. I would also like to sincerely thank my other colleagues at the IAB for all their support. A big thank you also goes to Nick Linsel, who supported me during his internship in my final phase of this dissertation.

Most importantly, I am especially grateful for the support I received from my friends and family. They always had a listening ear for me, offered words of encouragement, and were understanding even when I had no time for them. I am particularly grateful to my mom, who has always supported me unconditionally and in every way possible. From the bottom of my heart and with the greatest love, I also want to thank my husband for his endless patience and loving support. He has been my anchor through it all.

To wrap up, I would like to lean on the wise words of Snoop Dogg (2018): *"Last but not least, I wanna thank me. I wanna thank me for believing in me. I wanna thank me for doing all this hard work. […] I wanna thank me for never quitting."*

Contents

List of Figures

List of Tables

1. Introduction

»THE PREVALENCE OF SO-CALLED "CHEATING" BY INTERVIEWERS IN THE PROCESS OF OBTAINING PUBLIC OPINION AND MARKET RESEARCH DATA HAS BECOME AN INCREASINGLY GRAVE CONCERN TO RESPONSIBLE OPINION RESEARCHERS. ANY PRECISE FIGURES UPON THE INCIDENCE OF FABRICATION ARE, IN THE NATURE OF THE CASE, DIFFICULT TO OBTAIN. BUT IT IS NO SECRET THAT THE PREVALENCE AND AMOUNT HAS BEEN IN MANY INSTANCES FAR FROM NEGLIGIBLE, AND IT IS WIDELY AGREED THAT THE PROBLEM MUST BE SOLVED IF THE OPINION RESEARCH TECHNIQUE IS TO PRESERVE ITS STATUS AS A RELIABLE TOOL OF INQUIRY. «

Leo Crespi (1945)

Nearly eight decades ago, Leo Crespi was the first to openly discuss the interviewer "cheating" problem, namely, interviewers purposefully deviating from their interview instructions in surveys (Crespi 1945; Groves 2004). Since then, an increasing number of conference papers, reports, and studies have addressed the topic of interviewer falsification. These contributions have facilitated a more comprehensive understanding of the prevalence of the problem, the underlying causes of such interviewer behavior, and its impact on data quality (e.g., see contributions in Winker, Menold, and Porst 2013). Further studies have delineated strategies for preventing or identifying interviewer falsification (see Bredl, Storfinger, and Menold 2013; Robbins 2018; DeMatteis et al. 2020 for detailed overviews). Despite the fact that many survey organizations and methodologists have identified methods for addressing the issue of interviewer falsification, several of Crespi's assertions remain pertinent in the present day: Obtaining precise figures upon the prevalence of cheating interviewers remains challenging, yet sporadic reports of extremely high rates of falsifications emphasize the extent of the problem (Turner et al. 2002; Bredl, Storfinger, and Menold 2013). Even for smaller amounts of fabricated data, studies found that these cases hold the potential of severely biasing analysis results (Schnell 1991; Schräpler and Wagner 2005; Brüderl, Huyer-May, and Schmiedeberg 2013), confirming the statement that the problem is far from neglectable. In addition, reports of data falsification hold the potential to affect the credibility of surveys and related research (Werker 1981; DeMatteis et al. 2020).

However, as surveys—and in particular interviewer-administered surveys—play a crucial role in various fields of study, including sociology, economics, demographics, and public opinion research, the credibility and reliability of survey data is of vital importance (Olson et al. 2020). These data are for instance employed to establish new academic insights, to evaluate policy programs and political measures, and serve as the foundation for policy advice (Thissen and Myers 2016). In some cases, the results of surveys can directly influence

political decisions and the allocation of funding, which further underlines the importance of ensuring high data quality. Despite the outlined possibility of interviewer falsification, interviewer-administered survey modes are often considered superior data collection methods in terms of data quality. Interviewers have in many respects a positive effect on data quality (Japec 2006; Olson et al. 2020): For example, interviewers can assist in creating sampling lists, even in settings where no register information is available (Eckman and Koch 2019). While contacting the respondents, interviewers can counteract refusal and therefore increase response rates (Biemer and Lyberg 2003; West and Blom 2017). In the process of administering the questionnaire, they can facilitate the completion of lengthy or complicated questionnaires, and address respondents' inquiries (Fowler 2013). Further, interviewers can collect additional information, like information on the respondents' housing situation or biological data (Pashazadeh, Cernat, and Sakshaug 2020; West et al. 2020).

At the same time, previous research has also found negative effects on the data quality attributable to the interviewer (see, e.g., Fowler and Mangione 1990; Groves 2005; West and Blom 2017), including interviewer effects, unintended interviewer error, and—interviewer falsification. Importantly and despite the growing body of literature, interviewer falsification is still an understudied topic, especially in contrast to, for example, interviewer effects. Among the studies focusing on interviewer falsification, most contributions present their particular method that was used to identify falsifications in a specific survey (see, e.g., Stokes and Jones 1989; Hood and Bushery 1997; Turner et al. 2002; Murphy et al. 2004; Porras and English 2004; Li et al. 2011; Bredl, Winker, and Kötschau 2012; Bergmann, Schuller, and Malter. 2019; Kosyakova et al. 2019). Although these studies are valuable contributions and have helped scholars to improve quality controls, they have two important limitations: First, because of their nature, one cannot assess the external validity of the results for other survey settings. Second, one cannot determine which of the proposed methods is most effective in detecting interviewer falsification, as the studies lack an evaluation of multiple methods. As Crespi emphasized, the problem of "cheating" interviewers must be solved if researchers want to preserve credibility of survey data and fully exploit the advantages of interviewer-administered surveys (Crespi 1945). Thus, identifying the appropriate quality control strategy to identify interviewer falsification is of crucial importance. This dissertation aims to contribute to this goal by providing guidance on the detection of various interviewer falsification forms in face-to-face surveys when using statistical identification methods. This includes the selection of the appropriate methods to enhance quality controls and, thus, preserving the credibility of survey data.

1.1 Previous Findings on Interviewer Falsification

1.1.1 What Is Considered to Be Interviewer Falsification?

According to the American Association for Public Opinion Research (AAPOR) every intentional deviation from the designed instructions or guidelines, i.e., any interviewer misbehavior, which goes unreported by the interviewer, is considered interviewer falsification, if it holds the potential to impact the data or its quality (Groves 2004). This definition of interview falsification encompasses a variety of different types of fraudulent interviewer behavior (Blasius and Friedrichs 2012; Robbins 2018; Schwanhäuser et al. 2020). These different types of behavior vary in some respects; 1) the stage in the data collection process in which they occur, 2) the interviewers' motivation for this behavior, 3) their frequency, 4) their impact on the data quality, and 5) potential prevention and identification strategies necessary to counteract the behavior. **Figure 1.1** provides an overview of possible interviewer misbehavior along the data collection process, following the distinction given by DeMatteis et al. (2020).

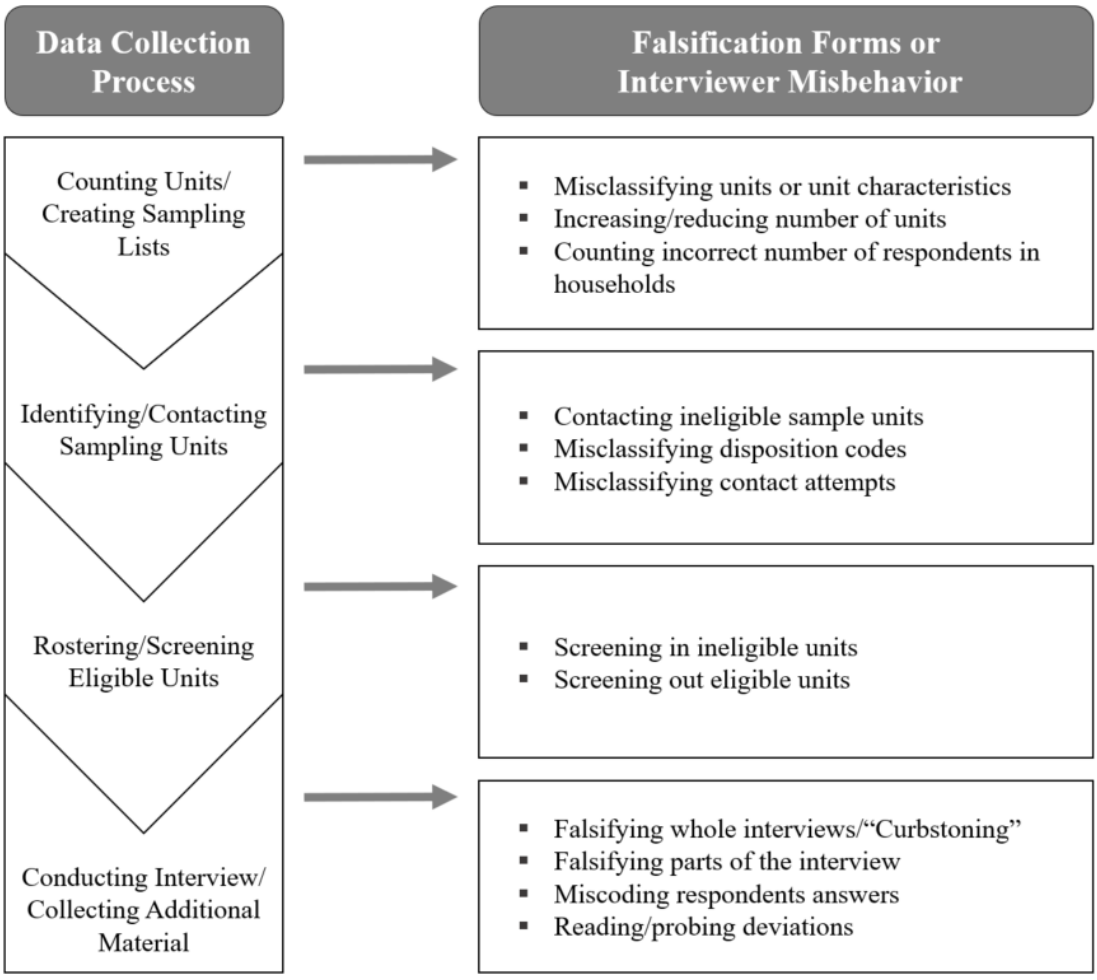

Figure 1.1: Overview of possible interviewer misbehavior along the data collection process. *Source*: Own illustration.

In principle, interviewers can deviate from their instructions in any step of the data collection process in which they are involved. Dependent on the interview mode and sampling method (Menold 2014), this may include the task of creating sample lists, the identification or screening of sampling units, and the interview itself (DeMatteis et al. 2020). As outlined in **Figure 1.1**, the type of potential interviewer misbehavior changes accordingly. In some surveys, interviewers are involved in the creation of the sample by listing or counting sampling units. This includes listing households or other units in specific geographical regions or counting units with particular characteristics (Biemer and Lyberg 2003; Groves et al. 2011; Menold 2014; West and Blom 2017). These procedures follow defined rules, guaranteeing the probability-based nature of the sampling process (Fowler 2013). Interviewers might deviate by counting an incorrect number of units, misclassifying certain units to include or exclude them in the sample, or miscounting persons within household units (DeMatteis et al. 2020). Such deviations can be motivated by a desire to avoid certain buildings, neighborhoods, or respondents with specific characteristics, as well as by the inaccessibility of an interview locations (Gwartney 2013; Robbins 2018; Davis and Wilfahrt 2024). However, these deviations may result in a biased sample.

In addition, interviewers frequently perform the task of identifying and contacting sampling units or screening for eligible units (Biemer and Lyberg 2003; Menold 2014; West and Blom 2017). In this process, interviewers may be inclined to interview ineligible but willing units (Eckman and Koch 2019), manipulate disposition codes or the contact attempt history (DeMatteis et al. 2020), or deviate from the screening procedure (Turner et al. 2002; Murphy et al. 2016). Interviewers may utilize such deviations to circumvent uncooperative or complicated respondents (Eckman and Koch 2019) and to enhance or achieve required response rates. Such deviations can bias contact or response rates, which is particularly problematic if these data are used to adjust for nonresponse (Wagner, Olson, and Edgar 2017). They may also influence the sample composition (Menold 2014).

Probably the most common understanding of interviewer falsification is the fabrication of survey data instead of conducting the interview as instructed. Most studies focus on this stage in the survey process (DeMatteis et al. 2020). The vast majority of these studies shed light on complete falsifications, i.e., the fabrication of the entire interview. Here, interviewers neither contact the target respondents nor obtain any information by these respondents; instead, they fill questionnaires with fictitious data (DeMatteis et al. 2020; Schwanhäuser et al. 2020). Some authors refer to this behavior as “curbstoning” (Stokes and Jones 1989; Schäfer et al. 2004b; Li et al. 2011). This form of interviewer falsification is

often regarded as the most blatant form (Bredl, Storfinger, and Menold 2013). Because those fabricated interviews do not contain any valid information from target respondents, predictions or statements regarding the target population can be biased (Schreiner et al. 1988; Biemer and Stokes 1989; Koch 1995).

As a further variation, falsifiers may fabricate single questions or questionnaire modules, which is also referred to as partial falsification. In this case, interviewers are actually conducting an interview, but are only asking parts of the questionnaire during a so-called "short interview" (Blasius and Thiessen 2012; DeMatteis et al. 2020). The remaining questions are answered by the interviewer to complete the interview (Storfinger and Winker 2013), resulting in a mixture of real survey data and fabricated responses. Similar to complete falsifications, statements on the target population can be severely biased (Biemer and Stokes 1989; Koch 1995; Schäfer et al. 2004b). In addition to this fraudulent behavior, falsifying interviewers may deliberately miscode respondents' proper answers to manipulate filter questions and thereby reduce the length of the interview (Brüderl, Huyer-May, and Schmiedeberg 2013; Kosyakova, Skopek, Eckman 2015, Josten and Trappmann 2016).

Further information: The process of fabricating data

When fabricating interviews, falsifiers can follow different strategies. The simplest, but probably easiest to detect, strategy is to randomly pick answers (Hülser 2013). As such a behavior produces implausible combinations and answers, or produces suspicious paradata (e.g., too short time stamps), they are easily detected (Menold et al. 2013). Some more sophisticated falsifiers try to copy "typical" respondent behavior, imitating "stereotypical" respondents while fabricating answers (Blasius and Friedrichs 2012; Menold et al. 2013). However, this behavior in turn can result in specific patterns in the data, which allow the identification of these cases. Other falsifiers might get support by accomplices like family, friends, or neighbors (DeMatteis et al. 2020), which makes the identification of these cases especially challenging. One specific strategy to produce complete falsifications is the duplication of records (Koczela et al., 2015; Sarracino and Mikucka, 2016; Winker, 2016), i.e., copies which share identical responses (Slomczynski, Powalko, and Krauze 2017). Here, falsifiers may simply copy all or most answers given by a real respondent (Sarracino and Mikucka, 2016). As these copies do not lead to suspicious answer patterns in the data, they are hard to identify by most data-based detection methods (Koczela et al., 2015; Kuriakose and Robbins 2016). Related to this, falsifiers could use the data of any person willing to take the survey, resulting in real data coming from the wrong respondent (DeMatteis et al. 2020).

It is important to note that many of the mentioned types of interviewer misbehavior are not binary outcomes, but rather follow along a continuous spectrum (Murphy et al. 2004; Murphy et al. 2016). An interviewer might fabricate every question (complete falsification) in every interview. Another interviewer might manipulate answers to filter questions or fabricate some items (partial falsification) in a very small number of interviews. This spectrum of deviations is illustrated in **Figure 1.2**, which shows that interviewer behavior can move along these lines: The number of falsified items and the number of interviews being affected by fraud.

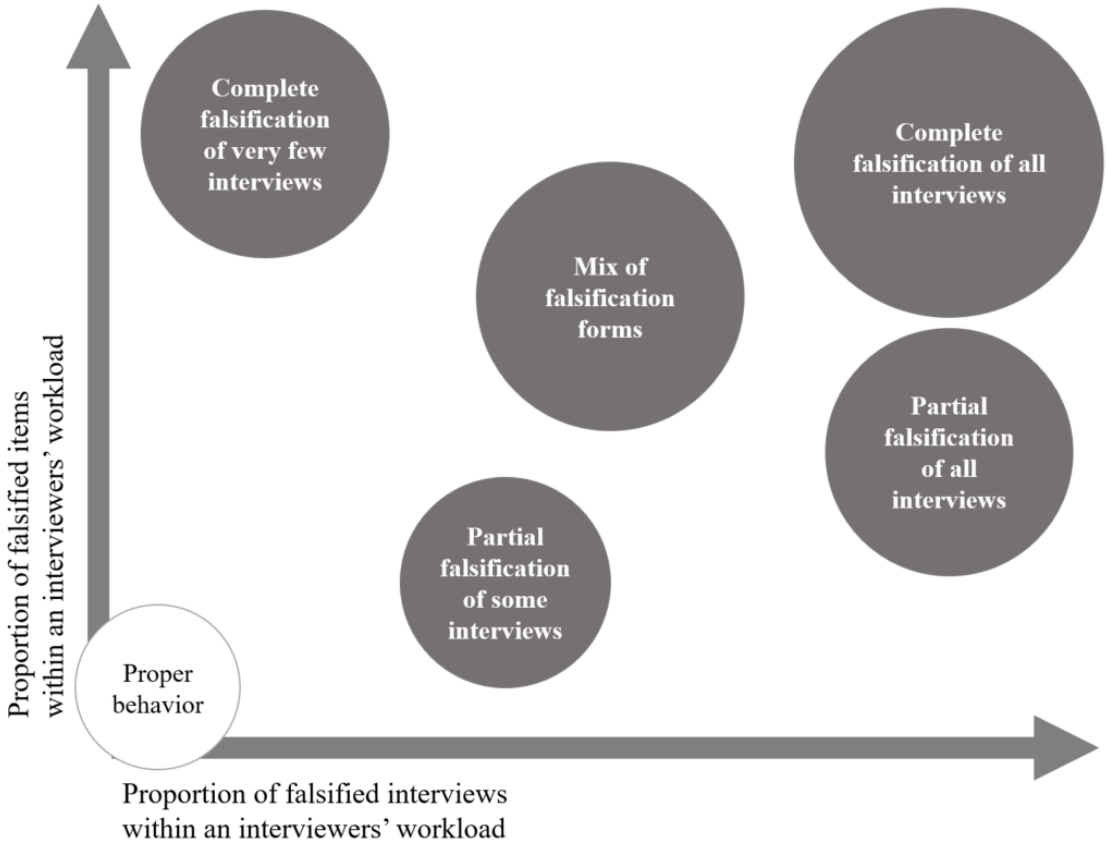

Figure 1.2: Illustration of possible falsification forms along a continuous spectrum.
Source: Own illustration.
Note: The size of the circles represents the total quantity of fabricated information attributable to a fabricating interviewer.

An interviewer decides from interview to interview whether to falsify or not and even if the interview is properly started, they might decide from question to question whether they rely on standardized interview procedures or not. An interviewer might fake only one interview (e.g., to meet a tough deadline), occasionally fabricate interviews (e.g., if the population includes hard-to-reach respondents), or decide to fabricate all interviews (e.g., to get the highest possible monetary outcome with the least amount of work). An interviewer may select the right respondents, but avoids one single item battery (e.g., because it is

difficult for respondents to answer), or never follow the reading guidelines (e.g., using a more conversational interview style). Interviewers might even rely on multiple falsification strategies, e.g., counting vacant buildings as eligible and then fabricate interviews for these "fictional" households (see e.g., Kindred and Scott 1993).

Further information: Differences between survey modes

Further differences may arise between the different interviewer-administered modes. Complete and partial falsifications as well as other fraudulent interviewer behavior can occur in any interviewer-administered survey mode, such as telephone or face-to-face interviews (Blasius and Friedrichs 2012). However, due to differences between the modes, for example with regard to the monitoring processes or the interviewers' payment, fraudulent interviewer behavior is much less likely in telephone surveys. Hence, most studies refer to interviewer falsification in face-to-face surveys. Nevertheless, there are exceptions, such as Nelson and Kiecker (1996), who documented fraudulent behavior by telephone interviewers.

1.1.2 Random Error or Intended Interviewer Behavior?

In addition to the fluid transitions between falsification forms, interviewer falsification may also get conflated within another survey methodological concept; namely the Total Survey Error (TSE) framework (see Groves et al. 2011 for further information). In practice, the boundaries between the concepts of interviewer misbehavior and random interviewer errors may be blurred in some situations. For example, interviewers could either misclassify an answer because of an accidental typing error or because they wanted to avoid follow-up questions triggered by a filter question. Previous research has highlighted numerous such potential errors attributable to the interviewer (Fowler and Mangione 1990; Biemer and Lyberg 2003; Groves et al. 2011), and often summarized them within the context of the TSE. Thus, some authors also define interviewer falsification as a part of the measurement error concept within the TSE framework (see, e.g., Biemer and Lyberg 2003; Winker 2016), as the measurement error describes the differences between the underlying true value and the obtained response in the survey (Groves et al. 2011). But unlike most error sources within the TSE, interviewer falsification is less of an accidental "error" than a deliberate decision by the interviewer (Gwartney 2013; Robbins 2018). Further, the measurement error considered in the TSE framework usually impacts every observation and is assumed to be randomly generated from a zero-mean distribution (DeMatteis et al. 2020). Interviewer misbehavior and error also clearly differ with regard to the interviewer's motivation and intent. Hence, unintentional errors may affect the variance of a measure,

whereas interviewer falsification may introduce a systematic error leading to biased measures (Groves 2004; Gwartney 2013; DeMatteis et al. 2020). This has even led to criticism regarding the TSE framework, as it does not sufficiently include the incident of interviewer falsification, even though it makes the claim that the framework considers all factors that reduce data quality and validity (Spagat 2016). Although it is not always possible to distinguish between unintentional error and intentional fraudulent behavior (Robbins 2018), as the result could be the same, both types of behavior vary regarding the interviewers' motivation for showing this behavior (Gwartney 2013). Therefore, it is important to ascertain reasons promoting interviewer falsification, as identifying and addressing these reasons may help to prevent fraudulent behavior in the first place.

1.1.3 Why Do Interviewers Falsify?

There are various factors that can lead interviewers to become falsifiers or show specific types of misbehavior. The vast majority of interviewers are not inherently fraudulent, and those who engage in falsifying data are often a minority among otherwise trustworthy interviewers (Schwanhäuser et al. 2020). In many cases, the motivation to falsify data is simply the result of certain circumstances or situations that daunt interviewers from fulfilling their roles adequately (Blasius and Friedrichs 2012; Gwartney 2013; Menold et al. 2018). For instance, interviewers may be confronted with complex and lengthy questionnaires, hard-to-reach populations, or specific survey locations; hence, factors increasing interviewer burden (Winker 2016; Robbins 2018). The various demoralizing factors can be classified into intrinsic and extrinsic factors of the interviewers' work, which collectively undermine the morale of the interviewers and thus increase the probability of falsifications (Gwartney 2013; Jesske 2013; Koczela et al. 2015).

Intrinsic factors include conditions related to the sampling design, the survey instrument, the respondents, or working conditions (Crespi 1945; Nelson and Kiecker 1996; Gwartney 2013; Schwanhäuser et al. 2020). Hence, they frequently lie within the control of the researcher or the survey organization. Demoralizing factors of the sampling design include difficult selection rules, which increase the probability of deviations from these rules (Gwartney 2013). On the side of the survey instrument, poorly designed questionnaires, programming errors, or lengthy questionnaires can increase the probability of falsifications (Crespi 1945). This may lead interviewers to shorten the interview, skip parts of the questionnaire, or rephrase the wording of questions. The same holds if interviewers have to deal with challenging respondents or have to navigate complicated interview situations (Crespi 1945; Gwartney 2013). It is, however, important to note that fraudulent behavior is

not always due to the interviewer's burden, but can also be caused by potential respondent burden (Robbins 2018). Some interviewers want to ease the respondent's task by avoiding cognitively demanding and lengthy item batteries, particularly in high-frequency panel surveys where interviewers can anticipate the respondent's answers (Schreiner, Newbrough, and Pennie 1988; Koczela et al. 2015). In addition, poor working conditions such as extremely high workloads, poorly communicated standards, poor payment or contract schemes, or pressure from supervisors (Nelson and Kiecker 1996; Bredl, Storfinger, and Menold 2013; Gwartney 2013) may result in falsifications, as interviewers attempt to make their work more profitable.

Extrinsic factors—factors that are related to the external environment—include the interview location, the interviewer's personal situation, but also the interviewer's knowledge about controlling actions and monitoring procedures (Gwartney 2013; Koczela et al. 2015; Robbins 2018; Schwanhäuser et al. 2020). Circumstances, such as working in dangerous or inaccessible areas, favor falsification (Robbins 2018; Davis and Wilfahrt 2024). A lot of these factors can be considered during the process of designing the survey, minimizing interviewer burden, using monitoring procedures and hence reduce the overall falsification likelihood (Crespi 1945; Biemer and Stokes 1989; Blasius and Friedrichs 2012; Gwartney 2013; Koczela et al. 2015; Schwanhäuser et al. 2020).

Further information: Theoretic perspectives on deviant interviewer behavior

From the perspective of behavioral theory, the different intrinsic and extrinsic demoralizing factors are embedded within a broader moral hazard problem that exists between a principal (here the researcher or a survey organization) and the agent (here the interviewer) (Winker et al. 2013; Kosyakova, Skopek, Eckman 2015; Winker 2016, Robbins 2018). On the one hand, researchers or survey organizations are unable to fully observe the actions of interviewers in the field (Kosyakova, Skopek, Eckman 2015). At the same time, interviewers and their employers may pursue conflicting goals: Where researchers and organizations may prioritize high data quality and sufficient numbers of interviews, interviewers may be more concerned with optimizing their time and effort, reducing their burden, and increasing their monetary output (Menold et al. 2013; Kosyakova, Skopek, Eckman 2015; Olbrich et al. 2023). Consequently, it is likely that interviewers will deviate from established guidelines if such deviation is expected to increase their own utility (Gwartney 2013). As a solution, researchers and survey organizations must ensure that the utility gained from adherence to the established guidelines is greater than the utility gained from deviating. This can be

achieved by adjusting three principal parameters within the survey design: First, the survey design can sufficiently address the aforementioned demoralizers in order to minimize the number of such factors (Winker et al. 2013). Second, principals can implement monitoring procedures to enhance the probability of detection, thereby increasing the risk associated with deviating. Finally, principals may adjust the interviewers' payment schemes in order to shift the focus from completing as many interviews as possible to completing high quality interviews (Kosyakova, Skopek, Eckman 2015; Winker 2016). In summary, the principal-agent framework combined with a rational choice theoretical perspective illustrates the interviewers' personal trade-off between complying with instructions or deviating, highlighting the importance of demoralizers, monitoring, and the incentive structure. See Harrison and Krauss (2002) for a systematic overview of the utility process within the interviewer's trade-off.

1.1.4 What Is the Size of the Problem?

Empirical findings suggest that the proportion of falsified data is low, and that interviewer falsification is a rare event (Blasius and Friedrichs 2012; Schwanhäuser et al. 2020). Nevertheless, studies have occasionally reported particularly high falsification rates. A commonality among these surveys is that they only employed a limited number of interviewers and did not limit the maximum number of interviews per interviewer (e.g., Turner et al. 2002; Bredl, Storfinger, and Menold 2013). In order to ascertain the frequency of fraud, **Table 1.1** provides a broad overview of face-to-face survey projects which reported on different forms of interviewer falsification. As **Table 1.1** illustrates, instances of falsification have been documented in a variety of studies conducted over different years and across a range of geographical regions worldwide. Most studies report on complete falsifications, or a combination of different deviant interviewer behavior, whereas very few studies explicitly report on partial falsifications (e.g., Beste, Olbrich, and Schwanhäuser 2021; Bossler et al. 2022). The majority of reports indicate falsification rates between less than 1 percent up to 10 percent (see **Table 1.1**).

Table 1.1: Overview of survey projects reporting incidents of interviewer falsifications.

Survey/Project	Region	Year	Falsification type	Percentage/number of falsifiers/falsifications	Source
Current Population Survey (CPS)	USA, North America	1982-1985	Mixed (complete, partial, other)	~3-5 % of interviewers	Schreiner et al. (1988); Biemer and Stokes (1989)
National Crime Survey (NCS)	USA, North America	1982-1985	Mixed (complete, partial, other)	0.4 % of interviews	Schreiner et al. (1988); Biemer and Stokes (1989)
American Housing Survey (AHS)	USA, North America	1982-1985	Mixed (complete, partial, other)	20 interviews	Schreiner et al. (1988)
Survey of Income and Program Participation (SIPP)	USA, North America	1982-1985	Mixed (complete, partial, other)	15 interviews	Schreiner et al. (1988)
National Health Interview Survey (NHIS)	USA, North America	1983-1996	Complete	0.1-3.6 % of interviews	Hood and Bushery (1997); Bushery et al. (1999)
New York City Housing Vacancy Survey (NYC-HVS)	USA, North America	1984-1987	Mixed (complete, partial, other)	6.5 % of interviews	Schreiner et al. (1988)
German Socio-Economic Panel Study (SOEP)	Germany, Europe	1984-2000	Complete	0.1-2.4 % of interviews	Schräpler and Wagner (2003); Schäfer et al. (2004a, 2004b)
KwaZulu-Natal Income Dynamics Study (KIDS)	South Africa, Africa	1993/1998	Complete	39 interviews	May et al. (2007); Finn and Ranchhod (2017)
General Population Survey of the Social Sciences (ALLBUS)	Germany, Europe	1994	Mixed (complete, partial, other)	2.7 % of interviews	Koch (1995)
Baltimore STD and Behavior Survey (BSBS)	USA, North America	1997-1998	N.A.	6 interviewers	Turner et al. (2002)
Survey of Program Dynamics (SPD)	USA, North America	1998	N.A.	0.8-13 % of interviewers	Bushery et al. (1999)
Integrated Coverage Measurment/Post Enumeration Survey (ICM)	USA, North America	2000	Mixed (complete, partial, other)	0-0.7 % of interviews	Krejsa et al. (1999)
National Survey on Drug Use and Health (NSDUH)	USA, North America	2002	Complete	~ 0.3 % Screening interviews, 0.5 % interviews	Murphy et al. (2004); Murphy et al. (2005)

Table 1.1 (continued)

Current Population Survey (CPS)	USA, North America	2004-2006	Mixed (complete, other)	0.09 % of interviews	Li et al. (2011)
Small household survey	Multiple countries, Eurasia	2007-2008	Complete	Likely all interviews	Bredl et al. (2012)
Cape Area Panel Study (CAPS)	South Africa, Africa	2009	Mixed (complete, partial, other)	9 % of interviews	Lam et al. (2012); Finn and Ranchhod (2017)
National Income Dynamics Study	South Africa, Africa	2010-2011	Mixed (complete, partial)	10 % interviewers, 7,3 % of interviews	Finn and Ranchhod (2017)
Survey on fairness of earnings	Germany, Europe	2010-2011	Complete	5.4 % of interviews	Walzenbach (2021)
Panel Study Labour Market and Social Security (PASS)	Germany, Europe	2012-2020	Mixed (partial, other)	0.17-1.30 % of interviews	Beste et al. (2021)
Field experiment in the Niger Delta region	Nigeria, Africa	2013-2014	Mixed (complete, other)	14 % of interviews	Gomila et al. (2017)
Survey of Health, Ageing and Retirement (SHARE)	Multiple countries, Europe	2016	Complete	~ 9 % of interviews	Bergman et al. (2019)
IAB-BAMF-SOEP Survey of Refugees in Germany	Germany, Europe	2016-2017	Complete	1.3-7.5 % of interviews	Kosyakova et al. (2019); Schwanhäuser et al. (2020)
National Venezuelan Survey	Venezuela, Latin America	2016-2017	Other	~ 30 % of interviews	Castorena et al. (2023)
National Peruvian Survey	Peru, Latin America	2017	N.A.	4.4 % of interviews	Castorena et al. (2023)
Americas Barometer Study	Multiple countries, America	2016-2017	Other	7 % of interviews	Cohen and Warner (2020)
IAB Job Vacancy Survey	Germany, Europe	2020-2021	Partial	16.4-28 % of interviews	Bossler et al. (2022)

Source: Own literature research.
Note: The table only includes well-documented cases of interviewer falsification. Percentages of falsifiers/falsification were used if available from the literature. Ranges of these values were used if the values were distributed over multiple waves or instruments.

Especially in panel surveys, reported falsification rates are often very low. Importantly, partial falsifications are more likely to occur in panel surveys than complete falsifications. This is because the specific structure of panel surveys makes complete falsification rather complicated and easy to detect (Schräpler and Wagner 2005; Blasius and Friedrichs 2012; Schwanhäuser et al. 2020). As many of the listed studies do not provide information on the specific proportion of falsifications for each fabrication type, it is hard to ascertain the prevalence of each type. It is also important to note that, as Crespi (1945) discussed, there is no way of knowing how many instances of falsification go unnoticed. Further, some survey projects fail to systematically document instances of falsification (Winker 2016). The combination of these two factors makes it challenging to accurately estimate the frequency of interviewer falsification.

1.1.5 How Do Falsifications Impact Estimation Results and Data Quality?

Despite the typically low incidence of fraudulent interviews in survey data, research has shown that even small proportions of falsified interviews can result in significant bias (e.g., Schnell 1991; Schräpler and Wagner 2005; Gomila et al. 2017; Sarracino and Mikucka 2017). Given the systematic nature of fraudulent interviewer behavior, falsified interviews can easily result in biased estimates and misleading inferences (Gwartney 2013; Schwanhäuser et al. 2020). Although univariate and bivariate statistics, such as means, proportions, and correlations, are mostly only slightly distorted (Landrock 2017a; Castorena et al. 2023), larger biases are evident for multivariate statistics (Schnell 1991; Schräpler and Wagner 2005; DeMatteis et al. 2020; Schwanhäuser et al. 2020).

In the context of univariate statistics, prior research found that falsifiers are adept at "estimating" marginal distributions present in the target population (Robbins 2018; Castorena et al. 2023). Furthermore, the size of the bias for a univariate estimate cannot exceed the share of falsified records (Schnell 1991; Schräpler and Wagner 2005). Nevertheless, falsifications retain the potential to distort the actual distribution of variables (DeMatteis et al. 2020). Because of these possible distortions in distributions and the falsifiers' inability to reproduce complex multidimensional relationships (Murphy et al. 2005; Schräpler and Wagner 2005; Bredl, Storfinger, and Menold 2013), effects on multivariate statistics are often even more severe. In the case of Schräpler and Wagner (2005), a falsification rate of only 0.6 to 4.7 percent led to misleading regression results, with regression coefficients becoming insignificant or effect sizes increasing/decreasing considerably. This finding is confirmed by Sarracino and Mikucka (2017), who found that only 5 percent of duplicates, i.e., doublets in survey data, had severe effects on estimates and

standard errors. The potential effects of falsifications on various statistical measures can also be derived from theoretical measurement error models. DeMatteis et al. (2020) present a comprehensive overview of formal derivations for different statistical measures and estimation techniques. They demonstrate that even if the introduced bias is equal to zero, falsifications still impact the estimation precision, i.e., inflating variance and standard errors. This impact is comparable to the effect of the intra-interviewer correlation coefficient. Especially for regressions in which both dependent and independent variables include falsified data, the impact on estimates and standard errors becomes unpredictable.

As the specific impact of interviewer falsification on the data depends on a multitude of different factors, it is difficult to make a generally valid statement about this impact (Robbins 2018). Although some results indicate a low impact of falsifications in single surveys (Castorena et al. 2023), the effect may vary from situation to situation. The impact of falsifications might depend on the type of deviant behavior, the falsifiers' ability to approximate population means and distributions, the quantity of falsifications in the survey, and even the interviewer's intention behind the fraudulent behavior (i.e., willingly distorting the data versus other motivations) (Robbins 2018; DeMatteis et al. 2020). To illustrate, consider a situation in which an interviewer uses filter questions to shorten the length of an interview (see, e.g., Kosyakova, Skopek, Eckman 2015; Josten and Trappmann 2016). In such cases, subgroup analysis or analysis of the surveys' key variables may be subject to substantial bias, as certain answers are only available for respondents that triggering the questions. An example of such a situation is the German Job Vacancy Survey, in which two interviewers used a filter question on the number of vacant jobs in the firms to shorten the interviews, leading to significant distortions regarding this key statistic (Bossler et al. 2022). In the light of these considerations, there is a clear rational for implementing sophisticated quality controls in surveys.

1.2 Methods for Detecting Interviewer Falsification

Researchers and survey practitioners have introduced a multitude of different control and detection measures to mitigate the risk interviewer falsification poses to data quality. Early on, these detection methods focused on reducing the information asymmetry regarding the interviewers' behavior, as described in the principal agent framework. For example, this is achieved through re-interviewing procedures to verify information obtained from the initial interview, or through the utilization of observational methods like monitoring (Winker 2016; Robbins 2018; DeMatteis et al. 2020). In particular, monitoring methods have

expanded considerably in recent years due to the advent of technology (see, for example, Thissen and Myers 2016; Edwards, Maitland, and Connor 2017; Wagner, Olson, and Edgar 2017; Edwards, Sun, and Hubbard 2020). In addition to these procedures, a wide range of studies have focused on statistical identification strategies. These methods analyze the available data (e.g., interview data or paradata) to identify suspicious or outlying patterns, which in turn help to distinguish between real interviews and the ones that are falsified (Schwanhäuser et al. 2020). The specifics of the different identification strategies and methods are detailed in the following section.

1.2.1 Non-Statistical Identifications Strategies

1.2.1.1 Re-interviewing and Re-contact

One of the most established methods for identifying falsifications and interview deviations is the re-interview or re-contacting method (Crespi 1945; Groves 2004; Winker 2016; Robbins 2018; DeMatteis et al. 2020; Schwanhäuser et al. 2020). Today, this method is commonly used as standard practice in face-to-face surveys. Re-interview data is collected by re-contacting a subset of the respondents after the initial interview using postal, telephone, or personal re-interviews (Schwanhäuser et al. 2020). This is done to verify whether the interview actually happened (Groves 2004). The method can, for example, assist in verifying whether the interview was conducted with the target respondent under the correct conditions. In some cases, organizations may even attempt to contact nonrespondents or ineligible cases to confirm that interviewers followed the selection rules (Robbins 2018). Usually, re-interview procedures include questions about the composition of the household or further selection criteria, the interview mode, the estimated duration of the interview, whether incentives or computers were used, and some of the key topics or items asked in the questionnaire (Groves 2004; Schwanhäuser et al. 2020). Furthermore, questions pertaining to the respondent's demographics, the use of auxiliary interview aids (e.g., show cards, lists, etc.), and evaluation questions regarding the quality of the interview may prove beneficial (Koch 1995). As re-contacting all respondents would be both burdensome and inefficient in terms of costs, verification is often conducted on a subsample of interviews (Groves 2004; Schwanhäuser et al. 2020). As a purely random selection of interviews would be very inefficient and carries the risk of overlooking falsifications (Storfinger and Opper 2011; DeMatteis et al. 2020), contemporary strategies frequently include “focused re-interviews”, whereby interviews with a higher likelihood of being falsified are oversampled (Winker 2016; Schwanhäuser et al. 2020). This higher likelihood is determined by falsification models or checks for any anomalies or suspicious patterns in the data, which will be

discussed later on (Biemer and Stokes 1989; Hood and Bushery 1997; Bushery et al. 1999; Krejsa, Davis, and Hill 1999; Murphy et al. 2004; Li et al. 2011). Overall, these targeted methods improve the effectiveness of the re-contacting procedures and increase the probability of identifying falsifiers (Bredl, Storfinger, and Menold 2013).

1.2.1.2 Monitoring and Observational Methods

Another commonly used tool for identifying instances of interviewer falsification is through observational methods or monitoring. As monitoring allows to directly observe parts of the interview process, it is an effective way of verifying the behavior of interviewers (Robbins 2018). In addition to its utility as an identification tool, monitoring also serves as a deterrent to interviewers, preventing interviewer misbehavior (Schwanhäuser et al. 2020; Castorena et al. 2023). Telephone surveys conducted in centralized call centers offer optimal conditions for monitoring (Gwartney 2013; Robbins 2018; DeMatteis et al. 2020). Here, supervisors monitor interviewers in real time via so-called "silent monitoring" (Schwanhäuser et al. 2020). This involves listening to the interviewer-respondent interaction and observing data entry (Groves 2004). Interviewers may be selected for monitoring randomly, or based on key statistics like the number of interviews, the interview duration, response and cooperation rates, or the number of contact attempts and refusals (Jesske 2013; Gwartney 2013; Schwanhäuser et al. 2020). As a result, interviewer falsification is easily detected in telephone surveys, which also discourages potential fraudulent behavior in the first place.

In contrast to telephone surveys, face-to-face surveys present a more complex environment for monitoring. Consequently, the possible monitoring strategies employed also vary in comparison to telephone surveys. In some face-to-face surveys, supervisors may accompany less experienced interviewers at the beginning to directly observe their actions (Robbins 2018). More commonly, monitoring is conducted using audio recordings of the interview or parts of it. These recordings can be collected using external devices or built-in microphones in the interviewer's laptop or mobile phone (CARI) (Koczela et al. 2015; DeMatteis et al. 2020; Schwanhäuser et al. 2020). Anomalies within the recordings—such as noises without an interview or audibility of only the interviewer's voice—may indicate fraudulent behavior. Additionally, analysis of the interviewer-respondent-interaction allows providing interviewers with feedback regarding their performance (Thissen and Myers 2016; Edwards, Maitland, and Connor, 2017; Edwards, Sun, and Hubbard 2020; Sun and Yan 2023). However, as capturing such recordings requires the respondents' consent due to legal

and ethical restrictions, interviewers are aware of the fact that they are recorded. Consequently, the recordings may provide only a limited insight into the actual behavior of the interviewer (Fee et al. 2015; Fee, Fields, and Marlay 2016). Still, recordings may serve as a prevention tool.

Additionally, modern techniques have enabled survey practitioners to collect paradata, such as GPS locations or photos of the interview location (Murphy et al. 2016; Thissen and Myers, 2016; Finn and Ranchhod 2017; Wagner, Olson, and Edgar 2017). In the case of the GPS data, the coordinates are either actively collected by the interviewer or passively collected by the software (DeMatteis et al. 2020). The use of these coordinates enables supervisors to ascertain the distance between the target household or unit and the actual GPS location, thereby confirming the interviewers' presence at a specific location (Thissen and Myers, 2016; Winker 2016; Edwards, Maitland, and Connor 2017; DeMatteis et al. 2020). Capturing photos of the environment during the interview also serves to corroborate the interviewers visit in the field (Thissen and Myers 2016; DeMatteis et al. 2020; Schwanhäuser et al. 2020). Such information regarding the location have proven to be of cumbersome importance for detecting interviewer falsification (Cohen and Warner 2020).

1.2.1.3 Validation with External/Administrative Data

Some sampling frames or linked surveys allow the verification of specific survey characteristics through administrative data (Schwanhäuser et al. 2020). Validating basic characteristics like age, gender, names, or addresses, which may be obtained from population registers, offers an effective way of identifying complete falsifications, deviations from the target respondent, or the selection rules (Koch 1995; Schnell 2012). It is important to note that this is only feasible if such external or administrative data are available, and that linking such data requires the respondent's consent in some circumstances. Moreover, the method assumes that information provided by both the registers and the respondents is free of errors, which would otherwise lead to unjustified suspicion regarding some interviewers (Schwanhäuser et al. 2020).

1.2.2 Statistical Identification Strategies

As outlined above, various studies mention a wide range of detection methods which are based on the analysis of the data generated during the survey process. This includes responses collected during the interview, and additional information like paradata. Usually, paradata are automatically generated by computer-assisted survey instruments or automatically collected during the survey process (Kreuter 2013), which limits the falsifiers

potential impact on paradata. Statistical or data-based methods, which make use of these data, are designed to identify outlying, repetitive, illogical, or otherwise suspicious patterns which are indicative of falsification behavior (Schwanhäuser et al. 2020). Importantly, these methods are only used to identify "at risk" interviewers, namely those with a higher likelihood of being falsifiers (Hood and Bushery 1997; Robbins 2018). Hence, these methods need to be combined with non-statistical identifications strategies to confirm fraudulent interviewer behavior (Winker 2016). Studies dealing with statistical methods can be divided into so-called "ex-ante" or "ex-post" studies, meaning that they either showcase how they applied certain methods during the field work (ex-ante), or evaluate the performance of detection methods based on known falsification cases (ex-post) (Bredl, Storfinger, and Menold 2013; Schwanhäuser et al. 2020).

The main data-based strategies may be divided into three categories: 1) falsification indicators, 2) multivariate analyses strategies or modelling approaches, and 3) approaches that focus on the identification of identical response patterns. The term falsification indicator is used to summarize a number of different measurable characteristics in survey data that may indicate fraudulent interviewer behavior. Even though these indicators measure very different concepts, they share the basic assumption that they allow to systematically distinguish between real and falsified data (Schwanhäuser et al. 2020). Even though these indicators measure very different concepts, they share the same basic assumption. Following a rational choice perspective, falsifiers are assumed to show specific response behavior when fabricating data, caused by their endeavor to maximize their (monetary) benefits while minimizing effort (Menold et al. 2013; Kosyakova, Skopek, Eckman 2015; Winker 2016). This frequently results in response behavior that systematically differs from the behavior of real respondents. The literature distinguishes between formal falsification indicators, content-related indicators, and indicators that are based on paradata (Bredl, Storfinger, and Menold 2013; Menold and Kemper 2014; Robbins 2018). Multivariate analysis strategies or modelling approaches are often an extension of this concept, as they combine the different indicators in order to increase the power of the result. Well-known examples include cluster analysis (e.g., de Haas and Winker 2014; Bergmann, Schuller, and Malter 2019; Schwanhäuser, Sakshaug, and Kosyakova 2022), multilevel or regression modeling (e.g., Li et al. 2011; Sharma and Elliott 2020; Olbrich et al. 2023), or machine learning methods (e.g., Birnbaum et al. 2013, Weinauer 2019; Jebreel et al. 2020). Lastly, some methods focus on the identification of identical response patterns, such as duplicates, near-duplicates, or item

batteries with low response variability (e.g., Kuriakose and Robbins 2016; Slomczynski, Powalko, and Krauze 2017; Blasius and Thiessen 2021).

1.2.2.1 Formal Falsification Indicators

Formal falsification indicators are designed to identify response patterns associated with specific types of survey questions (e.g., item scales, filter questions, or open responses) (Schwanhäuser, Sakshaug, and Kosyakova 2022). Consequently, they are closely related to quality indicators measuring response behavior, such as straightlining (i.e., tendency to provide identical answer in rating scales; Loosveldt and Beullens 2017) or item-nonresponse (Schwanhäuser et al. 2020). Other than these quality indicators, falsification indicators are often aggregated on the interviewer-level (Menold et al. 2013). These indicators have successfully been tested and used in different settings (see, for example, Bushery et al. 1999; Turner et al. 2002; Schräpler and Wagner 2005; Bredl, Winker, and Kötschau 2012; de Haas and Winker 2016; Schwanhäuser, Sakshaug, and Kosyakova 2022). Well known examples of formal falsification indicators include:

- *Share of triggered filters*: To keep interviews as short as possible, and therefore keep their effort low, falsifiers tend to use filter questions in order to take the shortest possible path through the questionnaire (Hood and Bushery 1997; Finn and Ranchhod 2017; Robbins 2018). Thereby, they trigger fewer follow-up questions compared to real respondents, which is measurable in the overall number of questions per interview, the number of triggered filters (triggering rate), or the number of answers that avoid follow-up questions (Hood and Bushery 1997; Storfinger and Opper 2011; Menold et al. 2013; Kosyakova, Skopek, Eckman 2015; Schwanhäuser, Sakshaug, and Kosyakova 2022).
- *Share of item nonresponse*: Falsifiers often tend to avoid the nonresponse categories in closed-ended questions, whereas real respondents usually show some level of refusal, measurable in different level of item nonresponse rates (Krejsa, Davis, and Hill 1999; Turner et al. 2002; Murphy et al. 2004; Bredl, Winker, and Kötschau 2012; Robbins 2018; Schwanhäuser, Sakshaug, and Kosyakova 2022).
- *Stereotypical responses*: Falsifiers tend to use their previous knowledge and stereotypical expectations to fabricate plausible responses (Inciardi 1981; Reuband 1990; DeMatteis et al. 2020), leading to homogeneous patterns especially across specific subgroups (e.g., migrants or minorities) (Storfinger and Opper 2011; Menold et al. 2013). One way of determining such behavior is through measures of scale consistency, e.g., Cronbach's alpha in item batteries (Schwanhäuser et al. 2020, 2022).

- *Acquiescent responding style*: A well-known phenomenon is that real respondents have a tendency of agreeing in surveys also known as acquiescent response behavior (Messick 1966). Falsifiers are less likely to show this type of behavior, resulting in lower level of agreement for opinion questions (Menold et al. 2013; de Haas and Winker 2016; Schwanhäuser, Sakshaug, and Kosyakova 2022).
- *Extreme and middle responding style*: Other than real respondents, falsifying interviewers have a stronger tendency of choosing the middle category rather than extreme values on ordinal response, resulting in measurable differences in the number of extreme and middle responding (Schäfer et al.2004a; Porras and English 2004; Bredl, Winker, and Kötschau 2012; Schwanhäuser, Sakshaug, and Kosyakova 2022).
- *Share of rounded answers*: Real respondents have the tendency of providing a relatively high share of rounded numbers to open numeric questions in order to reduce their cognitive effort, whereas falsifiers simply provide random numbers and hence are more likely to provide nonrounded numbers (Menold et al. 2013; Menold and Kemper 2014; Schwanhäuser, Sakshaug, and Kosyakova 2022).
- *Primacy and recency effects*: Presentation of questions varies for real respondents and falsifiers, as real respondents hear the questions and answers read by the interviewer (aural presentation), whereas a falsifier reads the questionnaire (visual presentation). This might lead to measurable differences in primacy and recency effects (Krosnick and Alwin 1987), with falsifiers having a higher tendency of choosing the first option of an answer list, and real respondents having a higher tendency of choosing the last option of an answer list (Menold et al. 2013; Menold and Kemper 2014; Winker et al. 2015; de Haas and Winker 2016; Schwanhäuser, Sakshaug, and Kosyakova 2022)
- *Open-ended questions and semi-open questions*: As falsifiers typically aim at reducing their time, effort, and possible inconsistencies, they tend to avoid open-ended items or open categories like “Other, specify”, measurable in a higher level of item nonresponse for this type of question (Storfinger and Opper 2011; Bredl et al. 2012; Menold et al. 2013; Menold and Kemper 2014; Winker et al. 2015; de Haas and Winker 2016; Schwanhäuser, Sakshaug, and Kosyakova 2022).
- *Response variability and nondifferentiation*: All mentioned behaviors contribute to another measurable pattern. Compared to real data, fabricated data is often characterized by a lower variance, both within an interview and between interviews (Porras and English 2004; Schäfer et al. 2004a, 2004b; Menold et al. 2013; Blasius and Thiessen 2015). This is for example measurable by calculating the standard deviation or variance

for item batteries or all categorical items of a questionnaire (see, e.g., Menold et al. 2013; Winker et al. 2013; Schwanhäuser, Sakshaug, and Kosyakova 2022).

Importantly, the assumptions measured via the indicators do not necessarily apply to all falsifiers. Falsifiers face a consistent trade-off between their falsification effort and the probability of being detected (Harrison and Krauss 2002; Menold et al. 2013; Menold and Kemper 2014). Consequently, the probability of detection and the interviewers' personal assessments of the situation or experiences may impact their concrete behavior or falsification strategy (Olbrich et al. 2023). For some indicators, this may imply that contrary values indicate fraudulent behavior, for example a higher share of item nonresponse, more acquiescent responding, or a higher share of rounded answers (Menold et al. 2013; Menold and Kemper 2014; Robbins 2018).

1.2.2.2 Content-Related Falsification Indicators

In contrast to formal indicators, content-related falsification indicators are employed to examine the distribution of variables or specific topics (Menold et al. 2013; Menold and Kemper 2014; Schwanhäuser et al. 2020). These indicators may also include the analysis of rare combinations, illogical patterns, or other distributional anomalies (Robbins 2018). However, as these indicators strongly depend on a survey's topic, they cannot be used in the same manner in every survey (Winker et al. 2013). Still, there are multiple examples for these types of indicators. The literature indicates that falsifiers are unable to reproduce the real distribution of rare or sensitive attributes (Schwanhäuser et al. 2020), such as the prevalence of minorities in the survey, household compositions, sexual behavior, drug use, or political participation of respondents (Hood and Bushery 1997; Turner et al. 2002; Murphy et al. 2004; Menold et al. 2013; Winker et al. 2013). Such information can be validated, by comparing means or frequencies between different interviewers or by comparing them to external sources. As a further detection tool, data can be checked for logical inconsistencies, or unusual combinations (Murphy et al. 2004; Porras and English 2004; Schwanhäuser et al. 2020). In the context of panel data, this concept can be extended to correlations of time-stable items between adjacent waves. As falsifiers are lacking in previous knowledge of the respondents' answers from previous waves, correlations between waves may be lower for falsifiers (Schräpler and Wagner 2005; Finn and Ranchhod 2017; Schwanhäuser et al. 2020).

Another widely utilized approach is Benford's Law, also known as Benford distribution (Benford 1938). Benford's Law describes the phenomenon that the first digit of

naturally occurring numbers (and thus some empirical numbers in surveys) follows a logarithmic probability distribution, the so-called Benford distribution (Hill 1999; Schäfer et al. 2004b; Walzenbach 2021). The distribution of open numeric answers, such as questions on income can be compared to the Benford distribution or slightly modified versions that precisely describe the empirical distribution of the survey (see, e.g., Schräpler and Wagner 2003; Swanson, Cho, and Eltinge 2003; Porras and English 2004; Schäfer et al. 2004a, 2004b; Bredl, Winker, and Kötschau 2012).

Some authors propose the addition of "trick" or control questions in surveys to identify falsifiers (see, e.g., Menold et al. 2013; Ziegler, Kemper, and Rammstedt 2013; Menold et al. 2013; Menold and Kemper 2014; Walzenbach 2021). These questions are special knowledge questions, which capitalize on the fact that falsifiers fail to reproduce real respondents' knowledge. For example, a list of fictitious and real response categories is presented to the respondents (e.g., a list of vocabulary or print media), asking them which items are known to them. As falsifiers may randomly select items, they are expected to select a higher number of fictitious items compared to real respondents (Menold et al. 2013; Winker et al. 2013; Menold and Kemper 2014; Winker et al. 2015, Schwanhäuser et al. 2020). In another example, interviewers received a wrong answer to a knowledge question during the interviewer training, assuming that falsifying interviewers would use the wrong answer from the training disproportionately often (Walzenbach 2021).

1.2.2.3 Paradata Falsification Indicators

Paradata are another valuable source for the identification of falsifications. One of the most important paradata for falsification identification are time stamps. Different studies have emphasized the usefulness of these time-based data (Cohen and Waren 2020; Schwanhäuser, Sakshaug, and Kosyakova 2022). Time stamps can be analyzed in multiple ways, for example by calculating the interview duration, relative duration per question, or module-level durations, identifying suspiciously short or long interviews or interview passages (Bushery et al. 1999; Li et al. 2011; Robbins 2018; Schwanhäuser, Sakshaug, and Kosyakova 2022). Additionally, short durations between two interviews, a large number of finished interviews shortly before the survey ends or on a certain day, as well as interviews conducted at suspicious day times might also indicate fraudulent behavior (Krejsa, Davis, and Hill 1999; Bushery et al. 1999; Murphy et al. 2004; Li et al. 2011; Robbins 2018).

Another paradata indicator is the number of available email addresses and telephone numbers (Stokes and Jones 1989; Turner et al. 2002). A high amount of missingness for this

information could indicate falsification behavior, as falsifiers might want to avoid that one can re-contact respondents and therefore verification of a falsification is not as easily possible (Stokes and Jones 1989; Bredl, Winker, and Kötschau 2012; Schwanhäuser, Sakshaug, and Kosyakova 2022). Additionally, other key statistics like response or success rates, and rates of eligible and ineligible households can be considered (DeMatteis et al. 2020). Schwanhäuser, Sakshaug, and Kosyakova (2022) have shown that consent rates to record linkage questions (i.e., linking data to external data sources) and the interviewers' post interview evaluation can also provide an indicator for fraudulent behavior.

1.2.2.4 Multivariate Analysis Strategies and Modelling Approaches

Common practice in data quality control procedures extends the concept of single indicators by combining various indicators in a multivariate fashion (Schwanhäuser et al. 2020). The joint analysis of indicators increases the possible evidence of fraudulent behavior and hence increases the power of the results. One popular technique is cluster analysis, which was first introduced by Bredl, Winker, and Kötschau (2012) in the context of falsification identification. Here, indicators are used as input to group interviewers based on their (dis-)similarities with regard to these indicators, allowing a group of "at risk" interviewers to be isolated (Bredl, Winker, and Kötschau 2012; DeMatteis et al. 2020; Schwanhäuser et al. 2020). Different studies have successfully demonstrated the use of a variety of different clustering algorithms, among others, Average Linkage, k-Means, Single Linkage, and Ward's Linkage (see, e.g., Bredl, Winker, and Kötschau 2012; Menold et al. 2013; Bergmann, Schuller, and Malter 2019; Schwanhäuser, Sakshaug and Kosyakova 2022; de Haas and Winker 2016). This method comes, however, with the drawback that it is not always easy to actually determine the "at risk" group (DeMatteis et al. 2020), as cluster analysis can result in multiple equally sized clusters.

Following the idea of combining multiple variables or levels, model-based methods like (multilevel) regression models open pathways for predicting an interviewer's falsification propensity (Li et al. 2011), identifying suspicious patterns by examining interviewer intraclass correlation coefficients (Landrock 2017b; Sharma and Elliott 2020), or examining potential fraudulent behavior by analyzing dynamics of indicators over time (Olbrich et al. 2023). These models can include a variety of variables, such as interviewer or respondent demographics, falsification indicators, or different levels (e.g., item-level, interview-level, interviewer-level) (DeMatteis et al. 2020). Importantly, these methods are flexibly adaptable to meet a project's specific quality control needs. For example, Olbrich et

al. (2023) were able to examine different types of fraudulent behavior by modelling interviewer effects on the intercept, scale, and slope of the interview sequence, capturing the dynamic behavior of interviewers over time. In a similar fashion, this approach can be used to identify outlying countries or regions in projects operating in multinational, multiregional, or multicultural contexts (Olbrich, Beckmann, and Sakshaug 2024). In addition to using regression models as a detection tool, some authors also evaluated the discriminative power of different indicators or identified new indicators using these models (Landrock 2017a; Walzenbach 2021).

Lastly, recent advancements in the area of machine learning open up new avenues for the detection of falsification (Buskirk et al. 2018), as machine learning methods can handle large datasets and identify complex patterns. One of the earliest examples of the application of machine learning is the study by Murphy et al. (2004). Using scoring models and anomaly detection, they identify suspicious patterns in response and paradata, and develop new falsification indicators specific to their data (Murphy et al. 2004). Other studies have demonstrated or proposed the use of machine learning tools like support vector machines, classification trees, naïve Bayes, ensemble-based regression trees, neural networks, density-based clustering method, or outlier detection methods (Birnbaum 2012; Birnbaum et al. 2013; Rosmansyah et al. 2019; Jebreel et al. 2020; Shah et al. 2020; Wienauer 2019). A major limitation of these studies is that they often lack comprehensive evaluation of their proposed methods or solely rely on simulated results. As a result, it is hardly possible to assess the effectiveness of these methods in the context of falsification detection. An exception are the contributions by Birnbaum (2012) and Birnbaum et al. (2013). Their results indicate that different algorithms (e.g., logistic regression, Bayesian network, and random forest) were able to precisely predict falsifications, with random forest being the best performing algorithm.

1.2.2.5 Identification of Identical Response Patterns

Another branch of the falsification detection literature focuses on the identification of duplicated answer patterns in the survey data. This includes duplicates, i.e., identical responses occurring across different interviews (Slomczynski, Powalko, and Krauze 2017), near-duplicates, i.e., records with a high number of identical responses across interviews (Kuriakose and Robbins 2016), and duplicated response patterns in item batteries (Blasius and Thiessen 2012; 2015). Especially in the case of (near-)duplicates it is important to note, that not only interviewers can be the source of them, but also other staff involved in the

survey, for example supervisors or project managers (Koczela et al. 2015; Sarracino and Mikucka 2016). Exact duplicates, which show the exact same answers to all responses, can be detected through simple duplicate analysis (Koczela et al. 2015; Kuriakose and Robbins 2016; Slomczynski, Powalko, and Krauze 2017; Schwanhäuser et al. 2020). The cumulated occurrence of duplicates within a single interviewer's workload may indicate interviewer falsification.

As changes in a single answer can disguise a duplicated record, Kuriakose and Robbins (2016) proposed strategies to identify near-duplicates. This "high-matching" or "percent-matching" method identifies data with a high correspondence of answers (usually a correspondence of 85 to 99 percent of answers is defined as near duplicate) (Kuriakose and Robbins 2016; Schwanhäuser et al. 2020). Even though the method was rightly criticized as it is sensitive to various characteristics of a survey, including the number of questions, the number of respondents, the homogeneity within the population and subgroups (Simmons et al. 2016), applications of the method also indicate its usefulness for detecting problems with the data quality (Cohen and Warner 2020; Schwanhäuser et al. 2020 Olbrich, Beckmann, and Sakshaug 2024).

Finally, some studies shift the focus towards duplicated, repetitive, and outlying response patterns in item batteries or parts of the questionnaire. Using (categorical) Principal Component Analysis (PCA) or Multiple Correspondence Analysis (MCA), it is possible to identify interviews with low variability, which is often – as described above – a characteristic of falsified interviews (Blasius and Thiessen 2012, 2013, 2015). By calculating factor scores, the methods reduce the dimensionality of the items but still preserve their true variability across answers (Blasius and Thiessen 2012, 2013, 2015, 2021; DeMatteis et al. 2020; Schwanhäuser et al. 2020). An extension of these approaches uses the Hamming distance instead of PCA and MCA to assess the distance (or similarity) between different interviews (Blasius and Sausen 2023).

Further information: Costs and benefits of different detection methods

It is important to emphasize that each of the described detection methods comes with certain benefits and problems. Some may be easy to implement, but their results require a lot of interpretation (Schwanhäuser, Sakshaug, and Kosyakova 2022). Others may work better on certain falsification types (Schwanhäuser et al. 2020). This underlines the importance of combining different approaches, to enhance quality controls and target different types of fraudulent behavior (Thissen and Myers 2016; DeMatteis et al. 2020). But with an increasing

number of control measures, expenditure and costs of controls also rise. In particular, detection and verification of milder falsification forms (e.g., deviations from selection rules) can be challenging. Therefore, practitioners need to weigh up the costs and benefits of different detection methods. To make informed decisions, they need information regarding the true costs and benefits associated with each method. This underscores the importance of practitioners making the outcomes of their quality controls, evaluations of different methods, and data including identified falsifications publicly available (Winker 2016; DeMatteis et al. 2020).

1.3 Focus and Agenda of This Dissertation

This dissertation contributes to the research on the detection of interviewer falsification using statistical detection tools. Even though literature has proposed numerous of such data-based tools for detecting interviewer falsification in surveys, most studies neglect testing these methods in systematic ways. Concrete evaluations of multiple methods are mostly missing from the literature and the few existing evaluations are often based on experimental data or simulated fraud cases. Further, most methods have a strong focus on complete falsifications and neglect other forms of falsifications. Hence, it is hard for practitioners to weigh up the costs and benefits of different methods and designing own targeted quality control procedures.

To close this gap, this dissertation presents a practical guide for improving quality controls with respect to interviewer falsification. The individual contributions (see **Table 1.2** for an overview) evaluate various statistical detection tools under different circumstances. The three evaluations encompass a multitude of falsification indicators, multivariate methods, namely cluster analysis, methods focusing on duplication (e.g., duplicate analysis, percent-matching method, and principal component analysis), and innovative machine learning methods. The analyses are based on a variety of datasets, including real-world survey data with known cases of interviewer falsification, as well as experimental data. This combination allows for a more comprehensive examination of the efficiency of identification methods under varying conditions. Additionally, this dissertation offers insight into the detection of falsifications in the context of different data collection designs. Rather than solely focusing on cross-sectional data or single waves of longitudinal data, this dissertation explicitly considers the structure of panel data. Lastly, this dissertation not only addresses the identification of complete falsifications but also examines the performance of the detection methods in the context of partial falsification.

Table 1.2: Overview of contributions in this dissertation.

Paper	Published/submitted	Coauthored by
How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification	Public Opinion Quarterly, Volume 86, Issue 1, 2022, https://doi.org/10.1093/poq/nfab066	Joseph W. Sakshaug Yuliya Kosyakova
How Falsifiers Make a Long Story Short: Identifying Partial Interviewer Falsification in Panel Surveys	Submitted November 9th, 2023	Jonas Beste Lukas Olbrich Joseph W. Sakshaug
Leaving No Data Unturned: Evaluating Machine Learning Algorithms to Detect Interviewer Falsification	To be submitted	Joseph W. Sakshaug Natalja Menold Peter Winker

As this dissertation covers a broad range of aspects regarding the data-based detection of interviewer falsification, it supports survey practitioners in selecting fitting tools for their data quality controls, or for improving existing ones. As different settings, survey designs, and types of fraudulent interviewer behavior are covered, the different contributions provide a rich guideline for various types of interviewer-administered surveys. The following chapters provide a detailed summary of the individual contributions.

1.3.1 How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification

The first paper entitled "How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification" focuses on the performance evaluation and comparison of different statistical detection methods in a real-world setting. Using data from the first wave of the IAB-BAMF-SOEP Survey of Refugees in Germany, including approximately 7.5 percent falsified interviews, the paper tests newly proposed and existing falsification indicators, their combined use in cluster analysis, and additionally introduces a new multivariate detection method (meta-indicator approach) that overcomes some practical limitations of other detection methods. The study showcases a broad number of indicators, using a total of 32 falsification indicators, based on person and household interviews as well as the paradata of the survey. These indicators are jointly used as input variables to Single-Linkage clustering, Ward's Linkage clustering, and the newly proposed meta-indicator. The basic idea of the meta-indicator is to summarize interviewer-level indicators into a single indicator, by summing up all standardized indicator values. To evaluate the performance of these three methods the paper considers five different quality measures (false-positive rate, false-negative rate, accuracy, error rate, and Cohen's kappa). To further test the robustness

of the multivariate methods, a leave-one-out procedure is applied, excluding one indicator after another to see if results change based on single indicators. Lastly, the paper evaluates the relative importance and explanatory power of each indicator for identifying the falsifiers, by applying a discriminant analysis. In that it is also possible to evaluate the directional assumptions behind the indicators (e.g., whether falsifiers really show less item nonresponse, less rounded values etc.). The main finding of the paper confirms the effectiveness of multivariate detection methods, including the newly proposed meta-indicator. These results were robust to leaving single indicators out. Further results show, that most indicators successfully differentiate between real and falsified interviews, with indicators based on time stamps being of highest importance. For practitioners, this implies that the combined use of indicators is an effective tool for detecting falsifiers.

1.3.2 How Falsifiers Make a Long Story Short: Identifying Partial Interviewer Falsification in Panel Surveys

The second paper, "How Falsifiers Make a Long Story Short: Identifying Partial Interviewer Falsification in Panel Surveys", aims at identifying effective falsification detection methods, which are suitable in the context of panel surveys, and that are not only capable of identifying complete falsifications but also partial interviewer falsification. The paper, hence, evaluates detection methods that could be used in a cross-sectional setting as well as methods which specifically make use of the panel structure, by comparing data between subsequent waves. Analyses and evaluations are based on data from nine waves of the German panel study "Labour Market and Social Security" (PASS). Enhanced quality control checks, implemented during the COVID-19 pandemic, brought to light two interviewers responsible for partial falsifications. As one of the key interests of the paper lies in the identification of partial falsifications, the analyses include detection methods on different levels, namely the interviewer-level (i.e., methods aggregating results within an interviewers' workload), the respondent-level (i.e., methods focusing on single interviews), and the item-level (i.e., methods focusing on parts of the interview). Detection methods in the cross-sectional setting include five indicators, cluster analysis, Isolation Forest which is an outlier detection method, (near-)duplicate analysis, and Principal Component Analysis. The longitudinal detection methods focus on the common notion that falsifiers produce less stable answers between adjacent waves which should be measurable in correlations between items from adjacent waves. Again, correlations are compared on the interviewer-, the respondent-, as well as on the item-level. The analyses reveal that most cross-sectional analysis are effective in identifying the partial falsifications in the later waves. Especially

falsification indicators, cluster analysis, and isolation forest showed high detection precision. However, this does not hold for earlier waves. This indicates that the falsifiers showed learning effects, i.e., that their falsification behavior got more careless over time, making their detection easier. This also explains why the falsifiers were only detected after nine waves of the survey. Less promising results were observed for the longitudinal analysis on correlations. Even though the falsifiers showed slightly lower correlations than most other interviewers, they are no clear outliers. Importantly, the results contradict the common notion that falsifiers would be easier to detect in panels surveys and emphasizes the importance of solid and effective data quality controls. Overall, this implies that even in the panel context, practitioners can use commonly used cross-sectional methods to detect various forms of falsification, including partial falsifications.

1.3.3 Leaving no Data Unturned: Evaluating Machine Learning Algorithms to Detect Interviewer Falsification

Lastly, the paper "Leaving no Data Unturned: Evaluating Machine Learning Algorithms to Detect Interviewer Falsification" evaluates the potentials of supervised machine learning in the context of falsification detection. Supervised algorithms are trained on existing falsification data and hence make use of the distinct patterns produced by falsifiers. Applying a total of 14 supervised algorithms, belonging into the broader groups of regression models, decision trees, support vector machines, or neural networks, the paper is able to exploit the potentials of various algorithm types in picking up falsifiers' true patterns in the data. As supervised algorithms require data including falsified interviews, the paper relies on two distinct data sources: real-world survey data form the IAB-BAMF-SOEP Survey of Refugees in Germany, as well as experimental data coming from a large experiment conducted at the University of Giessen. To enrich the scope of the paper, the different algorithms are tested in three different scenarios. In the first scenario, both datasets were analyzed separately. Each of the two datasets were randomly divided into a training and a test dataset, training the algorithms on the falsifications in the training data to predict the binary status of the interviews in the test data (real interview or falsification). Hence, this part evaluates the effectiveness of the algorithms when training them on falsification within the same survey. In the second scenario, the experimental data are split in a similar fashion, but ensuring that all interviews conducted by one interviewer are either assigned to the training or the test data. Hence, this part evaluates the effectiveness of the algorithms when training them on falsifications induced by different falsifiers within the same survey. Lastly, the final scenario utilizes the experimental data as training data, testing the algorithms'

ability to detect falsifications in the real-world refugee data. Hence, the scenario tests the effectiveness when training an algorithm on falsification from a completely different survey. As both datasets contain different variables and hence would not be comparable, training and testing is done based on eleven standardized falsification indicators, which were available for both datasets. The results demonstrate that multiple algorithms were capable of accurately identifying falsifications within the same survey, as evidenced in both the first and second scenarios. Especially algorithms based on decision trees showed solid outcomes. However, performance of all algorithms strongly decreases in the between-survey scenario and no algorithm was able to precisely identify the falsification in the other survey. The results indicate that supervised machine learning could be helpful tools to identify falsifications, provided that training data from the same survey is available. This may be the case if other instances of falsification were discovered within the field, or if falsifications were discovered in an earlier wave of a panel survey.

References

Benford, Frank. 1938. “The Law of Anomalous Numbers.” Proceedings of the American Philosophical Society 78(4):551–72.

Bergmann, Michael, Karin Schuller, and Frederic Malter. 2019. “Preventing Interview Falsifications During Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE).” Longitudinal and Life Course Studies 10(4):513–30.

Beste, Jonas, Lukas Olbrich, and Silvia Schwanhäuser. 2021. “Interviewer: innenkontrolle im Panel Arbeitsmarkt und soziale Sicherung (PASS).” Institut für Arbeitsmarkt- und Berufsforschung. Available at https://doku.iab.de/fdz/reporte/2021/MR_04-21.pdf.

Biemer, Paul P., and Lars E. Lyberg. 2003. Introduction to survey quality. John Wiley & Sons.

Biemer, Paul P., and S. Lynne Stokes 1989. “The Optimal Design of Quality Control Samples to Detect Interviewer Cheating.” Journal of Official Statistics 5(1):23–39.

Birnbaum, Benjamin. 2012. “Algorithmic Approaches to Detecting Interviewer Fabrication in Surveys.” Dissertation, University of Washington. Available at http://hdl.handle.net/1773/22011.

Birnbaum, Benjamin, Gaetano Borriello, Abraham D. Flaxman, Brian DeRenzi, and Anna R. Karlin. 2013. “Using behavioral data to identify interviewer fabrication in surveys.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Available at https://bbirnbaum.com/assets/publications/chi13.pdf.

Blasius, Jörg, and Lukas Sausen. 2023. “Detecting Fabricated Interviews Using the Hamming Distance.” Survey Research Methods 17(2):131–45.

Blasius, Jörg, and Jürgen Friedrichs. 2012. “Faked Interviews.” In Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt, edited by Samuel Salzborn, Eldad Davidov, and Jost Reinecke, 49–56. Wiesbaden: VS Verlag für Sozialwissenschaften.

Blasius, Jörg, and Victor Thiessen. 2012. Assessing the Quality of Survey Data. SAGE Publications.

Blasius, Jörg, and Victor Thiessen. 2013. “Detecting Poorly Conducted Interviews.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 67–88. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Blasius, Jörg, and Victor Thiessen. 2015. “Should we trust survey data? Assessing response simplification and data fabrication.” Social Science Research 52:479–93.

Blasius, Jörg, and Victor Thiessen. 2021. “Perceived corruption, trust, and interviewer behavior in 26 European Countries.” Sociological Methods & Research 50(2):740–77.

Bossler, Mario, Nicole Gürtzgen, Alexander Kubis, Benjamin Küfner, Lukas Olbrich, and Silvia Schwanhäuser. 2022. “Revision and new data quality concept due to deviant interviewer behavior in the IAB Job Vacancy Survey.” Institut für Arbeitsmarkt- und Berufsforschung. Available at https://doku.iab.de/fdz/reporte/2022/MR_05-22_EN.pdf.

Bredl, Sebastian, Nina Storfinger, and Natalja Menold. 2013. “A Literature Review of Methods to Detect Fabricated Survey Data.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 3–24. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Bredl, Sebastian, Peter Winker, and Kerstin Kötschau. 2012. “A Statistical Approach to Detect Interviewer Falsification of Survey Data.” Survey Methodology Journal 38(1):1–10. Available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012001/article/11680-eng.pdf.

Brüderl, Josef, Bernadette Huyer-May, and Claudia Schmiedeberg. 2013. “Interviewer behavior and the quality of social network data.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 147–60. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Bushery, John M., Jennifer W. Reichert, Keith A. Albright, and John C. Rossiter. 1999. “Using Date and Time Stamps to Detect Interviewer Falsification.” Proceedings of the Survey Research Method Section, American Statistical Association, 316–20. Available at http://www.asasrms.org/Proceedings/papers/1999_053.pdf.

Buskirk, Trent D., Antje Kirchner, Adam Eck, and Curtis S. Signorino. 2018. “An introduction to machine learning methods for survey researchers.” Survey Practice 11(1).

Castorena, Oscar, Mollie J. Cohen, Noam Lupu, and Elizabeth J. Zechmeister. 2023. “How worried should we be? The implications of fabricated survey data for political science.” International Journal of Public Opinion Research 35(2):1–9.

Cohen, Mollie J., and Zach Warner. 2021. “How to Get Better Survey Data More Efficiently.” Political Analysis 29(2):121–38.

Crespi, Leo P. 1945. “The cheater problem in polling.” Public Opinion Quarterly 9(4):431–45.

Davis, Justine M., and Martha Wilfahrt. 2024. “Enumerator Experiences in Violent Research Environments.” Comparative Political Studies 57(4):675–709.

de Haas, Samuel, and Peter Winker. 2014. “Identification of partial falsifications in survey data.” Statistical Journal of the IAOS 30(3):271–281.

de Haas, Samuel, and Peter Winker. 2016. “Detecting Fraudulent Interviewers by Improved Clustering Methods—The Case of Falsifications of Answers to Parts of a Questionnaire.” Journal of Official Statistics 32(3):643–60.

DeMatteis, Jill M., Linda J. Young, James Dahlhamer, Ronald E. Langley, Joe Murphy, Kristen Olson, and Sharan Sharma. 2020. “Falsification in Surveys: Task Force Final Report.” Washington, DC: American Association for Public Opinion Research. Available at https://www.aapor.org/wp-content/uploads/2022/11/AAPOR_Data_Falsification_Task_Force_Report-updated.pdf.

Eckman, Stephanie, and Achim Koch. 2019. “Interviewer involvement in sample selection shapes the relationship between response rates and data quality.” Public Opinion Quarterly 83(2):313–37.

Edwards, Brad, Aaron Maitland, and Sue Connor. 2017. “Measurement error in survey operations management: detection, quantification, visualization, and reduction.” In Total Survey Error in Practice, edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, Brady T. West, 253–77. John Wiley & Sons.

Edwards, Brad, Hanyu Sun, and Ryan Hubbard. 2020. “Behavior Change Techniques for Reducing Interviewer Contributions to Total Survey Error.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 77–89. Boca Raton, FL: Taylor & Francis Group.

Fee, Holly, Jason Fields, and Matthew Marlay. 2016. “Computer Audio-Recorded Interviewing and Data Quality: Findings from Wave 1 of the 2014 Survey of Income and Program Participation.” 2016 Population Association of America (PAA) Annual Meeting. Available at https://paa.confex.com/paa/2016/mediafile/ExtendedAbstract/Paper7889/Fee_PAA_2016.pdf

Fee, Holly, T. Andy Welton, Matthew Marlay, and Jason Fields. 2015. “Using Computer-Assisted Recorded Interviewing to Enhance Field Monitoring and Improve Data Quality.” Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference. Available at https://nces.ed.gov/fcsm/pdf/J1_Fee_2015FCSM.pdf

Finn, Arden, and Vimal Ranchhod. 2017. “Genuine fakes: The prevalence and implications of data fabrication in a large South African survey.” The World Bank Economic Review 31(1):129–57.

Fowler Jr., Floyd J. 2013. Survey research methods. SAGE publications.

Fowler Jr., Floyd J., and Thomas Mangione. 1990. Standardized Survey Interviewing. Newbury Park, London, and Greater Kailash: Sage.

Gomila, Robin, Rebecca Littman, Graeme Blair, and Elizabeth Levy Paluck. 2017. “The audio check: A method for improving data quality and detecting data fabrication.” Social Psychological and Personality Science 8(4):424–33.

Groves, Robert M. 2004. “Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects.” Survey Research 35(1):1–5.

Groves, Robert M. 2005. Survey errors and survey costs. John Wiley & Sons.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. Survey Methodology. Hoboken, NJ: Wiley.

Gwartney, Patricia A. 2013. “Mischief versus mistakes: Motivating interviewers to not deviate.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 195–215. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Harrison, David E., and Stefanie I. Krauss. 2002. “Interviewer cheating: Implications for research on entrepreneurship in Africa.” Journal of Developmental Entrepreneurship 7(3): 319–30.

Hill, Theodore P. 1999. “The Difficulty of Faking Data.” Chance 12(3):27–31.

Hood, Catherine C., and John M. Bushery. 1997. “Getting More Bang from the Reinterview Buck: Identifying ‘At Risk’ Interviewers.” Proceedings of the Survey Research Method Section, American Statistical Association, 820–24. Available at http://www.asasrms.org/proceedings/papers/1997_141.pdf.

Hülser, Oliver. 2013. “Automatic Interview Control of Market Research Studies” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 103–16. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Inciardi, James A. 1981. “Fictitious data in drug abuse research.” International Journal of the Addictions 16(2):377–80.

Japec, Lilli. 2006. “Quality issues in interview surveys-Some contributions.” Bulletin of sociological methodology/Bulletin de méthodologie sociologique 90(1):26–42. Available at https://journals.sagepub.com/doi/abs/10.1177/075910630609000104.

Jebreel, Najeeb Moharram, Rami Haffar, Ashneet Khandpur Singh, David Sánchez, Josep Domingo-Ferrer, and Alberto Blanco-Justicia. 2020. “Detecting bad answers in survey data through unsupervised machine learning.” In Privacy in Statistical Databases, UNESCO Chair in Data Privacy, International Conference Proceedings, edited by Josep Domingo-Ferrer and Krishnamurty Muralidhar, 309–20. Springer International Publishing.

Jesske, Birgit. 2013. “Concepts and Practices in Interviewer Qualification and Monitoring.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 91–102. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Josten, Michael, and Mark Trappmann. 2016. “Interviewer effects on a network-size filter question.” Journal of Official Statistics 32(2):349–73.

Kindred, G. Machell, and Jimmie B. Scott. 1993. “Fabrication during the 1990 nonresponse followup operation.” JSM Proceedings, Survey Research Methods Section. Available at http://www.asasrms.org/Proceedings/papers/1993_053.pdf.

Koch, Achim. 1995. “Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994.” ZUMA Nachrichten 19(36):89–105.

Koczela, Steve, Cathy Furlong, Jaki McCarthy, and Ali Mushtaq. 2015. “Curbstoning and Beyond: Confronting Data Fabrication in Survey Research.” Statistical Journal of the IAOS 31(3):413–22.

Kosyakova, Yuliya, Lukas Olbrich, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2019. “Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany.” Institut für Arbeitsmarkt- und Berufsforschung. Available at https://fdz.iab.de/187/section.aspx/Publikation/k190404302.

Kosyakova, Yuliya, Jan Skopek, and Stephanie Eckman. 2015. “Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach.” International Journal of Public Opinion Research 27(3):417–31.

Krejsa, Elizabeth A., Mary C. Davis, and Joan M. Hill. 1999. “Evaluation of the quality assurance falsification interview used in the census 2000 dress rehearsal.” Proceedings of the American Statistical Association (Survey Research Methods Section). Available at http://www.asasrms.org/Proceedings/y1999f.html.

Kreuter, Frauke. 2013. Improving surveys with paradata: Analytic uses of process information. John Wiley & Sons.

Krosnick, Jon A., and Duane F. Alwin. 1987. “An evaluation of a cognitive theory of response-order effects in survey measurement.” Public opinion quarterly 51(2): 201–19.

Kuriakose, Noble, and Michael Robbins. 2016. “Don’t Get Duped: Fraud Through Duplication in Public Opinion Surveys.” Statistical Journal of the IAOS 32(3):283–91.

Lam, David, Cally Ardington, Nicola Branson, Anne Case, Murray Leibbrandt, Brenda Maughan-Brown, Alicia Menendez, Jeremy Seekings, and Meredith Sparks. 2012. “The Cape Area Panel Study: Overview and Technical Documentation Waves 1-2-3-4-5 (2002-2009).” University of Michigan, University of Cape Town. Available at https://microdata.worldbank.org/index.php/catalog/895/download/30341.

Landrock, Uta. 2017a. “Explaining Political Participation: A Comparison of Real and Falsified Survey Data.” Statistical Journal of the IAOS 33(2):447–58.

Landrock, Uta. 2017b. “Differences between real and falsified data.” Justus-Liebig-Universität, Gießen. Available at https://core.ac.uk/download/pdf/153558393.pdf.

Li, Jianzhu, J. Michael Brick, Back Tran, and Phyllis Singer. 2011. “Using Statistical Models for Sample Design of a Reinterview Program.” Journal of Official Statistics 27(3):433–50.

Loosveldt, Geert, and Koen Beullens. 2017. “Interviewer effects on non-differentiation and straightlining in the European Social Survey.” Journal of official statistics 33(2):409–26.

May, Julian D., Jorge Agüero, Michael R. Carter, and Ian M. Timaeus. 2007. “The KwaZulu-Natal Income Dynamics Study (KIDS) third wave: methods, first findings and an agenda for future research.” Development Southern Africa 24(5):629–48.

Menold, Natalja, Uta Landrock, Peter Winker, Nathalie Pellner, and Christoph J. Kemper. 2018. “The impact of payment and respondents’ participation on interviewers’ accuracy in face-to-face surveys: Investigations from a field experiment.” Field Methods 30(4):295–311.

Menold, Natalja. 2014. “The influence of sampling method and interviewers on sample realization in the European Social Survey.” Survey Methodology 40(1):105–23.

Menold, Natalja, and Christoph J Kemper. 2014. “How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys.” International Journal of Public Opinion Research 26(1):41–65.

Menold, Natalja, Peter Winker, Nina Storfinger, and Christoph J. Kemper. 2013. “A Method for Ex-Post Identification of Falsification in Survey Data.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 25–47. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Messick, Samuel. 1966. “The psychology of acquiescence: an interpretation of research evidence 1.” ETS research bulletin series 1966(1):i–44.

Murphy, Joe, Paul Biemer, Chris Stringer, Rita Thissen, Orin Day and Y. Patrick Hsieh. 2016. “Interviewer falsification: Current and best practices for prevention, detection, and mitigation.” Statistical Journal of the IAOS 32(3):313–26.

Murphy, Joe, Joe Eyerman, Colleen McCue, Christy Hottinger, and Joel Kennet. 2005. “Interviewer Falsification detection using data mining.” Proceedings of Statistics Canada Symposium 2005, Methodological Challenges for Future Information Needs. Available at https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X20050019445.

Murphy, Joe, Rodney Baxter, Joe Eyerman, David Cunningham, and Joel Kennet. 2004. “A System for Detecting Interviewer Falsification.” Proceedings of the American Statistical Association and the American Association for Public Opinion Research. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000517.pdf.

Nelson, James E., and Pamela L. Kiecker. 1996. “Marketing research interviewers and their perceived necessity of moral compromise.” Journal of Business Ethics 15:1107–17.

Olbrich, Lukas, Elisabeth Beckmann, and Joseph W. Sakshaug. 2024. “Multivariate assessment of interviewer-related errors in a cross-national economic survey.” Working Paper No. 253. Österreichische Nationalbank (OeNB). Available at https://www.econstor.eu/handle/10419/286404.

Olbrich, Lukas, Yuliya Kosyakova, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2023. “Detecting Interviewer Fraud Using Multilevel Models.” Journal of Survey Statistics and Methodology 12(1):14–35.

Olson, Kristen, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West. 2020. “The Past, Present, and Future of Research on Interviewer Effects.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 3–16. Boca Raton, FL: Taylor & Francis Group.

Pashazadeh, Fiona, Alexandru Cernat, and Joseph W. Sakshaug. 2020. “Investigating the Use of Nurse Paradata in Understanding Nonresponse to Biological Data Collection.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 221–34. Boca Raton, FL: Taylor & Francis Group.

Porras, Javier, and Ned English. 2004. “Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys.” Proceedings of the Survey Research Method Section, American Statistical Association, 4223–28. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000879.pdf.

Reuband, Karl-Heinz. 1990. “Interviews, Die Keine Sind: ‘Erfolge’ Und ‘Mißerfolge’ Beim Fälschen Von Interviews.” Kölner Zeitschrift für Soziologie und Sozialpsychologie 42(4):706–33.

Robbins, Michael. 2018. “New frontiers in detecting data fabrication.” In Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC), edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop, and Brita Dorer, 771–805 Wiley & Sons, Inc.

Rosmansyah, Yusep, Ibnu Santoso, Ariq Bani Hardi, Atina Putri, and Sarwono Sutikno. 2019. “Detection of Interviewer Falsification in Statistics Indonesia's Mobile Survey.” International Journal on Electrical Engineering and Informatics 11(3): 474–84.

Sarracino, Francesco, and Malgorzata Mikucka. 2017. “Bias and efficiency loss in regression estimates due to duplicated observations: a Monte Carlo simulation.” Survey Research Methods 11(1):17–44.

Sarracino, Francesco, and Malgorzata Mikucka. 2016. “Estimation bias due to duplicated observations: a Monte Carlo simulation.” Available at https://mpra.ub.uni-muenchen.de/69064/.

Schäfer, Christin, Jörg-Peter Schräpler, Klaus-Robert Müller, and Gert G. Wagner. 2004a. “Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methodes.” DIW Discussion Paper No. 441. Berlin: DIW–German Institute for Economic Research. Available at http://hdl.handle.net/10419/18293.

Schäfer, Christin, Jörg-Peter Schräpler, and Klaus-Robert Müller. 2004b. “Identification, Characteristics and Impact of Faked and Fraudulent Interviews in Surveys.” European Conference on Quality and Methodology in Official Statistics. Available at https://www.diw.de/documents/dokumentenarchiv/17/41963/paper2004_schaeferetal.pdf.

Schnell, Rainer. 2012. Survey-Interviews. Springer.

Schnell, Rainer. 1991. “Der Einfluß gefälschter Interviews auf Survey-Ergebnisse.” Zeitschrift für Soziologie 20(1):25–35.

Schräpler, Jörg-Peter, and Gert G. Wagner. 2003. “Identification, characteristics and impact of faked interviews in surveys: An analysis by means of genuine fakes in the raw data of SOEP.” IZA Discussion Paper No. 969. Institute for the Study of Labor (IZA). Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=487402.

Schräpler, Jörg-Peter, and Gert G. Wagner. 2005. “Characteristics and Impact of Faked Interviews in Surveys: An Analysis of Genuine Fakes in the Raw Data of SOEP.” Allgemeines Statistisches Archiv 89:7–20.

Schreiner, Irwin D., Jennifer Newbrough, and Karen Pennie. 1988. “Interviewer falsification in Census Bureau surveys.” Proceedings from Section on Survey Research Methods, 491–96. Available at http://www.asasrms.org/Proceedings/papers/1988_090.pdf.

Schwanhäuser, Silvia, Joseph W. Sakshaug, and Yuliya Kosyakova. 2022. “How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification.” Public Opinion Quarterly 86(1):51–81.

Schwanhäuser, Silvia, Joseph W Sakshaug, Yuliya Kosyakova, and Frauke Kreuter. 2020. “Statistical identification of fraudulent interviews in surveys: improving interviewer controls.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 91–106. Boca Raton, FL: Taylor & Francis Group.

Shah, Neha, Diwakar Mohan, Jean Juste Harisson Bashingwa, Osama Ummer, Arpita Chakraborty, and Amnesty E. LeFevre. 2020. “Using machine learning to optimize the quality of survey data: protocol for a use case in India.” JMIR Research Protocols 9(8). Available at https://www.researchprotocols.org/2020/8/e17619/.

Sharma, Sharan, and Michael R. Elliott. 2020. “Detecting Falsifications in a Television Audience Measurement Panel Survey.” International Journal of Market Research 62(4):432–48.

Simmons, Katie, Andrew Mercer, Steve Schwarzer and Courtney Kennedy. 2016. “Evaluating a new proposal for detecting data falsification in surveys.” Statistical Journal of the IAOS 32(3): 327–38.

Slomczynski, Kazimierz Maciek, Przemek Powalko, and Tadeusz Krauze. 2017. “Non-Unique Records in International Survey Projects: The Need for Extending Data Quality Control.” Survey Research Methods 11(1):1–16.

Spagat, Michael. 2016. “Comment on ‘Don't get duped: Fraud through duplication in public opinion surveys’.” Statistical Journal of the IAOS 32(3):29–94.

Stokes, S. Lynne, and Patty Jones. 1989. “Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey.” Proceedings of the Survey Research Method Section, American Statistical Association, 696–98. Available at http://www.asasrms.org/Proceedings/papers/1989_127.pdf.

Storfinger, Nina, and Marie Opper. 2011. “Datenbasierte Indikatoren für potenziell abweichendes Interviewerverhalten.” Discussion Paper 58, ZEU, Giessen. Available at: http://geb.uni-giessen.de/geb/volltexte/2012/8559/pdf/ZeuDiscPap58.pdf

Storfinger, Nina, and Peter Winker. 2013. “Assessing the Performance of Clustering Methods in Falsification Using Bootstrap.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 46–65. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Sun, Hanyu, and Ting Yan. 2023. “Applying Machine Learning to the Evaluation of Interviewer Performance.” Survey Practice 16(1).

Swanson, David, Moonung Cho, and John Eltinge. 2003. “Detecting possibly fraudulent or error-prone survey data using Benford’s Law.” Proceedings of the Survey Research Method Section, American Statistical Association. Available at http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000205.pdf.

Thissen, M. Rita, and Susan K. Myers. 2016. “Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality.” Statistical Journal of the IAOS 32(3):339–47.

Turner, Charles F., James N. Gribble, Alia A. Al-Tayyib, and James R. Chromy. 2002. “Falsification in Epidemiologic Surveys: Detection and Remediation.” Technical Papers on Health and Behavior Measurement, No. 53. Washington, DC: Research Triangle Institute.

Wagner, James, Kristen Olson, and Minako Edgar. 2017. “The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata.” Survey Research Methods 11(3):218–33.

Walzenbach, Sandra. 2021. “Do falsifiers leave traces? Finding recognizable response patterns in interviewer falsifications.” Methods, data, analyses: a journal for quantitative methods and survey methodology (mda) 15(2):125–60.

Weinauer, Marlene. 2019. “Be a Detective for a Day: How to Detect Falsified Interviews with Statistics.” Statistical Journal of the IAOS 35(4):569–75.

Werker, Henry F. 1981. “Results of the 1980 US census challenged.” Population and Development Review 7(1):155–67.

West, Brady T., and Annelies G. Blom. 2017. “Explaining interviewer effects: A research synthesis.” Journal of survey statistics and methodology 5(2):175–211.

West, Brady T., Ting Yan, Frauke Kreuter, Michael Josten, and Heather Schroeder. 2020. “Examining the utility of interviewer observations on the survey response process.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 107–20. Boca Raton, FL: Taylor & Francis Group.

Winker, Peter. 2016. “Assuring the quality of survey data: Incentives, detection and documentation of deviant behavior.” Statistical Journal of the IAOS 32(3):295–303.

Winker, Peter, Karl-Wilhelm Kruse, Natalja Menold, and Uta Landrock. 2015. “Interviewer effects in real and falsified interviews: Results from a large scale experiment.” Statistical Journal of the IAOS 31(3):423–34.

Winker, Peter, Natalja Menold, and Rolf Porst. 2013. Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Winker, Peter, Natalja Menold, Nina Storfinger, Christoph J. Kemper, and Sabrina Stukowski. 2013. “A Method for Ex-Post Identification of Falsifications in Survey Data.” Paper presented at New Techniques and Technologies for Statistics (NTTS), Brussels, March 5–7. Available at https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_93.pdf.

Ziegler, Matthias, Christoph Kemper, and Beatrice Rammstedt. 2013. “The vocabulary and overclaiming test (VOC-T).” Journal of Individual Differences 34(1):32–40.

2. How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification

Abstract

Deviant interviewer behavior is a potential hazard of interviewer-administered surveys, with interviewers fabricating entire interviews as the most severe form. Various statistical methods (e.g., cluster analysis) have been proposed to detect falsifiers. These methods often rely on falsification indicators aiming to measure differences between real and falsified data. However, due to a lack of real-world data, empirical evaluations and comparisons of different statistical methods and falsification indicators are scarce. Using a large-scale nationally representative refugee survey in Germany with known fraudulent interviews, this study tests, evaluates, and compares statistical methods for identifying falsified data. We investigate the use of new and existing falsification indicators as well as multivariate detection methods for combining them. Additionally, we introduce a new and easy-to-use multivariate detection method that overcomes practical limitations of previous methods. We find that the vast majority of used falsification indicators successfully measure differences between falsifiers and nonfalsifiers, with the newly proposed falsification indicators outperforming some existing indicators. Furthermore, different multivariate detection methods perform similarly well in detecting the falsifiers.

2.1 Introduction

Interviewer-administered surveys are often treated as a superior form of data collection e.g., concerning response rates, communication with respondents, and administration of long questionnaires (Groves et al. 2009; Olson et al. 2020). By encouraging respondents' participation, answering their queries, and ensuring questionnaire completion, interviewers play a vital role for survey quality. However, previous research has emphasized numerous possible survey errors attributable to the interviewer (Fowler and Mangione 1990; Groves 2004). The falsification of survey interviews is one specific and understudied error associated with the interviewer. Interviewer falsification may take various forms such as intentional miscoding of respondents' eligibility status or answers, deviations from instructions, and, the most severe form, the fabrication of complete interviews (AAPOR 2003). Although empirical evidence suggests that complete falsification is a rare event (Blasius and Friedrichs 2012), even small amounts of undetected fraudulent data can

severely bias survey estimates, particularly in multivariate analyses (Schräpler and Wagner 2005; Landrock 2017; DeMatteis et al. 2020).

Accordingly, the ongoing improvement of strategies for detecting falsified interviews is crucial for optimizing and ensuring data quality. Statistical detection approaches offer an effective and cost-efficient means of complementing commonly used nonstatistical detection strategies (e.g., monitoring and re-interview procedures), by making those actions more focused on suspicious interviewers. Correspondingly, an increasing number of statistical detection methods (e.g., cluster analysis) and falsification indicators (e.g., interview duration) have been developed to identify potentially fraudulent interviewer behavior (Stokes and Jones 1989; Hood and Bushery 1997; Murphy et al. 2004; Li et al. 2011; Birnbaum 2012; Bredl, Winker, and Kötschau 2012; Blasius and Thiessen 2013; Slomczynski, Powalko, and Krauze 2017; Cohen and Warner 2020).

The multitude of proposed statistical methods, however, makes it difficult to identify the method(s) best suited for detecting falsifications. Empirical evaluations and comparisons of identification methods and falsification indicators using real-world data are rare as most studies rely on experimental data (Menold et al. 2013; Storfinger and Winker 2013) or small datasets with few falsified interviews (Bredl, Winker, and Kötschau 2012). Moreover, studies have mainly focused on evaluating only one method or few falsification indicators. Using survey data including around 600 verified falsifications for person-level and household-level interviews, we address the challenges practitioners face when deciding on an appropriate detection strategy by empirically examining and comparing the performance of different statistical detection methods and falsification indicators. First, we test different multivariate detection strategies, including cluster analysis under different clustering algorithms, as well as a newly-developed detection method we term the *meta-indicator*. Using different accuracy measures, we assess the performance of these detection tools. Second, we introduce some new falsification indicators, which are shown to be useful for the data used. Third, we compare the explanatory power of single indicators and test their directional assumptions pointing to suspicious interviewer behavior.

2.2 Detecting Falsifiers: Previous Research

2.2.1 Interviewer Falsification in Practice

There are various forms of interviewer falsification. The most blatant is the fabrication of entire interviews. A related form is the partial falsification of interviews.

Further forms of falsification include interviewers deviating from prescribed selection rules, interviewing any available person instead of the—maybe unwilling—target respondent, misclassifying non-cooperative target persons as ineligible cases, or deviating from the intended interview mode (AAPOR 2003; DeMatteis et al. 2020). Additionally, the intentional miscoding of a given answer to filter questions (Eckman et al. 2014; Kosyakova, Skopek, and Eckman 2015), in order to shorten the interview, is considered falsification.

The application of detection methods to identify falsifiers is an essential part of the quality control process. Traditionally, survey organizations use a wide range of non-statistical methods as part of their control routines, for example, validation of survey data with administrative data, interview monitoring, and re-interview routines (Hauck 1969; Koch 1995; Jesske 2013). Newer approaches use GPS data to verify interviewer travel routes, digital capture tools to collect screenshots or photos of the interview location (Finn and Ranchhod 2015; Thissen and Myers 2016; Wagner, Olson, and Edgar 2017), or rapid feedback systems to improve monitoring (Edwards, Sun, and Hubbard 2020). Nevertheless, these procedures have some limitations. For example, validation with administrative data is seldom possible and monitoring face-to-face interviews often requires respondent consent to record the interview. Re-interview methods are costly and can lead to erroneous suspicion against honest interviewers if respondents misremember the encounter (DeMatteis et al. 2020).

2.2.2 Statistical Methods for Detecting Interviewer Falsification

Statistical detection methods are increasingly being used to detect potential interviewer falsification, capitalizing on the notion that falsifiers tend to produce anomalous patterns in the survey data. Such methods aid in flagging suspicious interviewers, enabling more targeted monitoring and cost-efficient use of re-interviewing. Although the methods often share similar underlying assumptions, they differ in their concrete implementation and can be divided into two—sometimes overlapping—approaches: (1) data-driven approaches, focusing on conspicuous patterns in the data, and (2) behavior-oriented approaches, focusing on specific data patterns corresponding to assumptions regarding falsification behavior.

Data-driven approaches include outlier analysis, statistical modelling, and duplicate analysis. *Outlier analysis* compares outcomes of individual interviewers with the average outcome in the survey data using distance measures, or identifies outlying interviewers based on unusual or rare response patterns and response combinations (Murphy et al. 2004; Porras and English 2004). *Statistical modelling* relies on characteristics of interviewers (e.g., tenure

or individual response rates) and parameters from previous interviews or waves (e.g., response likelihood) to model the falsification likelihood for an interview (Biemer and Stokes 1989; Li et al. 2011). More recently, supervised machine-learning algorithms (Birnbaum 2012; Weinauer 2019) and multilevel models (Sharma and Elliott 2020) have been utilized to classify possible falsifiers. Duplicate analysis flags identical response patterns occurring in multiple interviews (Slomczynski, Powalko, and Krauze 2017), "near-duplicates," that is, data with an unusually high correspondence of identical response values (Koczela et al. 2015; Kuriakose and Robbins 2016), or duplicate response patterns across same-scaled item batteries (Blasius and Thiessen 2013), and is additionally suitable for identifying fraud by supervisors or other higher administrative-level staff.

The behavior-orientated approaches—which are of primary interest for our empirical investigation—focus on systematic differences in response behavior between real and falsified interviews. These differences are measured by *falsification indicators* (including, for example, the fraction of acquiescent responding, extreme responding, or item nonresponse), which rely on assumptions regarding the rational behavior of falsifiers. While falsification indicators can be analyzed separately, they are often analyzed jointly using multivariate methods to increase the reliability of the detection results. Bredl, Winker, and Kötschau (2012) used *cluster analysis* to divide suspicious and unsuspicious interviewers into subgroups based on a selection of falsification indicators (also see Winker et al. 2013; de Haas and Winker 2016; Bergmann, Schuller, and Malter 2019). Compared to the aforementioned data-driven methods, falsification indicators and cluster analysis can be applied to every survey regardless of the topic or population. It does not require prior knowledge on variables prone to outliers and unlikely response combinations, or the falsification likelihood and actual falsification status. Nevertheless, given the variety of clustering algorithms to choose from, it is unclear which are most suitable for identifying falsifiers in practice. Interpreting the results is not always straightforward since the optimal number of clusters is usually unknown: a two-cluster solution (suspicious versus nonsuspicious interviewers) is prone to falsely suspecting many interviewers, whereas allowing for more clusters may lead to ambiguous interviewer groups.

2.2.3 Falsification Indicators

Falsification indicators aim to identify patterns produced by fraudulent interviewer behavior. Hence, they are rooted in the idea of the rational behavior of falsifiers who try to maximize their monetary benefit and minimize their time expenditure and effort, while trying to remain undetected (Menold et al. 2013). The majority of falsification indicators are

analogous to data quality indicators used to study suboptimal respondent behaviors (e.g., straightlining, primacy/recency effects), but the difference is that each respondent-level outcome is aggregated to the interviewer-level to indicate suspicious behavior attributable to the interviewer. Various indicators have been successfully used in quality control processes (Stokes and Jones 1989; Bushery et al. 1999; Turner et al. 2002) and tested on data with known falsifications (Schräpler and Wagner 2005; Bredl, Winker, and Kötschau 2012; de Haas and Winker 2016). In the following paragraphs, we present the fabrication indicators used in this paper.

For example, timestamps are used to identify interviewers with suspiciously short interviews (Bushery et al. 1999; Li et al. 2011) and a high proportion of missing telephone numbers could indicate a falsifier's effort to prevent the survey organization from re-contacting the intended respondent (Stokes and Jones 1989). Further indicators focus on answers given to specific types of survey questions (e.g., scales, filter questions). In general, falsifiers tend to produce lower response variance within- and between interviews compared to honest interviewers (Schäfer et al. 2004; Menold et al. 2013). This is driven by a variety of strategies or behaviors. For instance, falsifiers rely on their preconceived opinions or group stereotypes to provide plausible answers for a particular respondent (e.g., student, homemaker, migrant) might provide during an interview (Reuband 1990). Falsifiers also have a tendency for choosing answers in the middle of ordinal response scales rather than extreme values to avoid suspicious inconsistencies (Porras and English 2004; Storfinger and Winker 2013). They tend to avoid item nonresponse by providing answers to all closed-ended questions (Bredl, Winker, and Kötschau 2012). To reduce implausible answer combinations, which could raise suspicion, falsifiers rarely show acquiescent response behavior i.e., the tendency to agree or answer "yes" to opinion items. To decrease their effort, falsifiers often choose answers which trigger fewer follow-up questions due to filtering (Hood and Bushery 1997; Eckman et al. 2014). Altogether, these behaviors lead to reduced variation in the data.

Furthermore, real respondents hear the questions, whereas falsifiers read and answer the questions as in a self-administered mode, which may lead to different primacy (choosing the first options of answer lists) and recency effects (choosing the last options of answer lists) (Menold et al. 2013). Respondents also show a higher rounding tendency in open numeric questions (e.g., income, working hours) compared to falsifiers (Menold et al. 2013). Additionally, falsifiers tend to avoid answering open-ended items leading to higher rates of nonresponse and less frequent selection of the "Other, specify"-option for semi-open-ended

questions, which is contrary to other question types (Bredl, Winker, and Kötschau 2012). Benford's Law is another example which states that the first digit of naturally occurring numbers follows a logarithmic distribution (Benford 1938; Hill 1999). It is often utilized to evaluate the veracity of numeric data since falsifiers are less likely to reproduce the Benford distribution (Schäfer et al. 2004).

New falsification indicators: In addition to the indicators from previous research described above, we propose four new falsification indicators: the rate of provided email addresses, a measure of the relative interview duration, the rate of respondent consent to link their survey data to administrative data, and the interviewers' evaluation of their interviews. The rate of provided email addresses follows the same logic as the paradata indicator on telephone numbers: falsifiers tend to produce more missing email addresses to prevent the verification of the interview. Relative interview duration (average interview duration per question) is expected to be lower for falsifiers as it reflects different types of time-saving behavior (e.g., avoidance of triggering follow-up questions to filter items, not reading/repeating questions out loud). Falsifiers are expected to produce higher linkage consent rates compared to real interviewers because granting linkage consent is viewed as a desirable research outcome and is an indication of cooperative response behavior that is unlikely to raise suspicion. Finally, given that falsifiers aim to produce inconspicuous and generally cooperative interviews in order to avoid detection, we expect falsifiers' post-interview evaluation of the interview (i.e., the interviewer evaluation) to be very positive compared to those of honest interviewers.

2.3 Data, Methods, and Evaluation Strategy

2.3.1 Data

We utilize data from the IAB-BAMF-SOEP Survey of Refugees in Germany (Brücker, Rother, and Schupp 2017), including verified falsifications (version SOEP.v33) (Kosyakova et al. 2019).[1] This is an annually conducted longitudinal household survey, launched in 2016. The target population includes refugees and asylum-seekers who arrived between 2013 and 2016, and their adult household members.[2] The sample was drawn from the German Central Register of Foreigners (Ausländerzentralregister) (Kroh et al. 2017). We

[1] For the analyses, we are using version SOEP.v33. All falsifications were excluded from the official data release (v34).

[2] Upon their arrival, refugees were distributed across Germany through a national dispersal allocation scheme (Königstein Key; Grote 2018).

use data of the first wave with a sample of 3,554 responding households and 4,816 respondents.[3]

The household-level response rate (Response Rate 2; AAPOR 2016) was 48.7 percent (Kroh et al. 2017). The survey included two types of questionnaires: person interviews, ideally conducted with every adult household member, and a shorter household interview with the anchor-person about the household's situation. A staff of 98 trained interviewers, who worked in specific regional areas, completed between 1 and 289 (mean≈49, median≈32) computer-assisted personal interviews (CAPI). Interviewing started at the end of June 2016 and was completed in December 2016 (Brücker, Rother, and Schupp 2017). Since the sample included refugees from various home countries—in part without German language proficiency—questionnaires were provided in seven languages (Arabic, English, Farsi/Dari, German, Kurmanji, Pashtu, and Urdu). Additionally, the questionnaires were complemented with audio files containing recordings of the questions and access to an interpreter hotline (Jacobsen 2018). Person-level questionnaires included principal topics on: migration history, education biographies, language acquisition and employment, life satisfaction, health, and attitudes (Brücker, Rother, and Schupp 2017).

Routine quality control checks by the survey organization detected a first suspicious interviewer, who was confirmed as a falsifier after a subsequent review of her wave 1 respondents (IAB 2017). We refer to this interviewer as 'F1'. F1 accounted for 289 person interviews and 217 household interviews, which must be considered as complete falsifications. Further investigations carried out by the survey organization and the IAB (including various statistical methods, re-contacting of respondents, questioning of supervisors and interviewers) confirmed two additional falsifiers responsible for a total of 62 person and 47 household interviews (DIW 2019; Kosyakova et al. 2019). These interviewers did not fabricate all of their assigned interviews. According to the survey organization, only in the latter half of the field period did these interviewers start fabricating complete interviews. The exact number of these falsified interviews could not be determined and is unknown. We refer to these interviewers as 'F2' and 'F3'. Consistent with the AAPOR definition of interviewer falsification (AAPOR 2003), we refer to interviewers F1, F2, and F3 as falsifiers and the data produced by these interviewers as falsifications. **Table 2.1**

[3] All analyses are based on the raw field data; therefore, no weights are used and no design effects are considered.

contains detailed information about the number of interviews (overall and for each falsifier) and response rates.

Table 2.1: Response outcomes for falsifiers and nonfalsifiers.

		Person interviews		Household interviews	
	Response rate	*N*	*Pct.*	*N*	*Pct.*
Falsifier					
F1	85.8%	289	6.0%	218	6.1%
F2	60.7%	46	1.0%	34	1.0%
F3	41.9%	16	0.3%	13	0.4%
Total for falsifiers	77.7%	351	7.3%	265	7.5%
Total for nonfalsifiers	48.4%	4,465	92.7%	3,289	92.5%
Total	48.7%	4,816	100%	3,554	100%

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Note: All response rates are calculated at the household level according to Response Rate 2 (AAPOR 2016).

2.3.2 Statistical Detection Methods

2.3.2.1 Cluster Analysis

Starting with cluster analysis (see, e.g., Kaufman and Rousseeuw 1990), the basic idea is to classify interviewers into smaller homogeneous subgroups that distinguish suspicious and nonsuspicious interviewers using grouping characteristics (i.e., the falsification indicators). To evaluate the distances between interviewers, we implement the commonly-used Euclidean distance:

$$d_{j,l} = \left[\sum_{k=1}^{n} (x_{jk} - x_{lk})^2\right]^{\frac{1}{2}} \tag{2.1}$$

with $d_{j,l}$ denoting the distance between a pair of interviewers *j* and *l*, and x_{jk} and x_{lk} denoting the values for the k^{th} (*= 1, 2, ..., n*) falsification indicator for the respective interviewer pair. Based on the resulting distance matrix, classification can take place using different clustering algorithms, which greatly differ with regard to the group formation. We compare two hierarchical-agglomerative algorithms: Ward's Linkage (Ward 1963) and Single-Linkage (McQuitty 1957).

In the context of falsification identification, Ward's Linkage has been successfully applied in previous research (Menold et al. 2013; Storfinger and Winker 2013). Ward's Linkage combines clusters such that the sum of squared errors is minimized. This allows varying cluster sizes, which enables a meaningful cluster solution even for—as we assume–

–a small group of potential falsifiers. In contrast to Bredl, Winker, and Kötschau (2012) and Menold et al. (2013), we allow for solutions with more than two clusters. The rationale for permitting solutions with more than two groups is that the falsification indicators could also capture different interviewing styles and behaviors (e.g., differences between experienced and inexperienced interviewers) that may not be fraudulent in nature. Hence, greater separation of these interviewing styles is enabled, minimizing the risk of unwarranted suspicions against honest interviewers that might occur in a forced two-group solution. However, this approach impedes direct identification of the suspicious group and requires further inspection of each group based on a comparison of their indicator values. In contrast to prior studies, we additionally apply Single-Linkage to address the problem of identifying suspicious interviewers. Single-Linkage[4] is particularly useful for identifying outliers, since it combines clusters that have the closest neighboring objects (Kaufman and Rousseeuw 1990).

To determine the optimal cluster solution for both Ward's Linkage and Single-Linkage, we visually inspect dendrograms and further consider the formal criteria of the Calinski-Harabasz index and the Duda-Hart index (Caliński and Harabasz 1974; Duda and Hart 1973). Optimal cluster solutions are indicated by large values of the Calinski-Harabasz pseudo F-index and Duda-Hart Je(2)/Je(1)-index as well as small values of the Duda-Hart pseudo T-squared. First, we derive from the dendrogram, which cluster solutions are plausible according to the shown dissimilarity measure. Second, we compare the values of the formal criteria for these cluster solutions.

2.3.2.2 Meta-Indicator Approach

As described above, the application of cluster analysis requires several decisions, which may affect the results. We therefore propose a simpler multivariate tool, which we refer to as the *meta-indicator approach*. Basically, it summarizes the interviewer-level values of all indicators into a single (meta-)indicator value per interviewer: First, to obtain comparable and continuous values for each of the indicators, each interviewer-level indicator value is standardized across all interviewers using the following equation:

[4] Note that Single-Linkage is prone to chaining effects, that is, an interviewer might be added to a cluster because of a high similarity with a single interviewer within the cluster, even though the added interviewer shows high dissimilarity with the other interviewers in the cluster (Everitt and Rabe-Hesketh 2006). However, in this particular application such effects are desirable, since we assume falsifiers to be strong outliers whereas honest interviewers may share some similarities but are still different from each other in other ways.

$$z_{i,k} = \frac{x_{i,k} - \bar{x}_k}{S_k} \tag{2.2}$$

with $z_{i,k}$ denoting the k^{th} standardized indicator value for interviewer i and $x_{i,k}$ denoting the unstandardized indicator value. Further, $\bar{x}_k$ denotes the mean value of indicator k and S_k the corresponding standard deviation. Second, all standardized indicator values are summed up for each interviewer. Note that indicator values are coded such that positive values indicate the assumed suspicious direction. Therefore, extreme positive values of the meta-indicator signal potential falsification behavior of interviewers. We consider three arbitrary thresholds, which flag interviewers as "suspicious" if their meta-indicator value exceeds it to demonstrate the sensitivity of the method under more inclusive and restrictive identification criteria. The first threshold is defined as 2 standard deviations (SD) above the mean, which is a commonly used "rule-of-thumb" for outlier detection, especially in relatively small samples.[5] The second and third thresholds are 1.75 and 2.25 SDs above the mean, which represent more liberal and conservative identification criteria, respectively, compared to the 2 SD rule. In practice, the actual threshold can be adapted flexibly, even after inspection of the overall distribution, depending on the user's preference for a more inclusive or restrictive controlling process.

2.3.2.3 Falsification Indicators

In total, we consider 32 falsification indicators: 21 based on person-level data (interview data, paradata, and interviewer's evaluation of the person interview) and 11 on household-level data (interview data and paradata). All indicators are standardized according to equation (2.2) and coded such that positive values indicate the suspicious direction; for example, interviewers with a lower share of item nonresponse—the assumed direction of falsification for closed-ended items—receive a larger positive indicator value compared to interviewers with a higher share of item nonresponse. Further, the interview-level values of a falsification indicator were aggregated to the interviewer-level by computing the mean indicator value across all interviews of an interviewer. **Table 2.2** provides a summary of the used indicators, their assumed direction for falsifiers, and a description of their construction. Further information about the used indicators is shown in **Table 2.3**.

[5] As further evaluation criteria, we considered the Interquartile Range (IQR), Tukey's Method, the Median Absolute Deviation (MAD) Method, the Z-Score, and the Modified Z-Score. We did not find strong differences in the results between these methods and the SD-Method. Hence, we only present results for the SD-Method.

Table 2.2: Overview of used falsification indicators and underlying assumptions.

Indicator	Description	Assumed direction of falsifiers	References
Acquiescent responding	Fraction of positive connotation ("Agree/ Strongly Agree") independent of content	Lower fraction of positive connotation independent of question content for falsifiers	Menold et al. (2013)
Benford's Law	Decreasing distribution of leading digit for numeric quantities	Poor fit of Benford's distribution to leading digits for falsifiers	Swanson, Cho, and Eltinge (2003)
Email	Fraction of email address provision	Lower fraction of provided email addresses for falsifiers	NEW
Extreme responses	Fraction of extreme responses to rating scales	Lower fraction of extreme responses to rating scales for falsifiers	Schäfer et al. (2005)
Filter questions	Fraction of responses which lead to follow-up questions	Lower fraction of responses which lead to follow-up questions for falsifiers	Hood and Bushery (1997)
Interview duration	Duration of completed interviews	Shorter duration of completed interviews for falsifiers	Hood and Bushery (1997)
Interview duration, relative	Duration of completed interviews relative to the triggered questions	Shorter duration of completed interviews relative to the triggered questions for falsifiers	NEW
Interviewer evaluation	Interviewer's evaluation of the interview situation	Higher fraction of very positive evaluation of the interview situation for falsifiers	NEW
Item nonresponse	Item nonresponse rate within an interviewer's workload of closed-ended questions	Lower item nonresponse rate for falsifiers	Schäfer et al. (2005)

Table 2.2 (continued)

Indicator	Description	Assumed direction of falsifiers	References
Middle category responses	Fraction of middle responses to rating scales	Higher fraction of middle responses to rating scales for falsifiers	Schäfer et al. (2005)
Non-Differentiation	Standard deviation within an item scale	Lower standard deviation within an item scale for falsifiers	Reuband (1990)
Primacy effects	Fraction of choosing the first two categories in non-ordered answer option lists	Higher fraction of choosing the first two categories in non-ordered answer option lists for falsifiers	Menold et al. (2013)
Recency effects	Fraction of choosing the last two categories in non-ordered answer option lists	Lower fraction of choosing the last two categories in non-ordered answer option lists for falsifiers	Menold et al. (2013)
Record linkage consent	Fraction of consent to record linkage	Higher fraction of consent to record linkage for falsifiers	NEW
Rounding	Fraction of rounding numbers in numerical open-ended questions	Lower fraction of rounded numbers in numerical open-ended questions for falsifiers	Menold et al. (2013)
Semi-Open responses	Fraction of responses to “other” in semi-open-ended question	Lower fraction of responses to “other” in semi-open-ended question for falsifiers	Hood and Bushery (1997)
Stereotyping	Strength of stereotypical response to attitudinal items	Higher strength of stereotypical response to attitudinal items for falsifiers	Reuband (1990)
Telephone number	Fraction of telephone number provision	Lower fraction of provided telephone numbers for falsifiers	Stokes and Jones (1989)
Response variance	Standard deviation of responses between interviews	Lower standard deviation of responses between interviews for falsifiers	Porras and English (2005)

Source: The table was adapted from Kosyakova et al. (2019).

The actual observed indicator values, shown separately for falsifiers and honest interviewers, are given in Appendix, **Table A 2.1**. Note that we do not account for area-level effects as this could hinder the identification of falsifiers collaborating in certain regions, as seen in Bergmann, Schuller, and Malter (2019). Likewise, we do not account for nonindependence within households, as all analyses are conducted at the interviewer level.

Table 2.3: Overview of used falsification indicators and labels.

Indicator	Data source	Label
Acquiescent responding	Person interviews	ACQ_P
Benford's Law	Person interviews	BFL_P
	Household interviews	BFL_H
Email	Household-level paradata	MAIL_H
Extreme responses	Person interviews	ERS_P*
	Household interviews	ERS_H
Filter questions	Person interviews	FILT_P
	Household interviews	FILT_H
Interview duration	Person interviews	DUR_P
	Household interviews	DUR_H
Interview duration, relative	Person interviews	RDUR_P
	Household interviews	RDUR_H
Interviewer evaluation	Person-level evaluation	EVAL_P
Item nonresponse	Person interviews	INR_P
	Household interviews	INR_H
Middle category responses	Person interviews	MRS_P*
	Household interviews	MRS_H
Non-Differentiation	Person interviews	ND_P
Primacy effects	Person interviews	PRIM_P
Recency effects	Person interviews	RECE_P
Record linkage consent	Person-level paradata	RLC_P
Rounding	Person interviews	ROUND_P
	Household interviews	ROUND_H
Semi-open responses	Person interviews	SOR_P
Stereotyping	Person interviews	STEREO_P
Telephone number	Household-level paradata	TEL_H
Response variance	Person interviews	VAR_P
	Household interviews	VAR_H

*Due to large differences in the number of scale categories between item batteries, three different indicators were created. Large scales with 10 or 11 answer categories (h), medium size scales with 7 categories (m), and small scales with 4 or 5 categories (l).

2.3.3 Evaluation Strategy

2.3.3.1 Comparison of Multivariate Detection Methods

To evaluate the performance of the different detection methods in identifying the falsifiers, we consider several quality measures: false-positive rate, false-negative rate, accuracy, error rate, and Cohen's kappa (Cohen 1960). The false-positive rate relates the number of falsely detected interviewers to the overall number of honest interviewers, whereas the false-negative rate measures the share of overlooked falsifiers. The accuracy captures the relationship between the false-negative and false-positive rates, whereas the error rate equals one minus the accuracy. Cohen's kappa adjusts the accuracy by accounting for the possibility of true predictions by chance. Corresponding formulas can be found in **Table 2.4**. We test the robustness of the cluster analyses and the meta-indicator results by applying a simple leave-one-out procedure, repeating the respective analyses excluding one indicator at a time.

Table 2.4: Overview of formulas for performance measures.

Performance measure	Formula	
False-positive rate (FP_{rate})	$FP/(TN + FP)$	(2.3)
False-negative rate (FN_{rate})	$FN/(TP + FN)$	(2.4)
Accuracy (A)	$(TP + TN)/(TP + TN + FP + FN)$	(2.5)
Error rate (E_{rate})	$1 - A$	(2.6)
Cohen's kappa (κ)	$\left(\Pr(a)_{obs} - \Pr(b)_{exp}\right)/\left(1 - \Pr(b)_{exp}\right)$	(2.7)

Note: FP = false-positive cases, *FN* = false-negative cases, *TP* = true-positive cases, *TN* = true-negative cases, $\Pr(a)_{obs}$ = observed agreement, $\Pr(b)_{exp}$ = expected agreement.

2.3.3.2 Comparison of Single Indicators

We use discriminant analysis to evaluate the relative importance of the single indicators for identifying falsifiers and to test the validity of the directional assumptions of the indicators (Bredl, Winker, and Kötschau 2012). Linear discriminant analysis is not used as an instrument to detect falsifiers but enables assessment of the goodness-of-falsification indicators in distinguishing falsifiers from the nonfalsifiers if falsifiers are known. Using a linear combination of the continuous standardized indicator variables z_k (k = 1, 2, …, n) as independent discriminating variables, we seek the canonical discriminant function that provides the maximal separation between the falsifier and nonfalsifier groups (Klecka 1980; McLachlan 2004). The discriminant function D takes the following form:

$$D = b_0 + b_1 z_1 + b_2 z_2 + \cdots + b_n z_n = b_0 + \sum_{k=1}^{n} b_k z_k \tag{2.8}$$

Due to the binary falsification status, only one discriminant function is determined. Maximal discrimination is achieved by determining the discriminant constant b_0 and the discriminant coefficients b_k such that the group specific $D_g = 1/I_g \sum_{i=1}^{I_g} D_{ig}$—with $g = 1$ for falsifiers, $g = 0$ for honest interviewers, and I_g the number of interviewers per group—are as different as possible (Klecka 1980; Bredl, Winker, and Kötschau 2012). Put differently, the aim is to maximize the between-group variance but minimize the within-group variance. The absolute sizes of the standardized discriminant coefficients identify the most important indicators for the distinction between falsifiers and nonfalsifiers. Since some of the indicators are highly correlated, we consider the canonical structure coefficients, which adjust for possible multicollinearity between indicators. Note that a comparison of the standardized coefficients and the structure coefficients deepens the understanding of the underlying relationships between the indicators and allows for assessing the importance of single indicators.

2.4 Results

2.4.1 Cluster analysis

2.4.1.1 Ward's Linkage

Figure 2.1 shows the full dendrogram according to the dissimilarity between the groups for Ward's Linkage. The dotted lines indicate plausible cluster solutions. Accordingly, a 2-, 3-, or 4-cluster solution seems plausible. Looking at the two formal indices (**Table 2.5**), we find contrary recommendations: Calinski-Harabasz suggests a 2-cluster solution whereas Duda-Hart suggests a 4-cluster solution. Looking at the number of interviewers per cluster, the 4-cluster solution with 26, 42, 25, and 5 interviewers rather than 68 and 30 interviews seems more plausible since we assume falsifiers to be the minority among interviewers. The dendrogram for the 4-cluster solution is presented in **Figure 2.2**.

To identify the suspicious group, inspection of the mean indicator values for each cluster is necessary. The results in **Figure 2.3** imply that Cluster 1 mostly includes interviewers with negative indicator values, while Cluster 2 mainly includes interviewers with indicator values around zero. Both groups are therefore associated with unsuspicious interviewer behavior. Cluster 3 includes interviewers with mixed indicator values, having a slight tendency for suspicious values. In practice, one might consider randomly sampling some interviews for re-interviews from this group of interviewers. More severe is Cluster 4, which includes interviewers with highly suspicious indicator values for most indicators. This group clearly

stands out as being suspicious, compared to the other groups, and would be a prime target for further investigation via re-interviews.

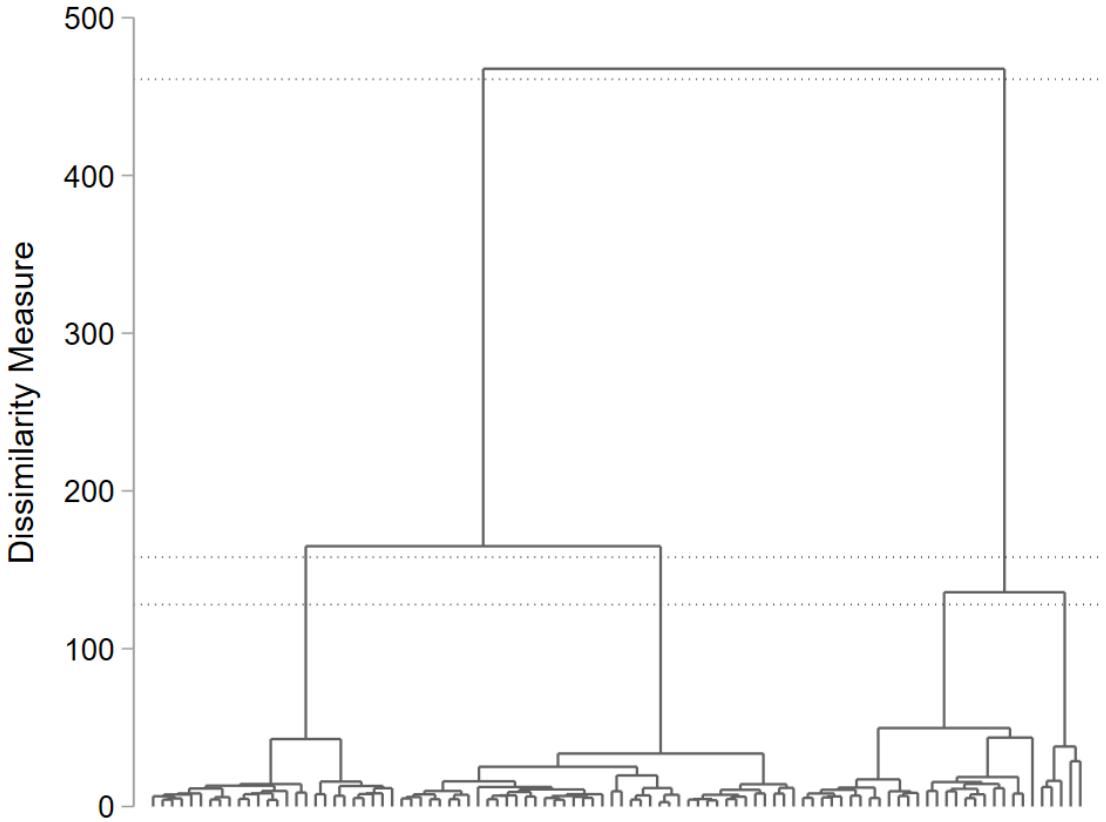

Figure 2.1: Full dendrogram for Ward's Linkage cluster analysis.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Note: Dotted lines indicate plausible cluster solutions.

Subsequent inspection revealed that the outlying cluster includes all three falsifiers (F1, F2 and F3) but also two further interviewers (I62 and I70). However, since these two interviewers conducted a very small number of (i.e., less than five) interviews, the indicator values could reflect respondents' answering behavior rather than deviant interviewing. Controls conducted by the survey organization did not confirm any suspicious behavior for these two interviewers.

Table 2.6 shows the false-positive rates, false-negative rates as well as the accuracy, error rate, and kappa statistic for the different detection methods. For the 4-cluster solution, Ward's Linkage (first column) results in a false-positive rate of 2.1 percent and a false-negative rate of 0 percent. Because of the low false-positive and false-negative rates, accuracy is very high (98.0 percent) and the error rate very low (2.0 percent), also resulting in a very good kappa statistic (0.74).

Table 2.5: Calinski-Harabasz and Duda-Hart Index for Ward's Linkage and Single-Linkage.

	Calinski-Harabasz	Duda-Hart	
Number of clusters	Pseudo F-index	Je(2)/Je(1) index	Pseudo T-squared
Ward's Linkage			
1	.	0.515	90.27
2	90.27	0.587	46.35
3	64.21	0.516	26.24
4	**79.93**	**0.777**	**6.60**
5	66.45	0.477	10.96
6	62.08	0.691	10.74
7	56.51	0.506	2.93
8	56.44	0.839	7.69
9	52.92	0.000	.
10	52.91	0.828	5.84
Single-Linkage			
1	.	0.959	4.07
2	4.07	0.886	12.19
3	8.37	0.833	18.81
4	12.89	0.788	25.02
5	**18.40**	**0.365**	**1.74**
6	14.90	0.000	.
7	12.44	0.962	3.59
8	11.48	0.968	2.91
9	10.62	0.969	2.85
10	9.96	0.987	1.16

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Note: The chosen cluster solution used for the evaluation is indicated in boldface.

2.4.1.2 Single-Linkage

Figure 2.4 shows the full dendrogram according to the dissimilarity measurement for Single-Linkage. As for Ward's Linkage, dotted lines indicate plausible cluster solutions, ranging between three to seven clusters. The figure further indicates that most interviewers (in total 92) share a high similarity, whereas six interviewers appear as outliers and therefore as suspicious.

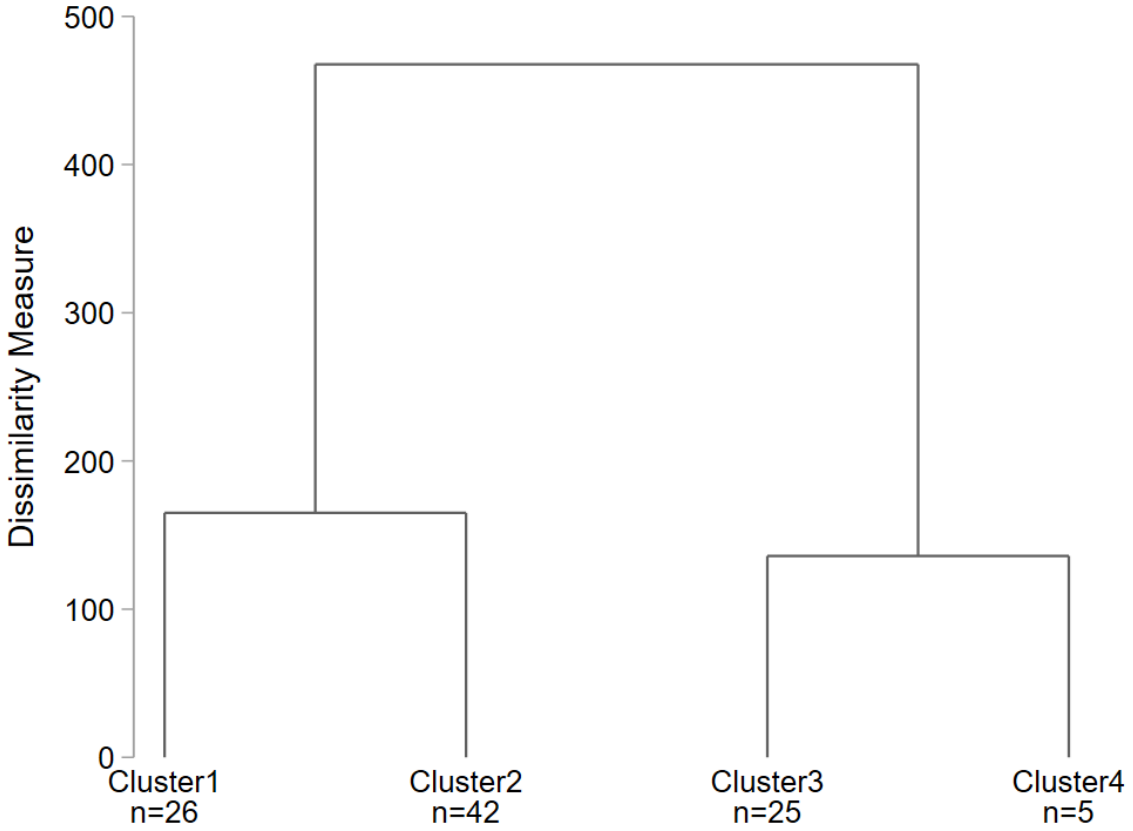

Figure 2.2: Dendrogram for Ward's Linkage cluster analysis with 4-cluster.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Considering the formal indices (**Table 2.6**), we find great support for a 5-cluster solution according to the Calinski-Harabasz index. The recommendation of the Duda-Hart index is ambiguous: the Pseudo T-squared of the index also supports the 5-cluster solution, whereas the Je(2)/Je(1)-index supports a 7-cluster solution. The decision between the two solutions is arbitrary as both identify the same outliers, with three outliers grouped together in the 5-cluster solution and placed in separate clusters in the 7-cluster solution.

Table 2.6: Performance measures of interviewer falsification detection methods.

	Ward's Linkage	Single-Linkage	Meta-indicator thresholds		
			1.75 SDs	2.00 SDs	2.25 SDs
False-positive rate	2.1%	3.2%	2.1%	2.1 %	2.1 %
False-negative rate	0.0%	0.0%	0.0%	0.0 %	40.0 %
Accuracy rate	98.0%	97.0%	98.0%	98.0%	95.9%
Error rate	2.0%	3.1%	2.0%	2.0%	4.1%
Kappa statistic	0.74	0.65	0.74	0.74	0.31

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

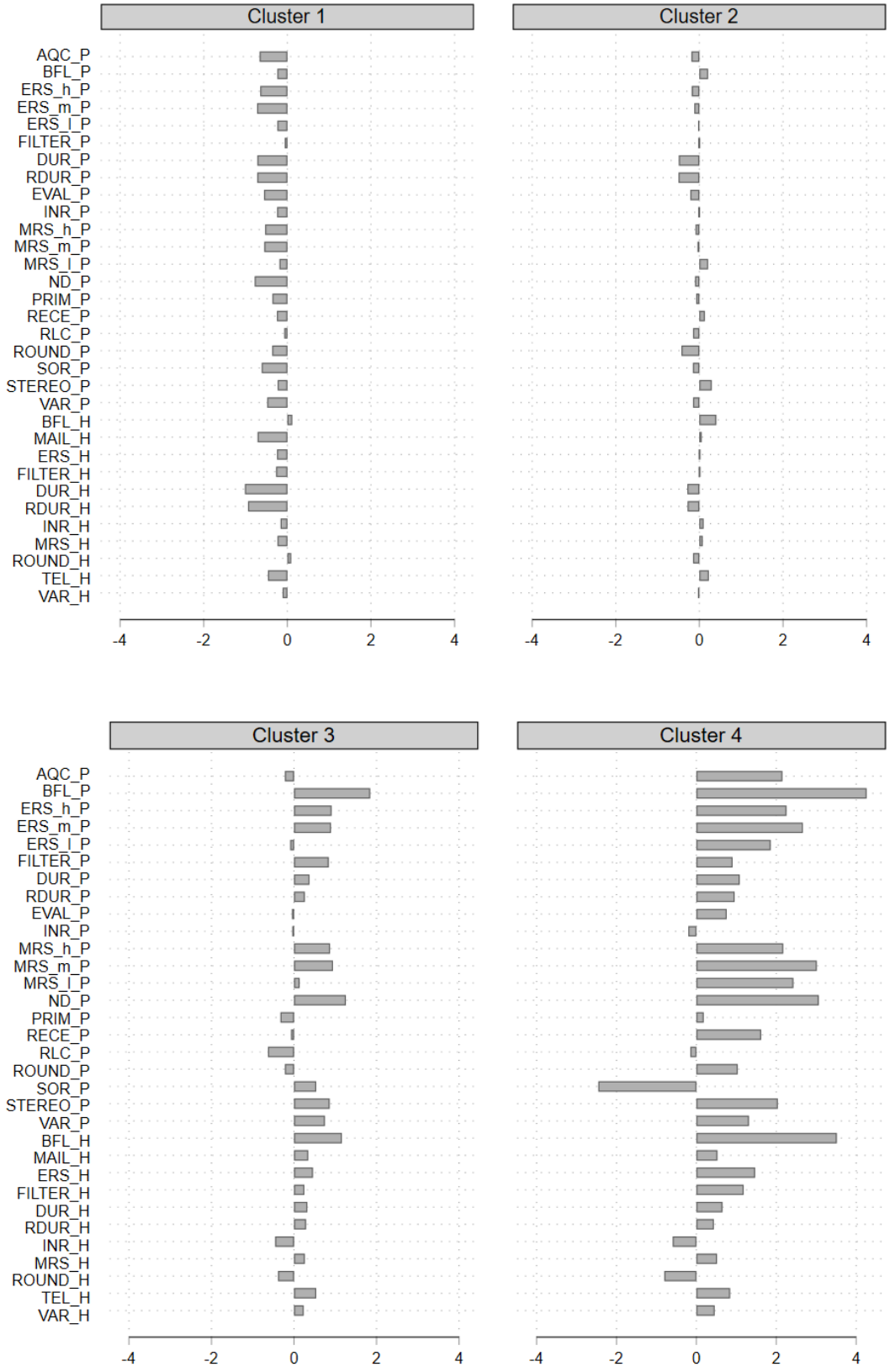

Figure 2.3: Mean indicator values per cluster for Ward's Linkage.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

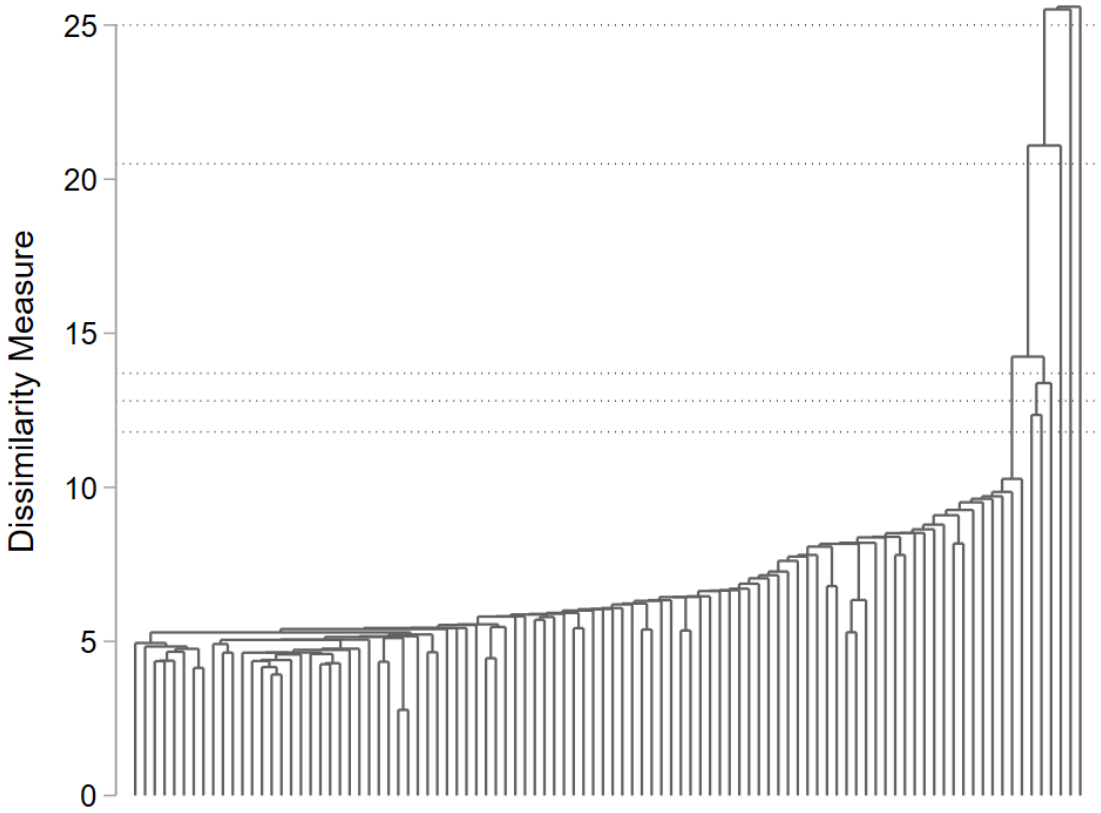

Figure 2.4: Full dendrogram for Single-Linkage cluster analysis.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Note: Dotted lines indicate plausible cluster solutions.

The dendrogram for the 7-cluster solution (**Figure 2.5**) reveals that, similar to Ward's Linkage, all three falsifiers (F1, F2 and F3) are identified as suspicious. Three further interviewers characterized by a small number of conducted interviews (I62, I70 and I88) are falsely suspected. The falsifiers seem to be more similar than the other outlying interviewers since they would be grouped together in a 5-cluster solution. Since the number of falsely suspected interviewers is slightly higher for Single-Linkage, accuracy, error rate and kappa statistic result in a worse evaluation (**Table 2.6**, second column).

2.4.2 Meta-Indicator Approach

Following our assumptions, the meta-indicator should produce extreme positive values for suspicious interviewers relative to the honest interviewers. As **Figure 2.6** shows, five outlying interviewers (including all falsifiers and two further interviewers) lie above the predefined threshold values of 1.75 and 2 SDs above the mean. I62 and I70 are again falsely

suspected. This is also confirmed using a boxplot, which can be found in Appendix, **Figure A 2.1**.

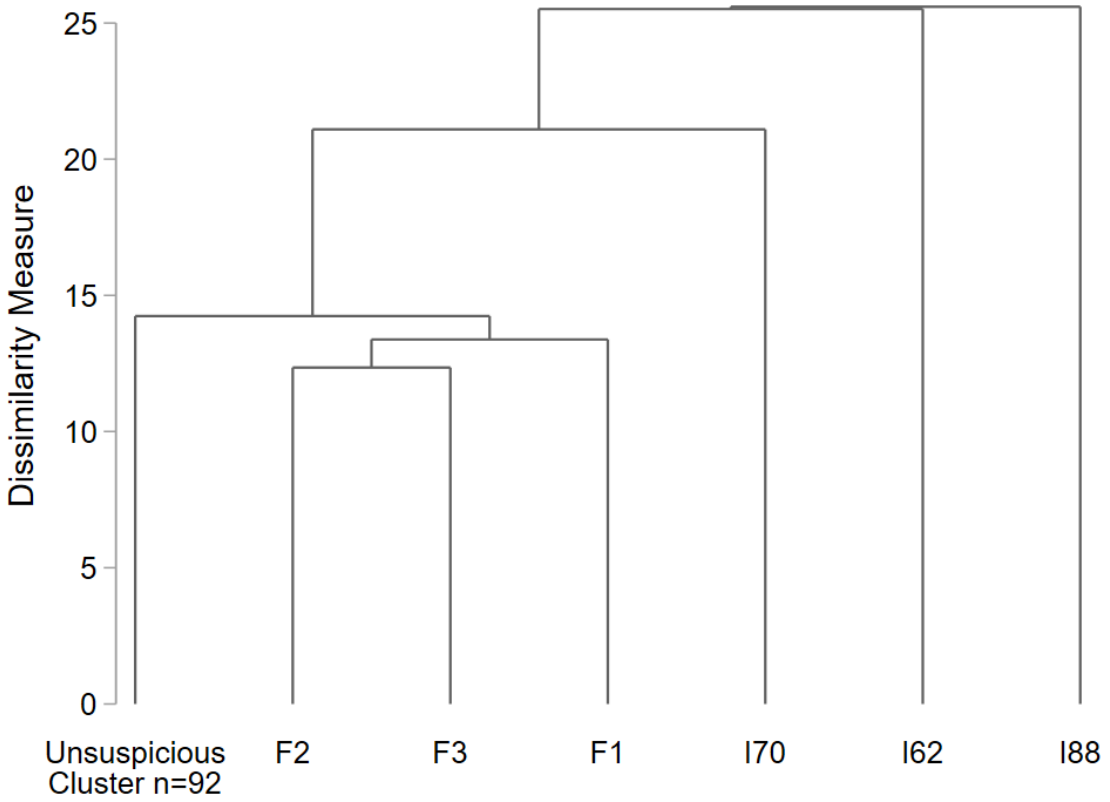

Figure 2.5: Dendrogram for Single-Linkage cluster analysis with 7-cluster solution.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Similar to Ward's Linkage, both meta-indicator thresholds (1.75 and 2 SDs) result in a false-positive rate of 2.11 percent and a false-negative rate of zero percent and therefore the same accuracy, error rate, and kappa statistic (**Table 2.6**, third and fourth columns). However, the more conservative threshold of 2.25 SDs above the mean results in poorer performance. Two falsifiers (F1 and F2) would be classified as unsuspicious, resulting in a high false-negative rate of 40 percent (**Table 2.6**, fifth column). This slightly affects the accuracy (95.9 percent) and the error rate (4.1 percent). However, the kappa statistic drops drastically from 0.74 to 0.31.

2.4.3 Sensitivity of Detection Methods by Indicator

2.4.3.1 Cluster Analysis

Repeating the analysis of the different cluster algorithms with a leave-one-out procedure for each indicator reveals very stable results. Regardless of which indicator is left out, both Single-Linkage and Ward's Linkage consistently identified all three falsifiers. All falsely

suspected interviewers are also identified as suspicious. Accordingly, the false-positive and false-negative rate and therefore also the other performance measures do not change.

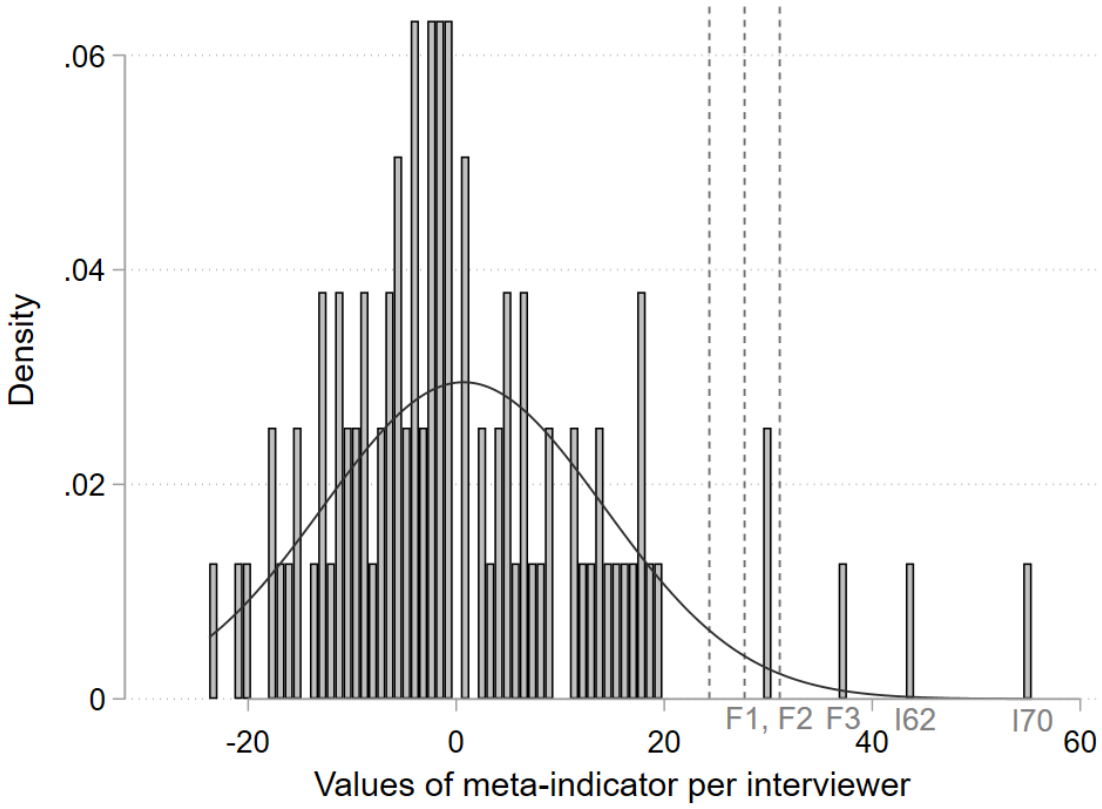

Figure 2.6: Distribution of the meta-indicator values.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Note: Dotted lines indicate the 1.75, 2.00, and 2.25 SD thresholds, respectively.

2.4.3.2 Meta-Indicator

For the meta-indicator, the results are mostly stable, depending on the selected threshold. Regardless of the indicator left out, all falsifiers are clearly identified as suspicious using the 1.75 SD threshold. Importantly, this does not increase the number of falsely suspected interviewers. The more conservative 2 SD threshold leads to a slightly worse performance. F3 is always identified as suspicious, however, F1 and F2 are not identified in all cases. Particularly, F1 is overlooked if the indicator for primacy effects (PRIM_P), interviewer evaluation (EVAL_P), or rounding tendency (ROUND_P, ROUND_H) is left out. F2 is not flagged if the indicator for semi-open responses (SOR_P), Benford's Law (BFL_P), nondifferentiation (ND_P), or middle-responding-style (MRS_h_P, MRS_m_P) is left out. This is reinforced by using the most conservative threshold of 2.25 SDs. Again, F3, I62, and I70 remain in the suspicious group regardless of the withdrawn indicator. F1 is labeled as suspicious

for only five (out of 32) versions of the reduced meta-indicator, and F2 for only nine versions of the reduced meta-indicator.

2.4.4 Comparison of Single Indicators Using Discriminant Analysis

To assess the relative importance of the single indicators, we turn to the discriminant analysis. The canonical correlation—which is equivalent to the Pearson correlation between the falsification status and the best linear combination of all indicators—is 0.757 (**Table 2.7**). Hence, the combination of indicators is highly correlated with the actual falsification status. This is also confirmed by Wilks' lambda (significant at an alpha-level of 0.000).

Table 2.7: Model-fit of the discriminant analyses.

	Canonical Correlation	Eigenvalue	Wilks' Lambda	F	df1	df2	*p*-value
Function D	0.757	1.346	0.426	2.734	32	65	0.000

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Note: P-values are based on a one-tailed significance test.

However, there are remarkable differences between the relative importance of single indicators for the used data. The resulting group-specific discriminant value D_g for the falsifier group ($g = 1$) amounts to -6.461 and 0.204 for the nonfalsifier group ($g = 0$). Accordingly, the group of falsifiers is associated with negative values on the canonical variables. This is important for the interpretation of the coefficients, given that negative coefficients indicate conformity of the directional assumptions of the indicators. **Table 2.8** presents the standardized discriminant coefficients as well as the canonical structure coefficients for all 32 indicators. The absolute magnitude of the standardized coefficients infers on the importance of the single indicators for the discrimination between falsifiers and nonfalsifiers in a joint model of all 32 indicators. The person-level interview duration indicator (DUR_P) and the newly-developed relative counterpart (RDUR_P) seem to be of utmost importance. However, due to their significant correlation (Appendix, **Table A 2.2**), the coefficient for duration is negative whereas the coefficient for relative duration is positive, since the effect of the relative duration is already captured by the duration indicator. Hence, it would probably suffice to use only one of these indicators in practice. Further, the number of triggered filter questions in person interviews (FILT_P) and the relative duration of the household interview (RDUR_H) are also crucial. All four indicators are related measures, highlighting the importance of time-related measures or measures indicating potential shortcutting for detecting falsifiers.

Table 2.8. Results of the discriminant analysis.

Indicator	Interview type	Standardized discriminant coefficients	Ranking	Canonical structure coefficients	Ranking
DUR_P	person level	–12.030	1	–0.205	6
RDUR_P	person level	10.608	2	–0.204	7
FILT_P	person level	2.630	3	–0.035	27
RDUR_H	household level	0.962	4	0.033	28
BFL_P	person level	–0.674	5	–0.155	14
MRS_h_P*	person level	–0.605	6	–0.330	2
ND_P	person level	–0.506	7	–0.327	3
ERS_m_P*	person level	0.484	8	–0.201	8
ROUND_P	person level	–0.462	9	–0.089	21
FILT_H	household level	–0.400	10	–0.080	24
ACQ_P	person level	–0.396	11	–0.161	13
MRS_H	household level	–0.392	12	–0.094	20
EVAL_P	person level	–0.384	13	–0.274	4
MRS_m_P*	person level	–0.351	14	–0.367	1
ERS_h_P*	person level	0.327	15	–0.194	10
DUR_H	household level	–0.313	16	0.003	32
PRIM_P	person level	–0.293	17	–0.179	12
VAR_H	household level	–0.282	18	–0.020	29
MRS_l_P*	person level	0.267	19	–0.049	26
RLC_P	person level	–0.241	20	–0.109	18
STEREO_P	person level	0.187	21	–0.192	11
INR_P	person level	0.162	22	–0.007	31
VAR_P	person level	0.146	23	–0.142	15
MAIL_H	household level	0.114	24	–0.077	25
ROUND_H	household level	0.091	25	0.109	17
SOR_P	person level	0.090	26	–0.084	22
ERS_l_P*	person level	–0.082	27	–0.114	16
TEL_H	household level	–0.060	28	–0.081	23
RECE_P	person level	–0.055	29	–0.197	9
BFL_H	household level	–0.041	30	0.014	30
ERS_H	household level	–0.030	31	–0.231	5
INR_H	household level	–0.013	32	0.100	19

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Note: *Due to large differences in the number of scale categories, three different indicators were created. Large scales with 10 or 11 answer categories (h), medium size scales with 7 categories (m), and small scales with 4 or 5 categories (l). New indicators are shown in italics.

We further observe that Benford's Law at the person-level (BFL_P) is central for the discrimination between falsifiers and nonfalsifiers. Another group of indicators plays an

essential role: middle-responding-style (MRS_h_P), extreme-responding-style (ERS_m_P), and non-differentiation (ND_P) in person interviews. Again, these indicators are correlated, since less extreme values automatically lead to more middle-category responses and therefore to more straightlining. Hence, ERS takes a positive value since the effect is already captured by MRS and ND. This demonstrates that item batteries serve as a crucial basis for falsification indicators. Turning to the newly-proposed indicators, we find that, in addition to the relative duration, the interviewer's evaluation (EVAL_P) of the interview serves as a valuable indicator. Although the indicator on record linkage consent (RLC_P) is inferior compared to the other new indicators, it is still useful in discriminating between falsifiers and nonfalsifiers and outperforms more than one-third of all indicators. The same is true for the number of provided email addresses (MAIL_H), which turned out to be of less importance relative to others, but still aided in the separation between the two groups.

To infer on the impact of single indicators without the influence of the other indicators, canonical structure coefficients—measuring the correlation between each indicator and the discriminant function—and their importance ranking are presented (**Table 2.8**). These coefficients allow testing the assumptions on the expected direction of the indicators (from **Table 2.2**). As the falsifier group is associated with negative function values, negative values of the canonical structure reveal that an indicator points in the assumed direction of suspicion. Again, very low values do not contribute much to the explanation and are of lower importance. A total of 27 (out of 32) indicators, including all new indicators, point in the assumed direction of suspicion. All of the 21 person-level indicators are consistent with the assumptions regarding their direction. In turn, five household-level indicators are not in the assumed direction: Benford's Law (BFL_H), interview duration (DUR_H), relative duration (RDUR_H), item-nonresponse (INR_H), and rounding tendency (ROUND_H). With the exception of ERS_H, most household-level indicators have very low coefficient values. It is important to note that some indicators (e.g., interview duration [DUR], rounding tendency [ROUND], and item-nonresponse [INR]) were generated for both interview types but resulted in contrary outcomes. Compared to the person-level interview, the household-level interview was much shorter with correspondingly fewer variables collected. Hence, indicators generated from a smaller set of variables might be characterized by lower explanatory power. Furthermore, answers to the household interview items were more homogeneous due to the special population,[6] which may

[6] Roughly 50 percent of refugees resided in shared accommodations, which are likely to be similar to each other (Brücker, Kosyakova, and Vallizadeh 2020). Given that most of the surveyed refugees arrived in 2015 and 2016, approximately one year or less before the interview, their households were likely less heterogeneous than had they resided in Germany for 5-10 years.

have limited the variation of these indicator values. Another possible explanation is the way in which the household data could be fabricated by the interviewers. Some household-related information might have been quite obvious for the falsifiers (e.g., composition, income, and accommodation type) or they might have conversed with the anchor-person but without a proper interview. This could have increased the "quality" of the household-level falsification and decreased the power of the indicators.

2.5 Discussion

Even though statistical falsification detection methods can be powerful tools for improving the quality control process, comparative evaluations of different methods performed on real-world data are rare. We addressed this research gap by using large-scale survey data with verified falsifications and evaluated the performance of different multivariate detection methods (Ward's Linkage clustering, Single-Linkage clustering, and the newly-proposed meta-indicator) and numerous falsification indicators. Consistent with the literature (Menold et al. 2013; de Haas and Winker 2016), the results revealed pronounced effectiveness of the different multivariate detection methods utilizing various indicators in identifying all three confirmed falsifiers. Ward's Linkage and the meta-indicator produced mostly the same accuracy, which was slightly higher than for Single-Linkage. By assessing the relative importance of single falsification indicators, we found—consistent with the literature (Hood and Bushery 1997; Li et al. 2011)—that time-related indicators are of crucial importance. This supports the notion that falsifiers aim to reduce their time investment when falsifying data. Furthermore, falsifiers failed in reproducing the Benford Distribution and were less successful in manipulating item batteries (Schäfer et al. 2004; Bredl, Winker, and Kötschau 2012; Menold et al. 2013). However, the importance of the indicators was sensitive to the level of interview data used to generate them. Indicators derived from person-level data were always in line with the directional assumptions and therefore proved to be of higher importance than those derived from household-level data.

2.5.1 Practical Implications of Results

What do these results imply for practitioners? First, while both cluster analysis and the meta-indicator performed similarly well, the meta-indicator approach proved to be more straightforward and produced less ambiguous results. Therefore, the meta-indicator might be preferred for an initial screening of the data. We recommend that users visually inspect the meta-indicator distribution and use a lenient threshold to minimize the risk of overlooking

falsifiers. Given the novelty of the approach, we encourage further applications and evaluations in other datasets to assess the generalizability of its performance and suitable thresholds. For a more thorough quality control, we recommend using both cluster analysis and the meta-indicator and compare their results. Note that statistical methods should be used in conjunction with routine non-statistical approaches (e.g., re-interviewing) for better targeting and more efficient use of resources for catching falsifiers, but also for confirming suspected falsifiers identified by the statistical methods. This is important as the premature removal of suspected falsified data without non-statistical confirmation could lead to serious bias.

Second, the relative importance of the time-related indicators (e.g., interview duration), item scale indicators (e.g., middle-responding style), and record linkage consent was particularly high. Thus, we recommend incorporating them into statistical detection methods. However, almost all falsification indicators pointed in the direction of falsification behavior and indeed proved to be essential for identifying falsifiers, even though household-level indicators were less important than person-level ones. Since some falsifiers scored very low on certain indicators while others scored very high, considering as many indicators as possible is a good strategy to identify falsifiers.

2.5.2 Limitations and Future Work

Although we showed that different detection methods performed similarly well in detecting falsifiers, each method has its drawbacks. While cluster analysis allows identifying different interviewer groups that may reflect different interviewer behavior, it does require some technical decision-making regarding clustering algorithms and may still lead to ambiguous cluster solutions requiring further inspection and expert judgment. Furthermore, cluster-analysis might not work as demonstrated if most of the interviewers are falsifiers. This also applies to the meta-indicator, which—while practically simple to implement—may also become difficult to interpret if the size of the interviewer staff is small.

We acknowledge that the results are based on a single dataset and data collection could be subject to specific opportunities and motives for the interviewers to falsify (Kosyakova et al. 2021). Hence, while the results are encouraging, these methods could work out differently for other datasets. Further, it is possible that the types of respondents assigned to an interviewer or the areas they worked in affected the results. However, such effects are unlikely for two reasons. First, due to the large number of indicators aggregated to the interviewer-level, it is improbable that an honest interviewer is flagged solely on the type or behavior of their respondents (with exception of interviewers with very few interviews). Second, upon their

arrival, refugees were distributed exogenously according to national dispersal policies, which reduces the potential for area effects. We further acknowledge that most falsifications in the used data were complete falsifications, which are easier to detect than partial falsifications (DeMatteis et al. 2020). Evaluating detection methods for partial falsification is a topic for future work. Further, the statistical methods were applied only at the end of the field period. Although the demonstrated methods could be applied in "real-time" during the field period, we are unable to assert how effective this would be. We encourage future studies to investigate this issue further. Future work should also consider the use of modern machine learning methods (e.g., random forests, generalized boosted models), which could provide additional insights on the importance of indicators and their correlations.

Appendix

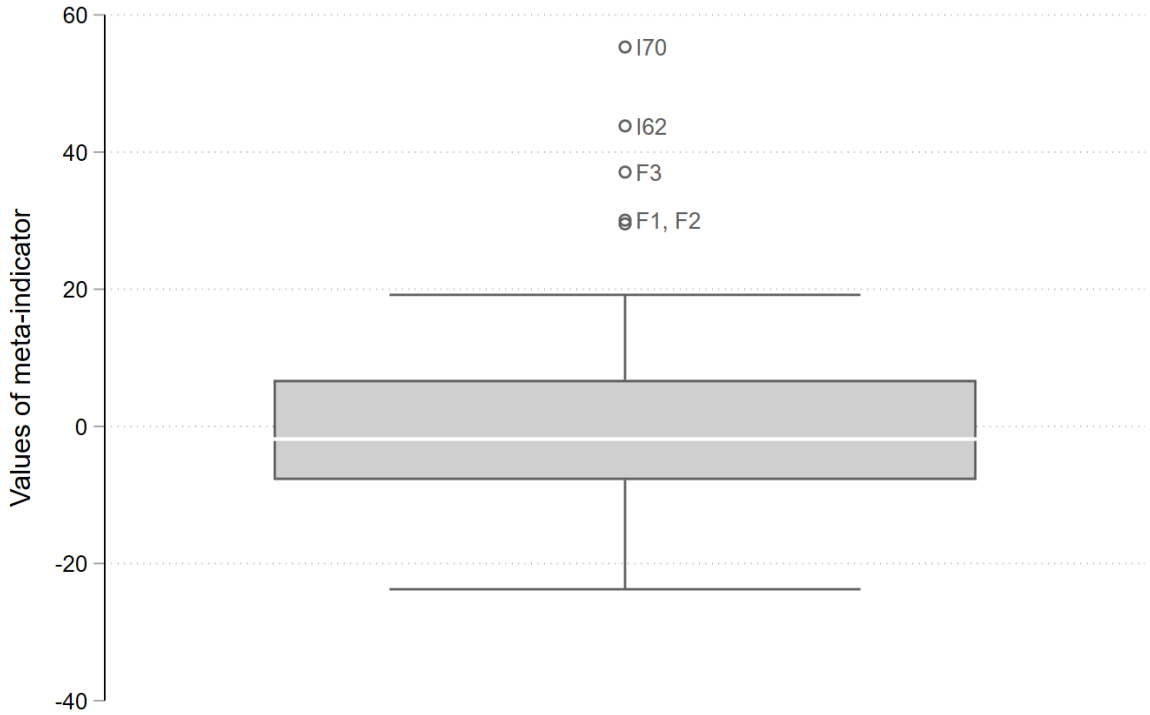

Figure A 2.1: Boxplot of the meta-indicator values.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 2.1 Interviewer-level indicator values for falsifiers and honest interviewers.

Indicator	Interview type	Total sample (excluding: F1, F2, and F3)	Falsifiers		
			F1	F2	F3
ACQ_P	person level	–0.2034	0.8292	0.7016	2.6225
BFL_P	person level	0.7150	–0.0306	6.5729	4.1294
BFL_H	household level	0.6794	0.1275	0.8487	0.5649
MAIL_H	household level	–0.0505	0.5243	0.5243	0.5243
ERS_l_P*	person level	–0.0045	2.1464	–1.0774	1.6614
ERS_m_P*	person level	0.1209	0.4889	2.5633	2.4227
ERS_h_P*	person level	0.0979	0.3935	2.4313	2.0076
ERS_H*	household level	0.1312	0.9695	1.9803	2.2602
FILT_P	person level	0.2408	–2.8745	2.8413	1.3776
FILT_H	household level	0.0575	–0.2697	1.8224	0.6541
DUR_P	person level	–0.2492	1.1268	0.9364	1.2291
DUR_H	household level	–0.2765	0.7292	–1.9405	0.3060
RDUR_P	person level	–0.2881	1.4631	0.4753	1.0810
RDUR_H	household level	–0.2733	0.7314	–2.5821	0.2031
EVAL_P	person level	–0.2136	2.7092	0.5304	1.2038
INR_P	person level	–0.0883	1.4933	0.1866	–1.7981
INR_H	household level	–0.1463	0.5096	–1.0332	–2.2182
MRS_l_P*	person level	0.1988	2.1032	–1.7229	1.5040
MRS_m_P*	person level	0.2318	0.4575	5.6283	2.8317
MRS_h_P*	person level	0.1554	–0.8409	5.5442	2.9924
MRS_H	household level	0.0649	0.9029	1.3673	0.3939
ND_P	person level	0.2260	0.4544	5.2077	3.4337
PRIM_P	person level	–0.2020	2.5178	–0.6834	1.2532
RECE_P	person level	0.0543	2.3236	–0.0324	1.9857
RLC_P	person level	–0.2528	0.8336	0.8783	0.0715
ROUND_P	person level	–0.2824	2.1264	–2.6465	1.3690
ROUND_H	household level	–0.1802	1.3899	–1.8232	–2.4090
SOR_P	person level	–0.2154	0.4853	1.6854	0.9337
STEREO_P	person level	0.3892	1.9909	1.8848	1.7101
TEL_H	household level	0.1548	1.0279	0.5217	0.4511
VAR_P	person level	0.0661	2.5035	–0.3120	0.7168
VAR_H	household level	0.0396	0.1848	–1.2022	1.6111

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

*Due to large differences in the number of scale categories, three different indicators were created. Large scales with 10 or 11 answer categories (h), medium size scales with 7 categories (m), and small scales with 4 or 5 categories (l). New indicators are shown in italics.

Table A 2.2: Correlations between falsification indicators.

	ACQ_P	BFL_P	ERS_h_P	ERS_m_P	ERS_l_P	FILT_P	DUR_P	RDUR_P	EVAL_P	INR_P
ACQ _P	1									
BFL_P	**- 0.41 (0.000)**	1								
ERS_h_P	**0.43 (0.000)**	0.01 (0.907)	1							
ERS_m_P	**0.55 (0.000)**	- 0.03 (0.765)	**0.90 (0.000)**	1						
ERS_l_P	**0.22 (0.031)**	- 0.19 (0.067)	**0.34 (0.001)**	**0.31 (0.002)**	1					
FILT_P	- 0.03 (0.746)	**0.46 (0.000)**	0.17 (0.102)	0.14 (0.169)	- 0.12 (0.245)	1				
DUR_P	0.11 (0.291)	**0.21 (0.038)**	0.13 (0.212)	0.19 (0.068)	- 0.14 (0.175)	**0.50 (0.000)**	1			
RDUR_P	0.12 (0.233)	0.16 (0.112)	0.13 (0.234)	0.19 (0.068)	- 0.14 (0.186)	**0.37 (0.000)**	**0.99 (0.000)**	1		
EVAL_P	0.13 (0.207)	- 0.06 (0.562)	0.02 (0.829)	0.04 (0.681)	0.07 (0.470)	- 0.02 (0.887)	0.20 (0.056)	**0.21 (0.042)**	1	
INR_P	**- 0.30 (0.003)**	- 0.12 (0.240)	- 0.15 (0.140)	- 0.17 (0.105)	0.03 (0.768)	- 0.30 (0.335)	- 0.11 (0.283)	- 0.07 (0.477)	0.07 (0.469)	1
MRS_h_P	0.18 (0.083)	0.07 (0.502)	**0.66 (0.000)**	**0.57 (0.000)**	**0.26 (0.011)**	**0.31 (0.002)**	0.01 (0.897)	- 0.02 (0.822)	- 0.03 (0.753)	0.03 (0.772)
MRS_m_P	0.25 (0.014)	0.10 (0.310)	**0.66 (0.000)**	**0.72 (0.000)**	**0.22 (0.033)**	**0.29 (0.004)**	0.13 (0.192)	0.11 (0.289)	0.10 (0.356)	0.19 (0.058)
MRS_l_P	- 0.07 (0.499)	0.18 (0.077)	0.16 (0.120)	0.14 (0.163)	**0.54 (0.000)**	- 0.04 (0.676)	- 0.00 (0.992)	0.01 (0.962)	- 0.04 (0.734)	0.18 (0.081)
ND_P	- 0.02 (0.847)	**0.46 (0.000)**	**0.61 (0.000)**	**0.64 (0.000)**	0.16 (0.124)	**0.48 (0.000)**	**0.32 (0.002)**	**0.28 (0.006)**	- 0.02 (0.864)	- 0.10 (0.341)
PRIM_P	**0.20 (0.049)**	**- 0.48 (0.000)**	0.07 (0.514)	0.09 (0.366)	0.12 (0.236)	**- 0.31 (0.002)**	- 0.09 (0.396)	- 0.05 (0.657)	- 0.16 (0.111)	0.06 (0.569)
RECE_P	0.16 (0.128)	0.06 (0.595)	0.04 (0.677)	0.04 (0.666)	0.07 (0.491)	- 0.07 (0.482)	- 0.11 (0.266)	- 0.12 (0.241)	- 0.18 (0.084)	0.18 (0.072)
RLC_P	- 0.13 (0.210)	0.03 (0.808)	**- 0.30 (0.002)**	**- 0.36 (0.000)**	- 0.16 (0.120)	- 0.18 (0.077)	- 0.16 (0.116)	- 0.14 (0.171)	- 0.14 (0.180)	- 0.11 (0.272)
ROUND_P	**0.24 (0.001)**	**- 0.26 (0.010)**	0.18 (0.072)	0.17 (0.105)	**0.22 (0.035)**	0.16 (0.123)	**0.26 (0.011)**	**0.25 (0.015)**	- 0.04 (0.679)	- 0.10 (0.335)
SOR_P	0.01 (0.925)	- 0.07 (0.513)	**- 0.22 (0.031)**	- 0.19 (0.063)	**- 0.48 (0.000)**	- 0.01 (0.956)	0.06 (0.554)	0.06 (0.555)	- 0.05 (0.614)	- 0.12 (0.243)
STEREO_P	**0.22 (0.029)**	0.19 (0.064)	0.17 (0.090)	0.17 (0.106)	0.07 (0.474)	**0.21 (0.044)**	- 0.02 (0.813)	- 0.07 (0.529)	0.07 (0.485)	- 0.09 (0.380)
VAR_P	**0.32 (0.001)**	0.12 (0.244)	**0.22 (0.033)**	**0.27 (0.007)**	0.16 (0.129)	0.03 (0.747)	**0.28 (0.006)**	**0.29 (0.004)**	**0.29 (0.004)**	0.07 (0.525)
BFL_H	- 0.02 (0.881)	**0.46 (0.000)**	0.17 (0.091)	0.17 (0.091)	**0.25 (0.012)**	**0.29 (0.004)**	0.07 (0.485)	0.03 (0.752)	- 0.01 (0.893)	- 0.13 (0.211)
MAIL_H	- 0.01 (0.921)	0.08 (0.439)	- 0.03 (0.786)	- 0.03 (0.786)	- 0.09 (0.409)	0.13 (0.190)	0.16 (0.120)	0.15 (0.156)	0.04 (0.698)	0.12 (0.238)
ERS_H	0.06 (0.586)	0.02 (0.821)	**0.31 (0.002)**	**0.27 (0.007)**	**0.25 (0.014)**	**0.31 (0.002)**	0.12 (0.251)	0.08 (0.465)	0.08 (0.438)	**- 0.30 (0.003)**
FILT_H	0.00 (0.989)	0.04 (0.671)	**0.24 (0.016)**	0.16 (0.131)	0.16 (0.109)	**0.27 (0.008)**	- 0.03 (0.741)	- 0.08 (0.432)	- 0.01 (0.921)	- 0.09 (0.356)
DUR_H	0.19 (0.063)	0.01 (0.891)	0.16 (0.123)	**0.21 (0.041)**	0.14 (0.161)	**0.22 (0.031)**	**0.59 (0.000)**	**0.58 (0.000)**	**0.22 (0.029)**	- 0.03 (0.748)
RDUR_H	0.20 (0.055)	0.02 (0.861)	0.13 (0.190)	**0.21 (0.039)**	0.14 (0.163)	0.17 (0.102)	**0.60 (0.000)**	**0.61 (0.000)**	**0.23 (0.022)**	- 0.04 (0.713)
INR_H	- 0.08 (0.413)	- 0.16 (0.129)	- 0.16 (0.125)	- 0.10 (0.360)	0.04 (0.679)	**- 0.40 (0.000)**	- 0.15 (0.152)	- 0.10 (0.336)	- 0.04 (0.699)	**0.65 (0.000)**
MRS_H	- 0.11 (0.307)	**0.21 (0.038)**	- 0.12 (0.257)	- 0.06 (0.588)	- 0.02 (0.830)	0.04 (0.724)	- 0.02 (0.858)	- 0.03 (0.773)	- 0.01 (0.956)	- 0.05 (0.618)
ROUND_H	- 0.17 (0.094)	0.04 (0.737)	- 0.09 (0.390)	- 0.17 (0.107)	0.04 (0.700)	- 0.16 (0.108)	- 0.13 (0.190)	- 0.11 (0.269)	0.02 (0.852)	0.11 (0.302)
TEL_H	0.06 (0.569)	0.16 (0.116)	0.10 (0.353)	0.10 (0.353)	0.03 (0.777)	0.15 (0.155)	0.14 (0.174)	0.13 (0.217)	0.09 (0.387)	0.01 (0.963)
VAR_H	0.28 (0.054)	- 0.06 (0.560)	0.08 (0.448)	0.08 (0.448)	- 0.12 (0.231)	- 0.10 (0.331)	0.09 (0.360)	0.13 (0.202)	0.12 (0.232)	0.08 (0.439)

Table A 2.2 (continued)

	MRS_h_P	MRS_m_P	MRS_l_P	ND_P	PRIM_P	RECE_P	RLC_P	ROUND_P	SOR_P	STEREO_P
MRS_h_P	1									
MRS_m_P	**0.72 (0.000)**	1								
MRS_l_P	**0.20 (0.050)**	0.18 (0.078)	1							
ND_P	**0.55 (0.000)**	**0.72 (0.000)**	0.19 (0.066)	1						
PRIM_P	- 0.02 (0.837)	0.00 (0.981)	- 0.03 (0.798)	- 0.07 (0.487)	1					
RECE_P	- 0.09 (0.395)	- 0.13 (0.218)	0.15 (0.137)	0.01 (0.930)	**0.30 (0.003)**	1				
RLC_P	- 0.11 (0.298)	- 0.20 (0.053)	0.02 (0.861)	**- 0.32 (0.002)**	0.16 (0.119)	0.15 (0.147)	1			
ROUND_P	- 0.00 (0.985)	- 0.00 (0.971)	0.11 (0.266)	- 0.05 (0.649)	0.10 (0.331)	0.16 (0.118)	- 0.11 (0.300)	1		
SOR_P	0.04 (0.696)	- 0.19 (0.065)	**- 0.30 (0.003)**	- 0.15 (0.153)	0.07 (0.472)	0.05 (0.610)	**0.28 (0.006)**	- 0.13 (0.196)	1	
STEREO_P	**0.25 (0.014)**	**0.23 (0.023)**	0.16 (0.128)	0.20 (0.053)	- 0.12 (0.243)	**0.22 (0.028)**	- 0.18 (0.085)	0.16 (0.114)	0.14 (0.185)	1
VAR_P	0.13 (0.224)	**0.23 (0.024)**	0.12 (0.226)	0.13 (0.193)	- 0.12 (0.231)	0.17 (0.101)	- 0.12 (0.226)	0.12 (0.255)	0.12 (0.228)	**0.36 (0.000)**
BFL_H	0.07 (0.473)	0.19 (0.067)	**0.37 (0.000)**	**0.24 (0.017)**	**- 0.24 (0.020)**	0.18 (0.074)	**- 0.22 (0.031)**	0.01 (0.959)	**- 0.36 (0.000)**	**0.40 (0.000)**
MAIL_H	- 0.02 (0.881)	0.02 (0.833)	- 0.02 (0.824)	0.14 (0.186)	0.04 (0.692)	0.08 (0.453)	- 0.08 (0.438)	0.07 (0.469)	0.09 (0.403)	0.01 (0.903)
ERS_H	**0.36 (0.000)**	**0.27 (0.007)**	0.15 (0.137)	**0.37 (0.000)**	0.00 (0.994)	0.10 (0.315)	**- 0.26 (0.011)**	0.14 (0.189)	0.02 (0.836)	**0.21 (0.042)**
FILT_H	0.19 (0.063)	0.16 (0.125)	0.03 (0.759)	0.15 (0.153)	- 0.06 (0.587)	- 0.17 (0.092)	**- 0.23 (0.026)**	0.08 (0.416)	**- 0.35 (0.000)**	0.02 (0.862)
DUR_H	- 0.02 (0.829)	0.07 (0.522)	0.07 (0.499)	0.18 (0.086)	0.01 (0.938)	- 0.03 (0.809)	- 0.11 (0.306)	**0.31 (0.002)**	- 0.16 (0.123)	- 0.08 (0.412)
RDUR_H	- 0.06 (0.580)	0.05 (0.646)	0.09 (0.394)	0.17 (0.092)	0.02 (0.872)	0.01 (0.951)	- 0.09 (0.396)	**0.30 (0.003)**	- 0.13 (0.209)	- 0.10 (0.348)
INR_H	- 0.16 (0.112)	- 0.10 (0.340)	0.17 (0.090)	**- 0.26 (0.010)**	0.12 (0.225)	0.06 (0.570)	0.06 (0.549)	- 0.11 (0.273)	**- 0.25 (0.015)**	- 0.15 (0.143)
MRS_H	0.01 (0.904)	- 0.10 (0.315)	0.11 (0.298)	0.10 (0.326)	- 0.13 (0.212)	**0.31 (0.002)**	0.10 (0.336)	- 0.12 (0.241)	**0.27 (0.007)**	0.15 (0.155)
ROUND_H	**- 0.21 (0.037)**	**- 0.21 (0.036)**	- 0.02 (0.833)	- 0.17 (0.106)	- 0.09 (0.369)	- 0.12 (0.236)	- 0.08 (0.445)	0.15 (0.148)	**- 0.29 (0.005)**	- 0.13 (0.198)
TEL_H	0.09 (0.369)	0.19 (0.063)	0.03 (0.797)	**0.21 (0.042)**	- 0.19 (0.068)	- 0.09 (0.411)	**- 0.22 (0.033)**	0.06 (0.561)	0.02 (0.840)	**0.27 (0.007)**
VAR_H	- 0.08 (0.448)	0.08 (0.417)	- 0.16 (0.129)	- 0.11 (0.307)	0.05 (0.603)	- 0.02 (0.837)	- 0.07 (0.480)	0.00 (0.974)	- 0.02 (0.879)	- 0.05 (0.646)

Table A 2.2 (continued)

	VAR_P	BFL_H	MAIL_H	ERS_H	FILT_H	DUR_H	RDUR_H	INR_H	MRS_H	ROUND_H
VAR_P	1									
BFL_H	**0.36 (0.000)**	1								
MAIL_H	**0.24 (0.020)**	0.14 (0.186)	1							
ERS_H	0.05 (0.602)	0.12 (0.241)	- 0.08 (0.411)	1						
FILT_H	- 0.19 (0.068)	**0.37 (0.000)**	0.16 (0.130)	0.02 (0.882)	1					
DUR_H	**0.24 (0.017)**	0.17 (0.092)	**0.23 (0.027)**	0.00 (0.986)	**0.28 (0.005)**	1				
RDUR_H	**0.31 (0.002)**	0.14 (0.180)	0.18 (0.084)	0.01 (0.936)	0.10 (0.353)	**0.98 (0.000)**	1			
INR_H	- 0.16 (0.130)	- 0.06 (0.571)	- 0.01 (0.911)	**- 0.45 (0.000)**	- 0.03 (0.805)	- 0.03 (0.790)	- 0.03 (0.773)	1		
MRS_H	**0.24 (0.020)**	- 0.04 (0.722)	0.05 (0.637)	**0.37 (0.000)**	**- 0.53 (0.000)**	- 0.08 (0.426)	0.03 (0.780)	- 0.12 (0.245)	1	
ROUND_H	- 0.16 (0.113)	0.07 (0.508)	- 0.09 (0.390)	**- 0.21 (0.039)**	0.16 (0.113)	- 0.13 (0.212)	- 0.14 (0.161)	**0.26 (0.011)**	- 0.20 (0.245)	1
TEL_H	**0.25 (0.013)**	**0.23 (0.024)**	**0.53 (0.000)**	0.09 (0.409)	0.11 (0.272)	0.12 (0.248)	0.09 (0.369)	- 0.16 (0.109)	- 0.05 (0.656)	- 0.08 (0.419)
VAR_H	**0.26 (0.009)**	0.01 (0.960)	0.06 (0.580)	- 0.37 (0.000)	0.03 (0.763)	0.15 (0.135)	0.18 (0.081)	0.12 (0.237)	**- 0.26 (0.010)**	0.02 (0.850)

Table A 2.2 (continued)

	TEL_H	VAR_H
TEL_H	1	
VAR_H	0.07 (0.482)	1

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).
Notes: P-values (provided in parentheses) are based on a two-tailed significance test.
Significant correlations ($p \leq 0.05$) are additionally marked in boldface.

References

AAPOR. 2003. “Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects.” American Association for Public Opinion Research, April 2003. Available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf.

AAPOR. 2016. “Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys”. American Association for Public Opinion Research. Available at https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf.

Benford, Frank. 1938. “The Law of Anomalous Numbers.” Proceedings of the American Philosophical Society 78(4):551–72.

Bergmann, Michael, Karin Schuller, and Frederic Malter. 2019. “Preventing Interview Falsifications During Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE).” Longitudinal and Life Course Studies 10(4):513–30.

Biemer, Paul P., and S. Lynne Stokes 1989. “The Optimal Design of Quality Control Samples to Detect Interviewer Cheating.” Journal of Official Statistics 5(1):23–39.

Birnbaum, Benjamin. 2012. “Algorithmic Approaches to Detecting Interviewer Fabrication in Surveys.” Dissertation, University of Washington. Available at http://hdl.handle.net/1773/22011.

Blasius, Jörg, and Jürgen Friedrichs. 2012. “Faked Interviews.” In Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt, edited by Samuel Salzborn, Eldad Davidov, and Jost Reinecke, 49–56. Wiesbaden: VS Verlag für Sozialwissenschaften.

Blasius, Jörg, and Victor Thiessen. 2013. “Detecting Poorly Conducted Interviews.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 67–88. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Bredl, Sebastian, Peter Winker, and Kerstin Kötschau. 2012. “A Statistical Approach to Detect Interviewer Falsification of Survey Data.” Survey Methodology Journal 38(1):1–10. Available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012001/article/11680-eng.pdf.

Brücker, Herbert, Yuliya Kosyakova, and Ehsan Vallizadeh. 2020. “Has There Been a ‘Refugee Crisis’? New Insights on the Recent Refugee Arrivals in Germany and Their Integration Prospects.” Soziale Welt 71(1-2):24–53.

Brücker, Herbert, Nina Rother, and Jürgen Schupp. 2017. “IAB-BAMF-SOEP-Befragung von Geflüchteten 2016: Studiendesign, Feldergebnisse sowie Analysen zu schulischer wie beruflicher Qualifikation, Sprachkenntnissen sowie kognitiven Potenzialen.” In IAB-

Forschungsbericht, Institut für Arbeitsmarkt und Berufsforschung. Available at https://www.iab.de/185/section.aspx/Publikation/k170918302.

Bushery, John M., Jennifer W. Reichert, Keith A. Albright, and John C. Rossiter. 1999. "Using Date and Time Stamps to Detect Interviewer Falsification." Proceedings of the Survey Research Method Section, American Statistical Association, 316–20. Available at http://www.asasrms.org/Proceedings/papers/1999_053.pdf.

Calinski, T., and J. Harabasz. 1974. "A Dendrite Method for Cluster Analysis." Communications in Statistics–Theory and Methods 3(1):1–27.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement 20(1):37–46.

Cohen, Mollie J., and Zach Warner. 2021. "How to Get Better Survey Data More Efficiently." Political Analysis 29(2):121–38.

de Haas, Samuel, and Peter Winker. 2016. "Detecting Fraudulent Interviewers by Improved Clustering Methods—The Case of Falsifications of Answers to Parts of a Questionnaire." Journal of Official Statistics 32(3):643–60.

DeMatteis, Jill M., Linda J. Young, James Dahlhamer, Ronald E. Langley, Joe Murphy, Kristen Olson, and Sharan Sharma. 2020. "Falsification in Surveys: Task Force Final Report." Washington, DC: American Association for Public Opinion Research. Available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Data_Falsification_Task_Force_Report.pdf.

DIW. 2019. "Quality Control in the IAB-BAMF-SOEP Survey of Refugees." Berlin: Deutsches Institut für Wirtschaftsforschung. Available at https://www.diw.de/en/diw_01.c.616027.en/quality_control_in_the_iab-bamf-soep_survey_of_refugees.html.

Duda, R. O., and P. E. Hart. 1973. Pattern Classification and Scene Analysis. New York: Wiley.

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser. 2014. "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys." Public Opinion Quarterly 78(3):721–33.

Edwards, Brad, Hanyu Sun, and Ryan Hubbard. 2020. "Behavior Change Techniques for Reducing Interviewer Contributions to Total Survey Error." In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 77–89. Boca Raton, FL: Taylor & Francis Group.

Everitt, Brian, and Sophia Rabe-Hesketh. 2006. Handbook of Statistical Analyses Using Stata. 4th ed. London: Capmann and Hall/CRC.

Finn, Arden, and Vimal Ranchhod. 2015. “Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey.” SALDRU Working Papers, South Africa Labour and Development Research Unit, University of Cape Town.

Fowler, Floyd, and Thomas Mangione. 1990. Standardized Survey Interviewing. Newbury Park, London, and Greater Kailash: Sage.

Grote, Janne. 2018. “The Changing Influx of Asylum Seekers in 2014–2016: Responses in Germany. Focussed Study by the German National Contact Point for the European Migration Network (EMN).” Bundesamt für Migration und Flüchtlinge (BAMF) Forschungszentrum Migration, Integration und Asyl. Available at https://nbn-resolving.org/urn:nbn:de:0168-ssoar-67637-6.

Groves, Robert M. 2004. Survey Errors and Survey Costs. Hoboken, NJ: Wiley.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. Survey Methodology. Hoboken, NJ: Wiley.

Hauck, Mathew. 1969. “Is Survey Postcard Verification Effective?” Public Opinion Quarterly 33(1):117–20.

Hill, Theodore P. 1999. “The Difficulty of Faking Data.” Chance 12(3):27–31.

Hood, Catherine C., and John M. Bushery. 1997. “Getting More Bang from the Reinterview Buck: Identifying ‘At Risk’ Interviewers.” Proceedings of the Survey Research Method Section, American Statistical Association, 820–24. Available at http://www.asasrms.org/proceedings/papers/1997_141.pdf.

IAB. 2017. “Revidierter Datensatz Der IAB-BAMF-SOEP-Befragung Von Geflüchteten.” Institut für Arbeitsmarkt und Berufsforschung. Available at http://doku.iab.de/grauepap/2017/Revidierter_Datensatz_der_IAB-BAMF-SOEP-Befragung.pdf.

Jacobsen, Jannes. 2018. “Language Barriers During the Fieldwork of the IAB-BAMF-SOEP Survey of Refugees in Germany.” In Surveying the Migrant Population: Consideration of Linguistic and Cultural Issues, edited by Dorothee Behr, 75–84. Köln: GESIS–Leibniz-Institut für Sozialwissenschaften.

Jesske, Birgit. 2013. “Concepts and Practices in Interviewer Qualification and Monitoring.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 91–102. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley.

Klecka, William R. 1980. Discriminant Analysis. Vol. 19, Quantitative Applications in Social Science Series. Newbury Park, London, and New Delhi: Sage.

Koch, Achim. 1995. “Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994.” ZUMA Nachrichten 19(36):89–105.

Koczela, Steve, Cathy Furlong, Jaki McCarthy, and Ali Mushtaq. 2015. “Curbstoning and Beyond: Confronting Data Fabrication in Survey Research.” Statistical Journal of the IAOS 31(3):413–22.

Kosyakova, Yuliya, Lukas Olbrich, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2019. “Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany.” Institut für Arbeitsmarkt- und Berufsforschung. Available at https://fdz.iab.de/187/section.aspx/Publikation/k190404302.

Kosyakova, Yuliya, Lukas Olbrich, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2021. “Positive Learning or Deviant Interviewing? Mechanisms of Experience on Interviewer Behavior.” Journal of Survey Statistics and Methodology. 10(2): 249–75.

Kosyakova, Yuliya, Jan Skopek, and Stephanie Eckman. 2015. “Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach.” International Journal of Public Opinion Research 27(3):417–31.

Kroh, Martin, Simon Kühne, Jannes Jacobsen, Manuel Siegert, and Rainer Siegers. 2017. “Sampling, Nonresponse, and Integrated Weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4)–Revised Version.” SOEP Survey Papers, No. 477. Berlin: DIW–German Institute for Economic Research. Available at http://hdl.handle.net/10419/172792.

Kuriakose, Noble, and Michael Robbins. 2016. “Don’t Get Duped: Fraud Through Duplication in Public Opinion Surveys.” Statistical Journal of the IAOS 32(3):283–91.

Landrock, Uta. 2017. “Explaining Political Participation: A Comparison of Real and Falsified Survey Data.” Statistical Journal of the IAOS 33(2):447–58.

Li, Jianzhu, J. Michael Brick, Back Tran, and Phyllis Singer. 2011. “Using Statistical Models for Sample Design of a Reinterview Program.” Journal of Official Statistics 27(3):433–50.

McLachlan, J. Geoffrey. 2004. Discriminant Analysis and Statistical Pattern Recognition. Hoboken, NJ: Wiley.

McQuitty, Louis L. 1957. “Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies.” Educational and Psychological Measurement 17(2):207–29.

Menold, Natalja, Peter Winker, Nina Storfinger, and Christoph J. Kemper. 2013. “A Method for Ex-Post Identification of Falsification in Survey Data.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 25–47. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Murphy, Joe, Rodney Baxter, Joe Eyerman, David Cunningham, and Joel Kennet. 2004. "A System for Detecting Interviewer Falsification." Proceedings of the American Statistical Association and the American Association for Public Opinion Research. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000517.pdf.

Olson, Kristen, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West. 2020. "The Past, Present, and Future of Research on Interviewer Effects." In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 3–16. Boca Raton, FL: Taylor & Francis Group.

Porras, Javier, and Ned English. 2004. "Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys." Proceedings of the Survey Research Method Section, American Statistical Association, 4223–28. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000879.pdf.

Reuband, Karl-Heinz. 1990. "Interviews, Die Keine Sind: 'Erfolge' Und 'Mißerfolge' Beim Fälschen Von Interviews." Kölner Zeitschrift für Soziologie und Sozialpsychologie 42(4):706–33.

Schäfer, Christin, Jörg-Peter Schräpler, Klaus-Robert Müller, and Gert G. Wagner. 2004. "Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methodes." DIW Discussion Paper No. 441. Berlin: DIW–German Institute for Economic Research. Available at http://hdl.handle.net/10419/18293.

Schräpler, Jörg-Peter, and Gert G. Wagner. 2005. "Characteristics and Impact of Faked Interviews in Surveys: An Analysis of Genuine Fakes in the Raw Data of SOEP." Allgemeines Statistisches Archiv 89:7–20.

Sharma, Sharan, and Michael R. Elliott. 2020. "Detecting Falsifications in a Television Audience Measurement Panel Survey." International Journal of Market Research 62(4):432–48.

Slomczynski, Kazimierz Maciek, Przemek Powalko, and Tadeusz Krauze. 2017. "Non-Unique Records in International Survey Projects: The Need for Extending Data Quality Control." Survey Research Methods 11(1):1–16.

Stokes, S. Lynne, and Patty Jones. 1989. "Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey." Proceedings of the Survey Research Method Section, American Statistical Association, 696–98. Available at http://www.asasrms.org/Proceedings/papers/1989_127.pdf.

Storfinger, Nina, and Peter Winker. 2013. "Assessing the Performance of Clustering Methods in Falsification Using Bootstrap." In Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 46–65. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Swanson, David, Moonung Cho, and John Eltinge. 2003. “Detecting possibly fraudulent or error-prone survey data using Benford’s Law.” Proceedings of the Survey Research Method Section, American Statistical Association. Available at http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000205.pdf.

Thissen, M. Rita, and Susan K. Myers. 2016. “Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality.” Statistical Journal of the IAOS 32(3):339–47.

Turner, Charles F., James N. Gribble, Alia A. Al-Tayyib, and James R. Chromy. 2002. “Falsification in Epidemiologic Surveys: Detection and Remediation.” Technical Papers on Health and Behavior Measurement, No. 53. Washington, DC: Research Triangle Institute.

Wagner, James, Kristen Olson, and Minako Edgar. 2017. “The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata.” Survey Research Methods 11(3):218–33.

Ward, Joe H. 1963. “Hierarchical Grouping to Optimize an Objective Function.” Journal of the American Statistical Association 58(301):236–44.

Weinauer, Marlene. 2019. “Be a Detective for a Day: How to Detect Falsified Interviews with Statistics.” Statistical Journal of the IAOS 35(4):569–75.

Winker, Peter, Natalja Menold, Nina Storfinger, Christoph J. Kemper, and Sabrina Stukowski. 2013. “A Method for Ex-Post Identification of Falsifications in Survey Data.” Paper presented at New Techniques and Technologies for Statistics (NTTS), Brussels, March 5–7. Available at https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_93.pdf.

3. How Falsifiers Make a Long Story Short: Identifying Partial Interviewer Falsification in Panel Surveys

Abstract

Interviewer-administered surveys are often seen as the gold-standard data collection form. Yet, some interviewers may be enticed to fabricate (parts of) interviews, leading to severe bias, especially in longitudinal analyses. Nonetheless, a common notion is that falsifications are straightforward to detect in panel surveys since large deviations between answers collected from the same respondents in adjacent waves are indicative of potentially fraudulent behavior. However, evaluations of this notion are missing from the literature. Additionally, the literature lacks methods for detecting partial falsifications. This study addresses these gaps using data from a German panel survey, including verified cases of partial falsifications. First, we assess whether various detection methods succeed in identifying partial falsifications. Second, we test the notion that falsifiers produce lower correlations between answers collected in adjacent panel survey waves. The results indicate that different data-driven methods aid in identifying partial falsifications, however, falsifiers did not produce significantly lower correlations.

3.1 Introduction

Evidence-based policymaking and associated research rely on high-quality population-based data. Interviewer-administered surveys are, in many respects, viewed as the gold-standard source of population-based data. Interviewers are tasked with ensuring data quality, for example, by contacting households, identifying target respondents, motivating them to participate, and answering their queries. Importantly, they are responsible for ensuring the standardized administration of the questionnaire (Groves et al. 2011). However, one drawback of interviewer-administered surveys, which undermines the accuracy of the collected data, is the incidence of interviewer falsification. That is, interviewers may be enticed to intentionally deviate from the prescribed interviewing guidelines and fabricate parts of interviews (otherwise known as *partial falsification*), or—in the worst case—fabricate complete interviews (also known as *complete falsification*; Groves 2004).

Even though interviewer falsification can lead to severe bias in statistical analyses (Schräpler and Wagner 2005) and various studies discuss methods of preventing or detecting

it with a focus on complete falsifications (e.g., Menold et al. 2013; Thissen and Myers 2016; Landrock 2017; Bergmann, Schuller, and Malter 2019; DeMatteis et al. 2020; Schwanhäuser et al. 2020), little is known about the effectiveness of methods for detecting partial falsifications. Addressing this neglected area is crucial for optimizing data quality in interviewer-administered surveys. Unlike complete falsifications, partial falsifications are much harder to detect, as only a subset of the data are affected (Blasius and Thiessen 2013). Hence, established detection methods may fail to identify this form of fabrication. In the context of panel surveys, a common notion is that falsifications are easier to detect compared to cross-sectional surveys, since less stable answers (e.g., in terms of lower correlations or other stability coefficients) between waves are highly suspicious (Schäfer et al. 2004a; Schräpler and Wagner 2005; Josten and Trappmann 2016). However, the literature rarely reports on the extent and form of falsification (complete or partial) in panel surveys nor the effectiveness of established detection methods in identifying partial falsifications. Further, studies rarely report on methods for assessing answer stability in panel surveys.

To address these issues, we analyze data from the German Panel Study "Labour Market and Social Security" (PASS), which includes confirmed cases of partial falsification by interviewers over multiple waves of the study. We first evaluate the performance of commonly-used detection methods and falsification indicators (i.e., indicators measuring systematic differences in response patterns between honest and dishonest interviewers) that are primarily used for identifying complete falsifications in cross-sectional surveys. Additionally, we evaluate an innovative and widely employed machine learning algorithm, Isolation Forest, that has mainly remained untapped in the interviewer falsification detection literature. Secondly, we evaluate the performance of methods specifically designed for detecting partial falsifications. Lastly, by comparing correlations between response patterns, answers to (time-)stable items, and falsification indicators we shift the focus to the longitudinal setting and evaluate the common assumption that falsifiers produce lower correlations between answers collected in adjacent waves of data collection. To date, this assumption has not been tested as detection method. A major strength of the present study is its inclusion of three different levels of analyses: interviewer-level, respondent-level, and item-level, which is more rigorous than previous evaluations.

Our findings show that detection methods typically used for identifying complete falsifications in cross-sectional survey data are also suitable for detecting some forms of partial falsifications in panel data. Further, we find that methods specifically designed for detecting partial falsifications were only effective to a limited extent in identifying the

confirmed PASS falsifications. Lastly, analyzing correlations between adjacent waves of data failed to detect the falsifiers. Taken together, simpler cross-sectional approaches were effective and sufficient for detecting the confirmed partial falsifications. However, applying a combination of detection methods was the most effective strategy for identifying the fraudulent behavior.

3.2 Interviewer Falsification: Previous Research

According to the American Association for Public Opinion Research (AAPOR), interviewer falsification is characterized by the deliberate and unreported deviation from standardized and well-considered interviewer instructions (Groves 2004). This definition, however, comprises a variety of forms of deviant interviewer behaviors such as the complete or partial fabrication of interviews, but also deviations from prescribed selection rules, or intentional miscoding of responses (Schreiner et al. 1988; Biemer and Stokes 1989; Groves 2004). These forms differ regarding their impact on data quality, the likelihood and means necessary to detect and prevent their occurrence, as well as their underlying motivations (DeMatteis et al. 2020). For a detailed overview of different falsification forms, their implications, and prevention measures, see DeMatteis et al. (2020).

3.2.1 Detecting Interviewer Falsification

Traditionally, survey organizations use a wide range of non-statistical strategies to detect interviewer falsification. For example, validation of survey data with administrative data (Koch 1995), observational methods like monitoring (Groves 2004; Jesske 2013; Robbins 2018), and re-interviewing of respondents (also known as re-contacting, validation, or verification methods) (Hauck 1969; Biemer and Stokes 1989; Groves 2004). Technological progress has further enabled survey organizations to apply advanced controlling procedures, e.g., the collection of audio recordings (CARI), the use of GPS data to verify interviewer travel routes, digital capture tools to collect screenshots or photos of the interview location (Keating et al. 2014; Thissen and Myers 2016; Finn and Ranchhod 2017; Wagner, Olson, and Edgar 2017), and rapid feedback systems to improve monitoring (Edwards, Maitland, and Connor 2017; Edwards, Sun, and Hubbard 2020).

In addition to non-statistical strategies, there have been considerable developments in statistical detection methods. Many studies focus on multivariate detection methods like cluster analysis (e.g., Menold et al. 2013; de Haas and Winker 2014; Bergmann, Schuller, and Malter 2019; Schwanhäuser et al. 2020; Schwanhäuser, Sakshaug, and Kosyakova

2022), multilevel modeling (e.g., Sharma and Elliott 2020; Olbrich et al. 2023), or machine learning (e.g., Birnbaum et al. 2013; Weinauer 2019; Jebreel et al. 2020; Cohen and Warner 2021). Further studies focus on the identification of duplicates by flagging identical response patterns that occur in multiple interviews (Slomczynski, Powalko, and Krauze 2017), or "near-duplicates", i.e., interviews with an unusually high correspondence of identical response values (Koczela et al. 2015; Kuriakose and Robbins 2016; Blasius and Thiessen 2021). These studies have partially shifted the focus from interviewer falsification to fraud committed by other parties involved in the data processing (e.g., survey institutes, supervisors, researchers).

Even though these statistical methods have proved useful in uncovering complete falsification of interviews and sometimes deviations in the household rostering or screening steps, there is near to no evidence regarding their performance for identifying partial falsification of interviews. Further, the literature largely neglects the relevance and specifics of interviewer falsification in panel surveys. There are different reasons for these two research gaps. First, complete falsifications are easier to detect and verify compared to partial falsifications (Blasius and Thiessen 2013), as partial falsifications can be subtle and affect only a small number of items; thus, many partial falsifications likely go undetected. Hence, evaluations of detection methods rely mainly on data from complete falsifications. Second, many surveys do not report on complete or partial falsifications or make the respective data available to researchers, which hinders the development and evaluation of tools for identifying falsifications. Third, a comprehensive analysis to identify partial falsifications might seem redundant from a cost-benefit perspective since the impact of partial falsifications is presumed to be low (Schräpler and Wagner 2005). Lastly, the conventional wisdom is that falsifiers are straightforward to identify in panel surveys without the need for sophisticated methods, as simple measures of response stability over adjacent waves should be able to identify suspicious behavior (Schäfer et al. 2004a; Schräpler and Wagner 2005); however, this notion has not been formally tested.

3.2.2 Detecting Partial Falsification of Interviews

To the best of our knowledge, no studies have reported on the extent of partial falsifications in panel or cross-sectional surveys. We refer to a partial falsification when an interviewer falsifies or skips some (but not all) questionnaire items or sections, intentionally miscodes some respondents' answers, or suggests answers instead of reading all response options properly. To identify the extent of partial falsifications it's important to consider the factors that drive their occurrence. Compared to complete falsifications, the motivation

behind partial falsifications can be quite different. Interviewers might be enticed to skip or falsify parts of the interview to keep the respondent engaged, avoid burdensome questions, and shorten the interview, while still receiving remuneration (e.g., Crespi 1945; Menold et al. 2013; Schwanhäuser, Sakshaug, and Kosyakova 2022). Thus, sensitive, complicated, long, or repetitive questions are at risk of being falsified. This fact lends itself to focusing on these types of questions when investigating potential falsification behavior. In the context of sensitive questions, Murphy et al. (2004) showed that falsifiers failed to reproduce actual patterns of substance use for subgroups. Further studies have used Benford's Law to examine the validity of income or other monetary distributions (e.g., Schäfer et al. 2004a; Schräpler and Wagner 2005; Schräpler 2011; Bredl, Winker, and Kötschau 2012). To identify falsified questionnaire sections, item-level timestamps can be useful as studies have demonstrated the general importance of pace- or duration-measures to identify deviant behavior (Bushery et al. 1999; Murphy et al. 2016; Schwanhäuser, Sakshaug, and Kosyakova 2022).

Among the few studies evaluating methods to identify partial falsifications, Blasius and Thiessen (2013, 2021) focus on, among other methods, using (Categorical) Principal Component Analysis (PCA) to detect repetitive or similar response patterns and outlying response structures in single item batteries. In an experimental setting, de Haas and Winker (2016) test whether cluster analysis—a method previously proposed for the identification of complete falsifications—is effective in detecting partial falsifications. Compared to identifying complete falsifications, the performance of the method was lower but still aided in identifying some partial falsifications. Murphy et al. (2016) report that some survey research organizations focus on partial falsifications by monitoring item-level indicators or prioritize monitoring of specific interviewers, but concrete details are not reported. One particular strand of literature focuses on interviewer effects on filter questions and question loops (Brüderl, Huyer-May, and Schmiedeberg 2013; Kosyakova, Skopek, and Eckman 2015; Josten and Trappmann 2016), or shortcutting, which can be considered a special form of partial falsification.

3.2.3 Detecting Interviewer Falsification in Panel Surveys

There are only a few reports on the extent of falsified interviews in panel surveys. The US Census Bureau reported falsification rates for multiple panel surveys between 1980 and 1987, varying between 0.4% in the Current Population Survey and National Crime Survey up to 6.5% in the New York City Housing Vacancy Survey (Schreiner et al. 1988). For the Survey of Health, Ageing and Retirement in Europe, Bergmann, Schuller, and Malter (2019) report an exceptional case where a regionally operating group of interviewers

falsified complete interviews which led to the deletion of 9% of the sample in the sixth wave (2015). Outside of this incident, they mention the occasional instance of interviewer falsification but do not quantify its extent. In the German Socio-Economic Panel (GSOEP), rates of complete falsification ranged from 0.1% to 2% between years 1984 and 2000 (Schäfer et al. 2004a; Schräpler and Wagner 2005). Another exceptional amount of complete falsifications was reported for the panel study IAB-BAMF-SOEP Survey of Refugees in Germany. In the first wave (2016), three falsifiers responsible for around 7% of the collected data were identified (IAB 2017; Kosyakova et al. 2019; Schwanhäuser, Sakshaug, and Kosyakova 2022). Most of these were complete falsifications rather than partial falsifications. Outside of the US and Europe, Finn and Ranchhod (2017) report total falsification rates in multiple South African surveys, e.g., 9% in the Cape Area Panel Study in 2009 and 7.3% in the second wave (2010/2011) of the National Income Dynamics Study. In a recent study, Castorena et al. (2023) reported that they had to replace 650 out of 1,500 interviews in their 2016/17 survey in Venezuela due to various interviewer-related quality concerns, like complete falsification or deviation from standardized interviewing practice.

The number of studies describing strategies for identifying falsifications in panel surveys is also small. Finn and Ranchhod (2017) utilize the panel data structure by comparing respondents' body mass index (BMI), signatures on paper-based consent forms, and the number of deceased respondents between waves. Schräpler and Wagner (2005) focus on the idea that falsifiers likely produce less stable responses between waves. By contrasting correlations of real and confirmed falsified data between waves, the authors find that correlations between 0.35 and 0.60 indicated real data from honest interviewers whereas falsifiers produced lower correlations. However, transferring this information to an appropriate detection strategy is lacking. Schäfer et al. (2004a) describe how all respondents in the GSOEP who have "considerable differences" between waves are asked to verify their interview data, but how this threshold is set is not explained. In addition, statistical process control charts aimed at monitoring interviewers' outputs for suspicious patterns (Bushery et al. 1999; Murphy et al. 2005; DeMatteis et al. 2020) can be adapted in the panel context to monitor deviations over time. An underutilized option for falsification identification is regression modeling that predicts the "falsification propensity" based on data from previous waves (Li et al. 2011; Murphy et al. 2016); however, this method requires labeled training data which are seldom available. In general, few studies seem to use the approaches outlined above.

3.3 Panel Study "Labour Market and Social Security"

With the present study, we address the aforementioned research gaps using data from the German panel study "Labour Market and Social Security" (PASS). We specifically focus on waves 7 to 15 as they include data produced by the interviewers responsible for the confirmed partial falsifications. Wave 6 is also used for longitudinal analyses. PASS is an annual household panel survey designed for academic- and policy-oriented research on the labor market, welfare state, and poverty in Germany. The PASS survey is conducted by the Institute for Employment Research (IAB), which is part of the German Federal Employment Agency (FEA), under the mandate of the Federal Ministry of Labour and Social Affairs. PASS was initiated to provide a longitudinal database for research on the German "Hartz-Reforms", in particular the introduction of a means-tested welfare benefit scheme (Unemployment Benefit II) in 2005. To allow for comparisons of recipients and non-recipients, the study is based on a dual frame. About half of households were sampled from a register of benefit recipients maintained by the FEA (Unemployment Benefit II sample) and the other half from a database of the residential population in Germany supplied by a commercial provider (general population sample), providing a representative cross-section of the population. The Unemployment Benefit II sample is refreshed yearly to include new welfare benefit recipients. The questionnaire covers topics such as material deprivation, welfare benefit receipt, employment, job search, participation in active labor-market programs, and income. The first wave of data collection started in 2006 with more than 12,000 households using a mixed-mode design of computer-assisted personal (CAPI) and computer-assisted telephone interviewing (CATI). In each household, an initial interview was conducted with the head of household (for the Unemployment Benefit II sample, the person registered as the contact person at the FEA; for the general population sample, the person most knowledgeable about household issues) followed by person-level interviews with each household member aged 15 years or older. For a more detailed description of the study design, see Trappmann et al. (2019).

While performing enhanced quality control checks—due to changes in the interviewing procedure because of the COVID-19 pandemic—two suspicious interviewers were detected between waves 14 and 15. The global onset of the COVID-19 pandemic had far-reaching consequences for face-to-face surveys in many countries, including Germany. Due to the imposition of government-mandated lookdowns and contact restrictions, CAPI interviewers in the PASS switched to telephone interviewing, the so-called CAPI-at-home mode. This mode was associated with a higher burden for interviewers as they were not

trained for telephone interviewing. Switching to this mode also made interview monitoring more challenging, as interviewers worked from home and recording the interview became technically more complicated. The enhanced quality controls included statistical identification methods and follow-up-checks performed by the survey organization (e.g., re-interviews, analysis of audio recordings) (Beste, Olbrich, and Schwanhäuser 2021). These quality control checks confirmed partial falsification behavior in the form of fabricating and skipping parts of the questionnaire and strongly deviating from the standardized interviewing procedures (e.g., suggestive probing, rephrasing answer scales). The respective interviews of both interviewers (referred to as F1 and F2 in the following) were then excluded from the officially released data.[7] Due to strict workload restrictions, the total number of excluded interviews amounted to only 0.64% of all household interviews (n = 505) and 0.71% of all person interviews (n = 814), distributed over waves 7 to 15. Affected interviews in wave 15 were replaced by re-interviews conducted by other interviewers. In the forthcoming analyses, we only consider interviews conducted by F1 and F2 as partial falsifications (see **Table 3.1** for a detailed breakdown). To evaluate different statistical methods for identifying the partial falsifications, we use preliminary PASS data which was later published as a Scientific Use File (PASS SUF W15; see Altschul et al. 2023).

Table 3.1: Number of partially falsified interviews, by falsifier, waves 7-15.

	Person interviews			Household interviews		
Waves	F1	F2	Total sample	F1	F2	Total sample
Wave 7	24	-	14,449	16	-	9,509
Wave 8	43	-	13,460	22	-	8,998
Wave 9	56	-	13,271	27	-	8,921
Wave 10	47	50	12,697	24	36	8,541
Wave 11	45	55	13,703	24	41	9,420
Wave 12	49	77	13,237	27	57	9,211
Wave 13	73	64	12,052	39	43	8,556
Wave 14	92	43	10,364	51	42	7,780
Wave 15	46	50	11,431	25	31	8,555
Total	475	339		255	250	

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

[7] Some further interviews conducted by nonfalsifying interviewers were also deleted due to standard data processing rules (see Beste, Olbrich, and Schwanhäuser 2021 for more details).

3.4 Statistical Detection Methods

The analysis is split into two parts. In the first part, we focus on analysis strategies to identify falsifications in a cross-sectional setting, analyzing each wave separately. First, we examine how well interviewer-level detection methods—namely univariate and multivariate analyses (cluster analysis) of falsification indicators on the interviewer-level (e.g., the proportion of rounded values or item nonresponse within a workload)—perform on the cross-sectional data. then, we focus on respondent-level analyses, including outlier detection (namely isolation forest) and duplicate analyses. lastly, we focus on the item-level using categorical principal component analysis (PCA). Single falsification indicators, their multivariate analysis using cluster analysis, different versions of duplicate analysis as well as categorical principal component analysis have previously been applied in the context of falsification identification. The usage of indicators has proven to be particularly useful for the identification of complete interviewer falsification (see, e.g., Bredl, Winker, and Kötschau 2012; Menold et al. 2013; de Haas and Winker 2016). In contrast, isolation forest has rarely been used as tool for falsification identification before (Jebreel et al. 2020; Olbrich, Beckmann, and Sakshaug 2024).

In the second part, we shift the focus to the longitudinal setting and strategies to identify falsifications through comparisons between waves. Here, we evaluate methods previously proposed or described in the literature that have not yet been applied in real-world settings. Namely, we make use of the panel data structure of the PASS to test the notion that falsifiers produce less stable answers between adjacent waves of data collection compared to nonfalsifiers. To focus on different dimensions for which differences in stability could occur, we compare correlations between waves for three different settings. First, we utilize interviewer-level falsification indicators—i.e., indicators measuring the prevalence of specific response styles, which are summarized across each interviewers' workloads—to identify interviewers with sudden changes in response patterns within their workloads between waves. Such changes could indicate changes in the interviewers' behavior. Second, we examine the falsification indicators similarly on the respondent-level, to identify single respondents with sudden changes in their response patterns between waves. Again, this could be caused by changes in the interviewers' behavior. Lastly, we compare correlations between various time-stable items across waves to focus on item-level changes. These correlations aid in identifying changes in the answers between waves. An overview of all methods used can be found in **Table 3.2**.

Table 3.2: Overview of all evaluated methods and rules for determining suspicion.

	Method	Description of Method	Rule/Criteria for Suspicion	Previously evaluated in the context of falsification
Cross-Sectional Identification Strategies	Falsification indicators			
	ACQ	Indicates interviews with a low share of acquiescent responding	Visual inspection of boxplot, identifying outlying data points based on 75%-quartile + 1.5*IQR	Yes
	MRS	Indicates interviews with a high share of middle responding		Yes
	INR	Indicates interviews with a low share of item nonresponse		Yes
	ROUND	Indicates interviews with a low share of rounded numbers		Yes
	DUR	Indicates interviews with a low interview duration		Yes
	Multivariate analysis of falsification indicators			
	Average Linkage	Hierarchical-agglomerative algorithm, minimizing the average distance of objects	Smallest resulting cluster for an optimal cluster solution, determined according to 30 different testing indices	Yes
	Complete Linkage	Hierarchical-agglomerative algorithm, maximizing the distance of objects		No
	Single Linkage	Hierarchical-agglomerative algorithm, combining the distance of objects		Yes
	Ward's Linkage	Hierarchical-agglomerative algorithm, minimizing the sum of squared errors		Yes
	Outlier detection – isolation forest			
	IsoForest response-data	Indicates outlying interviews based on the respondent data	95^{th} percentile of distribution	No
	IsoForest indicator-data	Indicates outlying interviews based on the indicator data		No

Table 3.2 (continued)

	Method	Description of Method	Rule/Criteria for Suspicion	Previously evaluated in the context of falsification
Cross-Sectional Identification Strategies	Analysis of duplicates			
	Duplicates analysis	Indicates identical copies of interviews, i.e., same response to every question	Number of complete duplicates	Yes
	Near-duplicates	Indicates exceptionally high shares of identical/matching responses to questions	95th percentile of the mean share of matching answers	Yes
	Categorical principal component analysis			
	PCA life satisfaction PCA work-life balance PCA Leisure Activities	Indicates low variation within and duplicated response patterns within an item battery based on factor scores	95th percentile of the share of duplicated factor scores	Yes
Longitudinal Identification Strategies	Correlations			
	Correlation between falsification indicators	Indicates low correlations between falsification indicators on interviewer-level	95th percentile of distribution	No
	Correlation between response styles	Indicates low correlations between falsification indicators on respondent-level		No
	Correlation between items	Indicates low correlations on the item-level		No

3.4.1 Cross-Sectional Identification Strategies

3.4.1.1 Interviewer-Level Analysis

Falsification indicators are based on the notion that response patterns of fabricated interviews are systematically different from response patterns of real interviews. Driven by the rational behavior of falsifiers—to remain undetected while maximizing outputs and minimizing time expenditure and effort—significant differences in response patterns can arise (Menold et al. 2013; Schwanhäuser, Sakshaug, and Kosyakova 2022). Like other data quality indicators used to study undesirable response behavior (e.g., rounding, extreme responding, acquiescent responding), falsification indicators are based on specific types of questions (e.g., rating scales, numeric questions, filter questions), and can be aggregated to the interviewer-level. In our analysis, we rely on five different indicators, described in more detail below: acquiescent responding (ACQ), middle responding style (MRS), item nonresponse (INR), rounding tendency (ROUND), and the interview duration (DUR); for more details, see Appendix **Table A 3.1**. These specific indicators are chosen for three reasons. First, they clearly allow for differentiating between low-quality data produced by respondents and possible falsification behavior of interviewers as they both manifest in different ways. For instance, the literature has found that, while respondents who simplify their response behavior tend to show more acquiescent responding behavior, often use extreme values, or frequently use "don't know" categories, falsifiers tend to avoid all these behaviors (Schäfer et al. 2004b; Bredl, Winker, and Kötschau 2012; Storfinger and Winker 2013; Menold et al. 2013; Schwanhäuser, Sakshaug, and Kosyakova 2022). Second, these indicators have been shown to work well for detecting falsifiers in survey data (Bredl, Winker, and Kötschau 2012; Menold et al. 2013; Schwanhäuser, Sakshaug, and Kosyakova 2022). Third, these indicators are likely to be available for most surveys, since most surveys document item nonresponse, interview duration, and include rating scales and open numeric items, which makes our analyses applicable to other surveys.

The literature suggests that some respondents show high acquiescence tendencies, i.e., the tendency to agree to a statement without considering its content or one's actual preference. This effect can be explained by respondents' predisposition to be agreeable, tendency to satisfice, or as a way of showing courtesy towards the interviewer (Krosnick 1999). Falsifiers, on the other hand, are not affected by this tendency to be agreeable. Further, they are often very familiar with the questionnaire and are therefore able to avoid inconsistencies in reverse-coded questions. This implies a lower rate of acquiescent responding for falsifiers compared to real respondents (Menold et al. 2013). Similarly, falsifiers have a higher tendency for choosing the

middle category in rating scales and a lower tendency for choosing the extreme categories, resulting in measurable middle responding tendencies (Porras and English 2004; Storfinger and Winker 2013). Again, this is motivated by avoiding inconsistencies and implausible combinations. In addition, falsifiers tend to provide substantive answers to every question which results in a lower rate of item nonresponse (Bredl, Winker, and Kötschau 2012), and have a lower rounding tendency for open numeric questions compared to real respondents (Menold et al. 2013). Lastly, because it is not necessary to read the questions out loud and due to the lack of respondent queries, the duration of a falsified interview is likely to be much shorter compared to a real interview, in effect increasing the interviewer's "hourly wage" if they are paid per completed interview. Taken together, compared to respondents' response styles that are typically associated with low data quality, we expect to observe effects in the opposite direction for falsifiers.

After aggregating the indicators to the interviewer-level, we check the data for interviewers with suspiciously high average indicator values. Furthermore, we use the falsification indicators in a cluster analysis to analyze them in a multivariate way (e.g., Menold et al. 2013; de Haas and Winker 2014; Bergmann, Schuller, and Malter 2019; Schwanhäuser et al. 2020; Schwanhäuser, Sakshaug, and Kosyakova 2022). In the context of complete falsifications, this procedure has shown promise as it increases the joint evidence of the falsification. However, in the context of partial falsifications, the joint analysis could hinder the identification of such falsifiers, as the indicators may sometimes rely on real data and other times on falsified data. For example, if an interviewer only fabricates answers to item batteries, most indicators would point towards an unsuspicious interview except for the indicators focusing on item batteries. These suspicious indicator values may be canceled out in a multivariate analysis. As the literature has used different clustering algorithms, we evaluate a variety of different algorithms, namely Average Linkage, Complete Linkage, Single Linkage, as well as Ward's Linkage. For details on the different algorithms and a detailed introduction to cluster analysis, see for example Kaufman and Rousseeuw (1990).

3.4.1.2 Respondent-Level Analysis

Previous literature has used different outlier detection methods to identify interviews with suspicious patterns or illogical combinations, ranging from simple thresholds to machine learning methods (e.g., Hood and Bushery 1997; Porras and English 2004; Murphy et al. 2005; Birnbaum et al. 2013; Weinauer 2019). We follow a similar approach by applying a well-known outlier detection method, namely Isolation Forest (Jebreel et al. 2020; Olbrich, Beckmann, and

Sakshaug 2024). Isolation forest is an unsupervised machine learning algorithm—namely a decision-tree-based method—that allows for the identification of anomalies or outliers in data frames. By randomly splitting the data according to certain variables, the algorithm allows for assessing the rarity of an observation—the fewer splits a tree needs to isolate the observation in a single branch, the rarer or more outlying an observation is (Liu, Ting, and Zhou 2008; Liu, Ting, and Zhou 2012). The average depth across multiple trees is translated into an anomaly score, ranging between 0 and 1, with values closer to one indicating outlying data points. Isolation Forest is especially suited for the identification of complete as well as partial falsifications as the method does not require assumptions regarding which responses or response patterns should be considered as suspicious. Further, it is insensitive to multi-modal distributions which could be produced by partial falsifications. Lastly, the method is fast and easy to fit and insensitive to the scales of the used variables (especially important when working with raw respondent data). We apply the algorithm first to the raw respondent data and, second, to the falsification indicators for each interview, i.e., the individual response styles of each respondent (not aggregated to the interviewer-level). In this way, outlying responses as well as outlying response behaviors on the respondent-level can be identified. To link back to the interviewer-level and identify possible interviewer falsification, we evaluate the median of the resulting anomaly scores within an interviewer's workload.

Alongside isolation forest, we also check the data for duplicates (i.e., completely identical data rows) and near-duplicates (i.e., cases with a suspiciously high correspondence between response values) (Koczela et al. 2015; Kuriakose and Robbins 2016; Slomczynski, Powalko, and Krauze 2017). We check the data for both kinds of duplicates, first, for all respondents of a wave, i.e., for (near)-duplicates between all interviews, and second, separately for all respondents assigned to a given interviewer, i.e., for (near) duplicates within an interviewer's workload.

3.4.1.3 Item-Level Analysis

To evaluate methods more suited for the identification of partial falsifications we follow the approach of Blasius and Thiessen (2012, 2015, 2021) and use (categorical) principal component analysis (PCA). Here, we focus on relatively sensitive, long, or complicated item batteries. The basic idea of the method is closely related to the falsification indicators described above, focusing on interviewers with suspicious response patterns e.g., straightlining or low variance of answer patterns. In line with Blasius and Thiessen (2021), we use the scaling method categorical PCA to obtain factor scores for different item batteries. Since we are solely

interested in identifying response patterns, only the first component of each item battery is extracted. In this way, categorical PCA produces a unique factor score for each response combination of an item battery, allowing to identify identical or similar response patterns. The respective scores allow for calculating a mean factor score as well as the share of duplicated factor scores on the interviewer-level. We assume that honest interviewers show various different factor scores, whereas falsifiers are characterized by a lower variation within item batteries (for example, because they use the same pattern to quickly click through an item battery without asking the questions). Hence, they should show a high share of duplicated scores and a clear tendency toward certain scores, which is therefore considered a suspicious pattern.

3.4.2 Longitudinal Identification Strategies

3.4.2.1 Interviewer-Level Analysis

For the longitudinal analysis, we first rely on the same falsification indicators used for the cross-sectional analysis. As these indicators are calculated on the interviewer-level, they are able to identify changes in the interviewer's behavior rather than changes within the response patterns of single respondents. As an example, if we observe a steep decrease in the average interview duration within an interviewer's workload, it is likely that the interviewer used a shortcutting strategy to shorten the interviews, rather than single respondents. To identify substantial changes in the falsification indicators, we rely on correlations between the five indicator values across the sequential waves. We note that, as most CAPI-interviewers stay within their local regions across different waves, it is unlikely that area effects distort the values of the indicators or their correlations.

3.4.2.2 Respondent-Level Analysis

Similarly, we utilize the falsification indicators to assess changes between waves in the respondents' response styles. Here, we rely on respondent-level measurements of ACQ, DUR, INR, MRS, and ROUND to assess individual response styles and changes in their respective response styles between waves. Based on the literature, we assume a high level of stability within individual respondent's response styles between waves (Billiet and Davidov 2008; Weijters Geuens, and Schillewaert 2010; Van Vaerenbergh and Thomas 2013). Hence, we posit that sudden changes within the individual response styles indicate potential falsifications. We assume that every respondent has their own tendency for item nonresponse, rounding, acquiescence, and so on. Even though respondents show tendencies of learning behavior in panels, these changes can also be attributed to interviewers if they happen for multiple

respondents of the same interviewer. Hence, we again rely on correlations—here, on the respondent-level—and summarize the results on the interviewer-level to identify interviewers with a suspicious number of respondents who change their response behavior between waves.

3.4.2.3 Item-Level Analysis

Focusing again on methods especially targeting partial falsifications, we lastly use correlations between single items. Relying on the basic idea of Schäfer et al. (2004b) as well as Schräpler and Wagner (2005), we examine the stability of answers from the same survey respondents for subsequent survey waves using correlations. Namely, we focus on the longitudinal correlation of items, especially items that are relatively time stable. For example, personality traits, satisfaction with different areas of life, or assessment of one's social position in society (see, e.g., Fujita and Diener 2005; Good, Willoughby, and Busseri 2011). To ensure that the correlations of the items used are indeed time-stable, we review the overall correlation for each item between waves. Only items with moderate (coefficients between 0.3 and 0.5) or strong (coefficients between 0.5 and 1) overall correlations are used in the analysis. To evaluate the utility of these correlations for detecting suspicious interviewers, we rely on the average correlation of each interviewer for each wave.

3.5 Results

Due to the extensive number of analyses and results, we mainly report the results for wave 15 for the cross-sectional analyses and adjacent waves 14 and 15 for the longitudinal analyses. Results for the other waves can be found in the Appendix with only key findings mentioned here. Note that the number of interviewers slightly varies across analyses, as some interviewers have to be excluded due to missing values depending on the method used.

3.5.1 Cross-Sectional Identification Results

3.5.1.1 Falsification Indicators

As mentioned above, we focus on the following key indicators: acquiescent responding (ACQ), interview duration (DUR), item nonresponse (INR), middle responding style (MRS), and rounding tendency (ROUND). We rely on normalized i.e., min-max-standardized versions of the indicator, scaling all indicators to values between 0 and 1 with values close to 1 being suspicious. To simplify the interpretation for some analyses, we further rely on the average value across all normalized indicators (AVER). All indicators were aggregated to the interviewer-level, meaning that there is one unique indicator value per interviewer and wave.

Figure 3.1 shows boxplots for the five falsification indicators and the average of all indicators for wave 15. Except for INR, most indicator values lie within the range from 0.25 to 0.75, with some outliers above or below. As there are only low rates of item nonresponse in the PASS survey, the INR indicator highlights that most interviewers had no item nonresponse. Therefore, this indicator might be less useful for identifying outlying interviewers. The average values of all indicators show a clear outlier in the suspicious direction, and two outliers closer to zero. Analogous distributional patterns are also found for the eight previous waves (see Appendix **Table A 3.2** and **Figure A 3.1**).

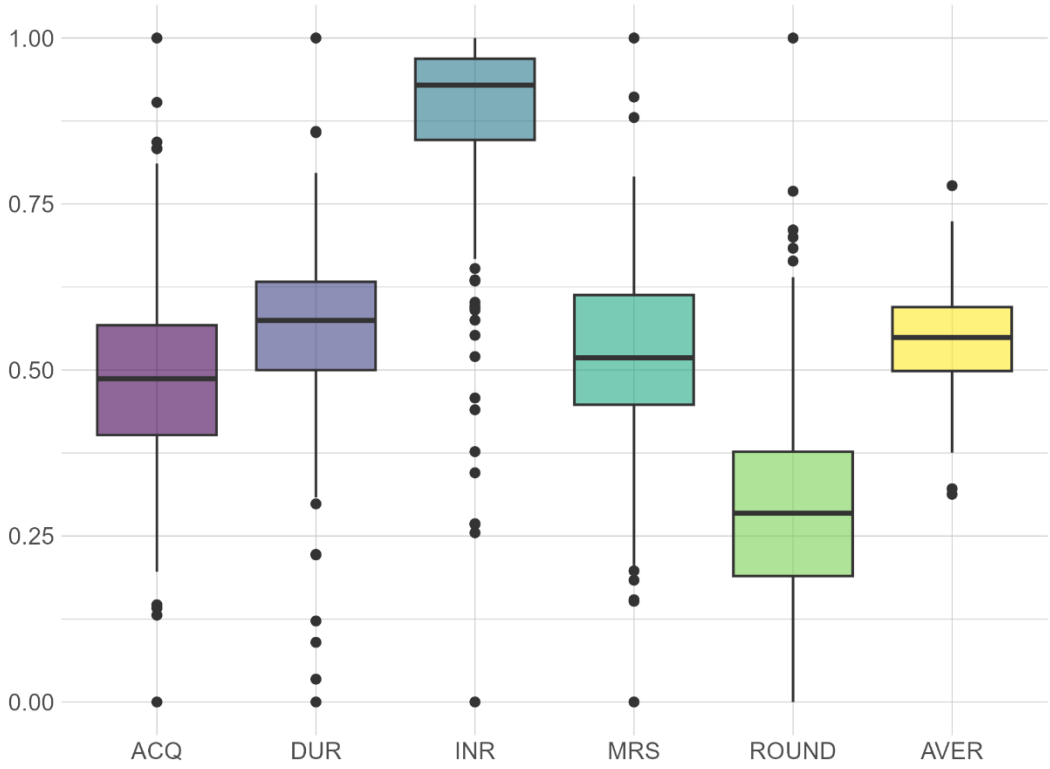

Figure 3.1: Boxplots of falsification indicators and average over all falsification indicators, wave 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: All falsification indicators are aggregated on the interviewer-level and normalized (i.e., ranging between 0 and 1) with suspicious values being close to 1. Note that most interviewers had no item nonresponse in the PASS study, which explains why the INR indicator displays many values close to 1.

To get a better overview of the outlier in the suspicious direction, we utilize a heat map (**Figure 3.2**). This plots all five indicator values of single interviewers ordered by the average overall indicator value (AVER), ranging from 1 (most suspicious) to 225 (least suspicious). The first line on the x-axis denotes the interviewer with the highest indicator average (0.78), whereas the last line denotes the interviewer with the lowest indicator average (0.31). As the heat map

reveals, the interviewer with the highest average indicator value shows suspicious falsification indicator values for item nonresponse (INR; 1.00), the duration of the interview (DUR; 0.86), acquiescence (ACQ; 0.84), as well as middle responding (MRS; 0.78). We find that this interviewer, with rank 1 out of 225, is F1. Contrary to F1 who clearly shows outlying indicator values in the suspicious direction, F2 is not labeled as suspicious, taking only rank 220 out of 225 (see Appendix **Table A 3.3** for more details). However, F2 shows outlying indicator values on the other side of the indicator scale, as two indicators—INR and MRS—are 0.00. Similar conclusions can be found for wave 14 (Appendix **Table A 3.4** and **Figure A 3.2**). For all other waves, both falsifiers show unsuspicious indicator values. Multiple other interviewers appear repeatedly with suspicious values, even without having been falsifiers (highlighted in gray in Appendix **Table A 3.4**). Hence, these indicators seem to be helpful for identifying interviewers for further controlling but are probably less stable for the precise identification of partial falsifications.

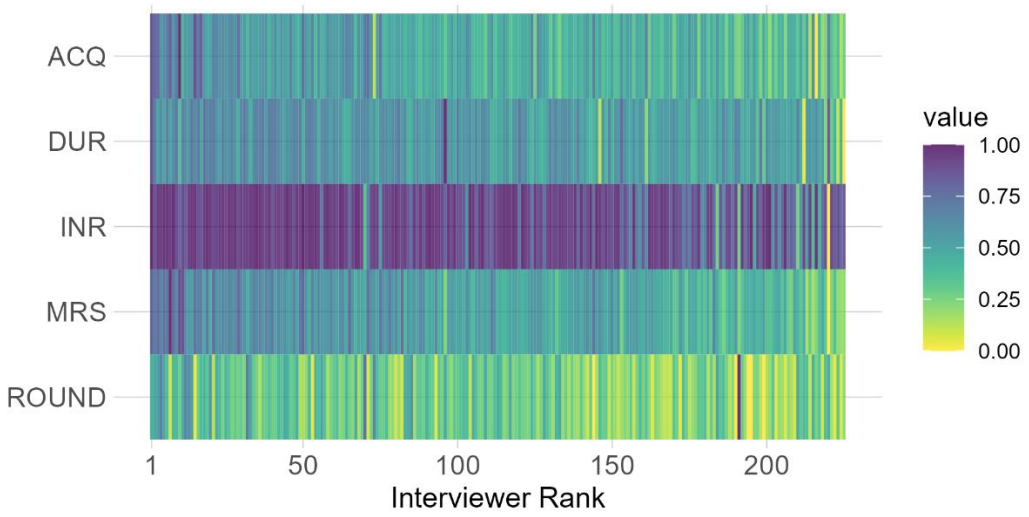

Figure 3.2: Heat map of falsification indicators per interviewer, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: Interviewers are ranked based on the average across all indicator values, ranging from 1 (most suspicious) to 225 (least suspicious).

3.5.1.2 Cluster Analysis

Combining the discussed indicators in the Cluster Analysis results in slightly different outcomes, depending on the respective clustering algorithm. However, one similarity is that all four algorithms identify at least one smaller cluster. For Average Linkage (**Figure 3.3**), Cluster 1 (purple) includes six interviewers. Similarly, Complete Linkages' Cluster 1 (purple) includes two interviewers (Appendix **Figure A 3.3**). Single Linkage (**Figure 3.4**) results in four smaller

clusters (Cluster 1, 2, 3, and 5) including six interviewers. Lastly, for Ward's Linkage (Appendix **Figure A 3.4**) Cluster 2 (turquoise) includes 26 interviewers. As falsifiers are usually a minority among interviewers and we are interested in finding outlying patterns, these clusters can be labeled as suspicious. **Table 3.3** further shows that all these suspicious clusters include F2 as well as another interviewer, I475 (for the full dendrogram of all algorithms, see Appendix **Figure A 3.5** to **Figure A 3.8**).

Table 3.3: List of interviewers within the suspicious clusters, wave 15.

Average Linkage Cluster 1 n = 6	Complete Linkage Cluster 1 n = 2	Single Linkage Cluster 1,2,3,5 n = 6	Ward's Linkage Cluster 2 n = 26
F2	F2	F2	F2
I475	I475	I475	I475
I539	-	I539	I539
I561	-	-	I561
I636	-	I636	I636
I708	-	I708	I708
-	-	I43	-
-	-	-	I20
-	-	-	I158
-	-	-	I174
-	-	-	I275
-	-	-	I345
-	-	-	I358
-	-	-	I408
-	-	-	I494
-	-	-	I543
-	-	-	I620
-	-	-	I642
-	-	-	I678
-	-	-	I686
-	-	-	I767
-	-	-	I768
-	-	-	I770
-	-	-	I781
-	-	-	I783
-	-	-	I822
-	-	-	I868

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

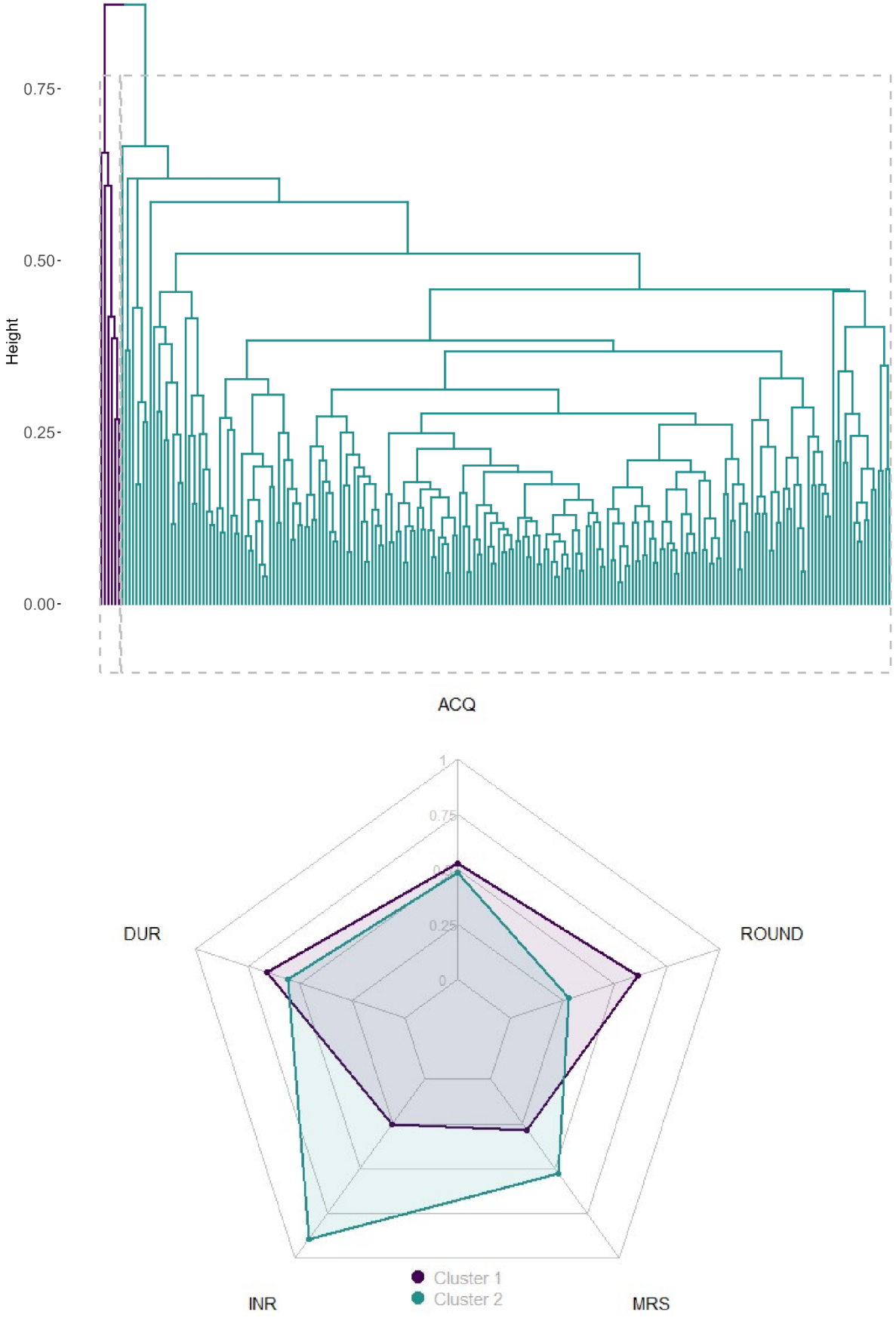

Figure 3.3: Dendrogram and radar plot for 2-cluster solution of Average Linkage, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

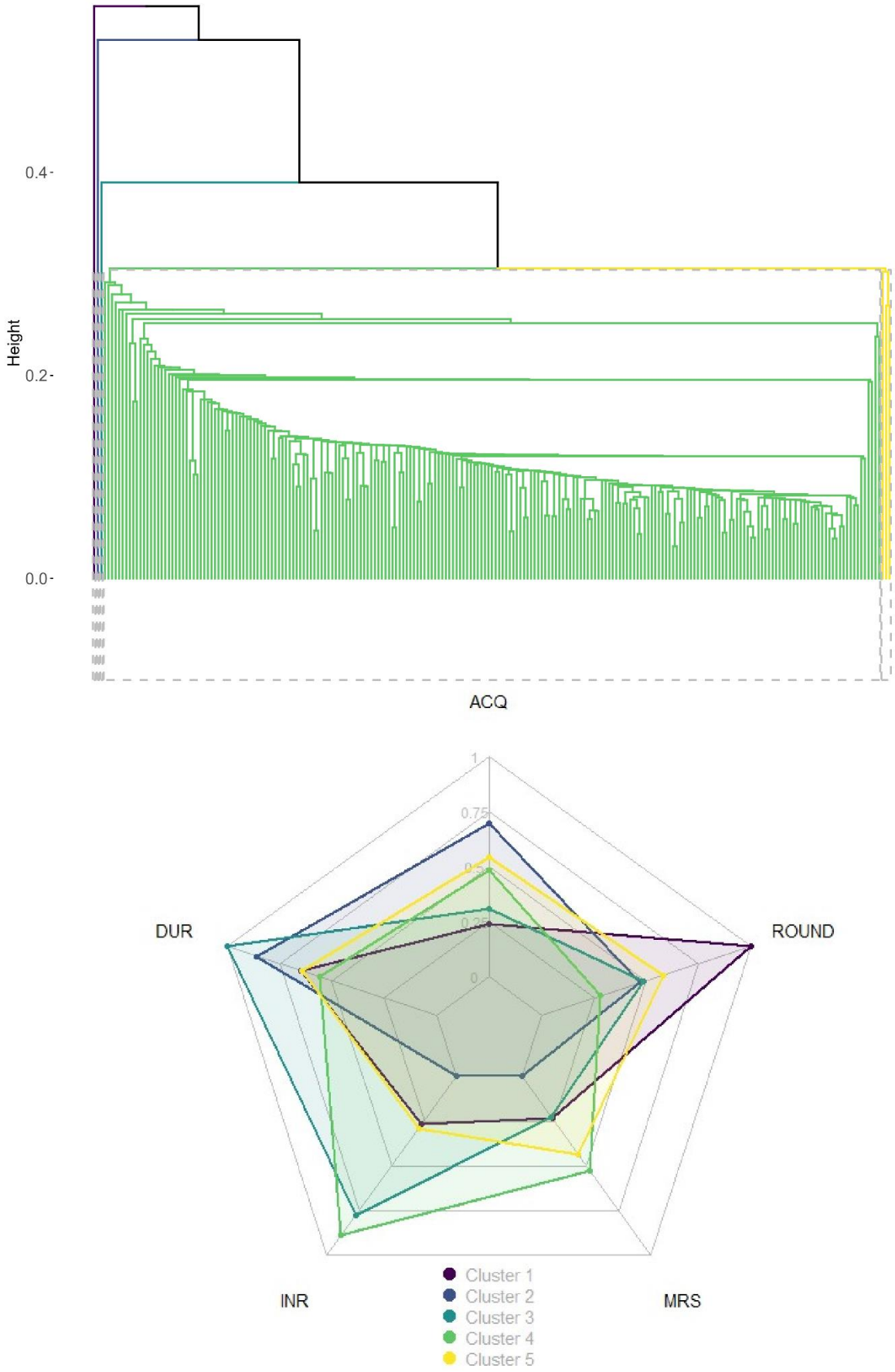

Figure 3.4: Dendrogram and radar plot for 5-cluster solution of Single Linkage, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

For Average Linkage, Single Linkage, and Ward's Linkage, the same three interviewers (I539, I636, and I708) are additionally included in the suspicious cluster(s). Further, it is important to note that Ward's Linkage produces a huge number of false-positive cases. However, F1 is included in none of these clusters. The corresponding radar plots provide an overview of the patterns of the individual clusters. We find that especially for the indicators DUR, INR, and MRS, F2 shows large deviations from the rest of the sample. The patterns contradict assumptions on the direction of falsification indicators (closer to 1 being the assumed suspicious direction). This could indicate an individual style of data fabrication, possibly related to the nature of the falsification strategy.

3.5.1.3 Isolation Forest

In the next step, we turn to the outlier detection method, relying on the machine learning algorithm Isolation Forest. Starting with the analysis of the actual respondent data, we use all categorical or continuous variables with less than 10% missing values (missings due to filter or item nonresponse). For wave 15, 71 different variables fulfilled these criteria. The algorithm calculates a unique anomaly score for each interview. To identify interviewers with outlying anomaly scores, we further calculated the median anomaly score for each interviewer. The results of this analysis, ordered by the rank of the median anomaly score per interviewer ranging from 1 (most suspicious) to 222 (least suspicious), are shown in **Figure 3.5**. The respective intervals denote the 10th and 90th percentiles of each interviewer's anomaly scores (Appendix **Figure A 3.9** and **Figure A 3.10** show the results for all waves).

We find that, on average, most interviewers produce very similar anomaly scores. The only clear outlier is F2, which is mainly driven by the large share of item nonresponse in F2's wave 15 data. Hence, answers within the workload of this falsifier differ systematically from the other respondent data. A robustness check replacing missing values with the mean value across an interviewer's workload demonstrates the effect of item nonresponse (see Appendix **Figure A 3.11**). In this case, a simple analysis of the item nonresponse indicator would have been sufficient. However, this is not true for all other waves, where F2 is sometimes identified as outlying independent of the coding of the missing values (see Appendix **Figure A 3.12**). In general, the observed data for F1 is more in line with the real respondent data.

This finding, regarding F1, slightly changes if we use Isolation Forest on the five different indicator values ACQ, MRS, INR, ROUND, and DUR on the respondent-level (**Figure 3.6**). Again, F2 is clearly identified as an outlier, however, this is followed by four

similar outlying interviewers who did not falsify data. F1 shows slightly outlying indicator values according to Isolation Forest. Hence, indicator values for both falsifiers differ from the indicator values of most interviewers.

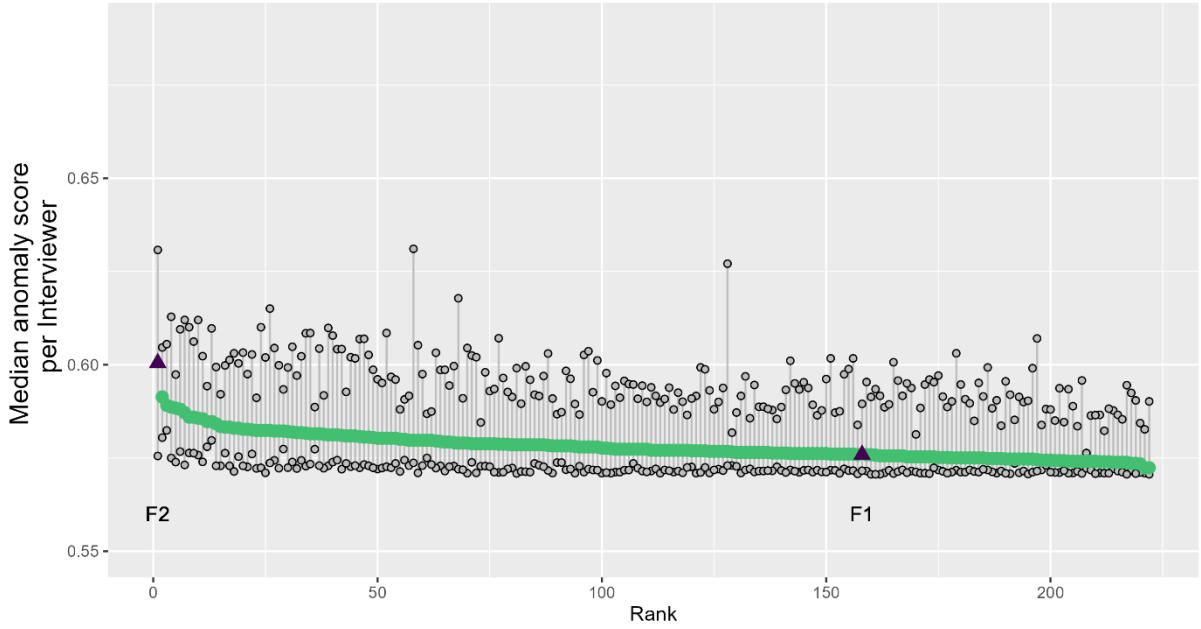

Figure 3.5: Median anomaly score per interviewer for respondent data, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: Interviewers are ranked and ordered based on their resulting median, ranging from 1 (most suspicious) to 222 (least suspicious).

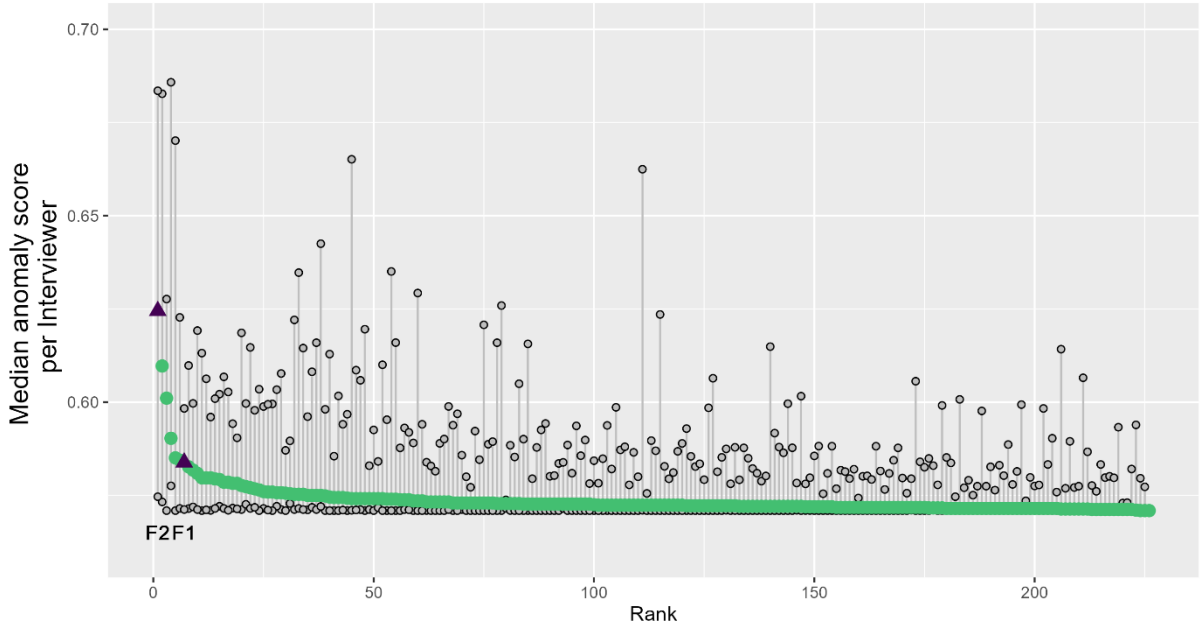

Figure 3.6: Median anomaly score per interviewer for respondent-level indicators, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: Interviewers are ranked and ordered based on their resulting median, ranging from 1 (most suspicious) to 225 (least suspicious).

Interestingly, similar results can again be found for wave 14. Between waves 7 to 13, anomaly scores seem to get increasingly suspicious for both interviewers (Appendix **Figure A 3.9**, **Figure A 3.10**, and **Figure A 3.12**). This could indicate that these interviewers either falsified more sections of the interview, falsified more frequently in their workload, or simply became more careless about falsifying over time.

3.5.1.4 Duplicate Analysis

The duplicate analysis results show that within and across all waves no exact duplicates––i.e., identical answers to all questions—exist. However, we find some interviews that share an exceptionally high proportion of answers with other interviews and interviewers that accumulate high shares of these near-duplicates. We use two different strategies to identify interviewers with an accumulation of near-duplicates, i.e., high shares of matching answers between two interviews. First, we compare each interview with all other interviews to identify the share of matching answers across all respondents and interviewers. A high share of matching answers between interviews coming from different interviewers could indicate task simplifications like straightlining or stereotyping resulting in reduced variance, often produced by falsifiers (Schäfer et al. 2004b; Menold et al. 2013). Second, we compare interviews within single interviewers' workloads. A high share of matching answers within single interviewers' outputs could indicate that the interviewer copied parts of the interviews or used the same falsification strategies (e.g., using specific filters) repeatedly. To identify interviewers with an accumulation of near-duplicates, we further calculated the mean share of duplicated answers for both measures. **Figure 3.7** shows a scatter plot of both measures.

For wave 15, we observe that F1 and F2 show suspiciously high shares of matching answers between interviews within their workloads[8]. Their values fall within the highest 5% of the whole sample (gray dashed line), with F2 having the highest share and F1 having the second highest share. This replicates our findings from the falsification indicators that revealed both falsifiers used the same type of response behavior repeatedly. Additionally, F1 and F2 show relatively high shares of matching answers between all interviewers—being among the 20 highest scoring interviewers. However, the results are less clear than for the near-duplicates within their own workloads, as some interviewers show similar patterns without being falsifiers. For the results of waves 7 to 14, we once again observe increasingly suspicious patterns for both falsifiers over time (Appendix **Figure A 3.13**).

[8] Note that for larger interviewer workloads higher rates of matching answers are expected by design.

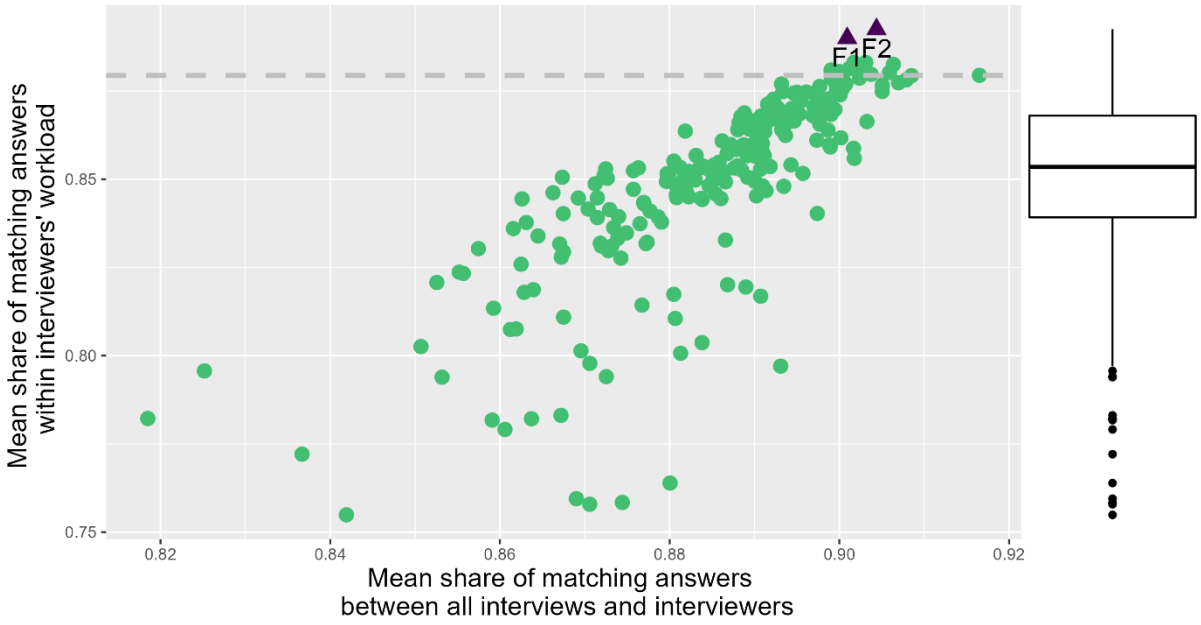

Figure 3.7: Mean share of matching answers within interviewers' workload and between all interviews, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: The dashed line denotes the 95th percentile of the mean share of matching answers within interviewers' workload.

3.5.1.5 Principal Component Analysis

As the last method for the cross-sectional analyses, Categorical Principal Component Analysis is used to identify duplicated response patterns within different item batteries. **Figure 3.8** shows the share of duplicated factor scores per interviewer for five different item batteries of wave 15 (for further waves, see Appendix **Figure A 3.14**): satisfaction with different aspects of life (PA0100-PA0300), aspects of work-life balance (PQB0900), experiences with the German Federal Employment Agency (PTK2500), acceptance of disadvantages in the process of job searching (PAS1400), and frequency of leisure activities (PSK0600) (for further information on the item batteries, see Appendix **Table A 3.5**).

We find that, with exception of the two item batteries on acceptance of disadvantages while searching for a job and on experiences with the German Federal Employment Agency,[9] multiple interviewers show exceptionally high numbers of duplicated factor scores, indicating the repeated use of the same response patterns.

[9] These item batteries were only asked of unemployed people. Further, both include a larger number of items. Hence, these item batteries are less likely to include duplicated response patterns. This highlights the importance of questionnaire characteristics when interpreting the results from these kind of falsification methods (see Simmons et al. 2016, for a more in-depth discussion).

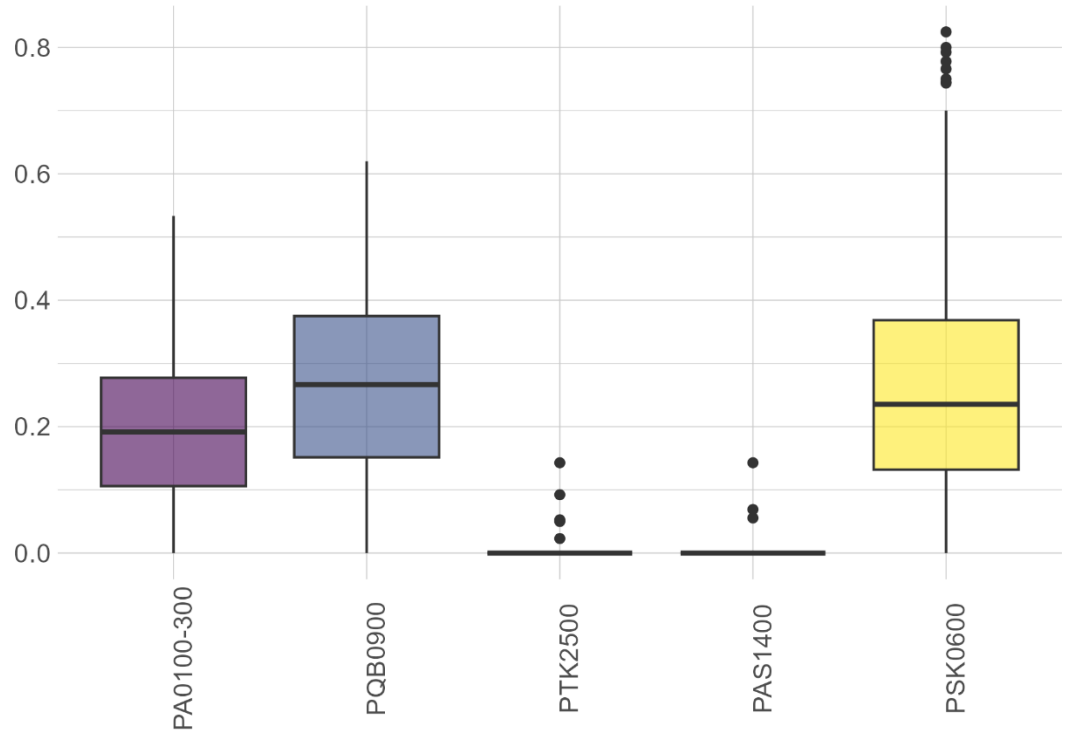

Figure 3.8: Boxplots of the share of duplicated factor scores per interviewer for different item batteries, wave 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

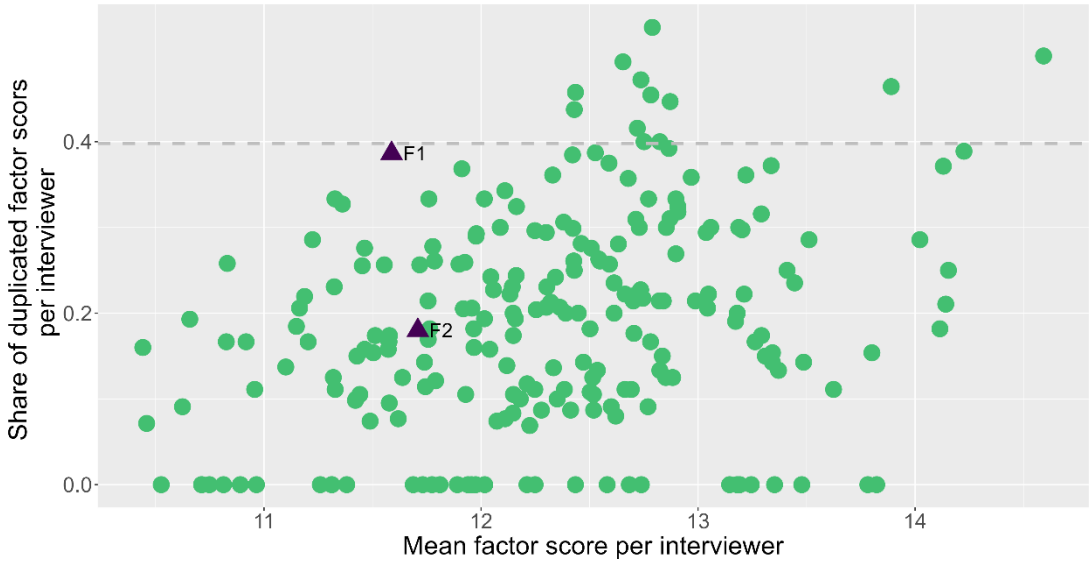

Figure 3.9: Scatter plots of the share of duplicated factor scores and the mean factor score per interviewer, item batterie life satisfaction, wave 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: The dashed line denotes the 95th percentile.

Figure 3.11, **Figure 3.10**, and **Figure 3.11** show the proportion of duplicated factor scores and the average factor score per interviewer for three respective item batteries life

satisfaction, work-life balance, and leisure activities. The most outlying values in terms of the share of duplicated factor scores are indicated by the dotted gray line. It denotes the 95th percentile across these shares. In general, F1 shows more suspicious patterns in terms of the share of duplicated factor scores compared to F2. For the item battery on leisure activities, both F1 and F2 produce suspicious patterns in terms of duplicated scores and the mean factor scores. This could indicate that only answers for the last item battery were falsified.

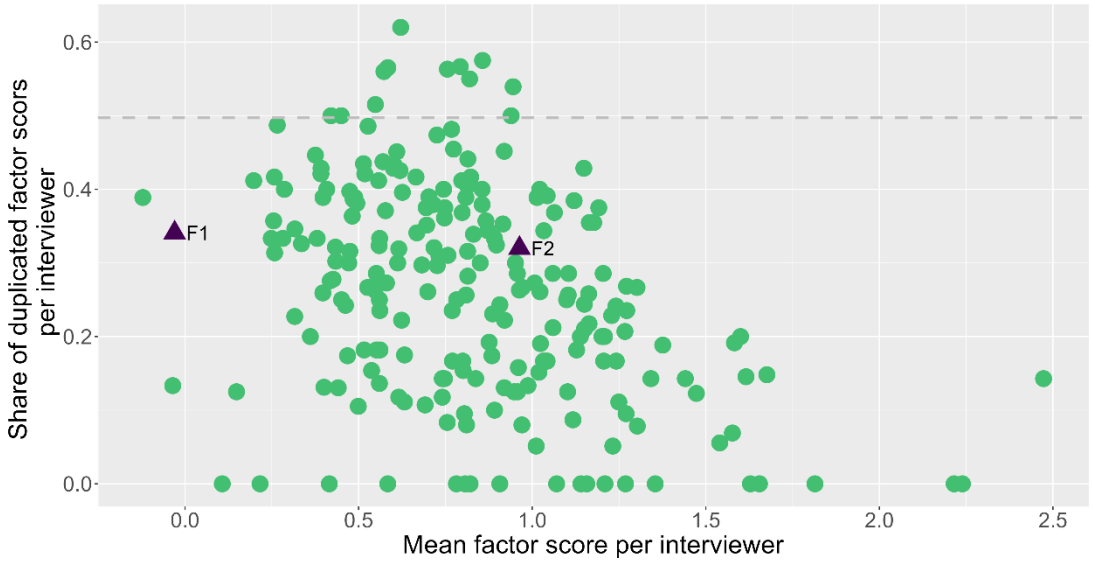

Figure 3.10: Scatter plots of the share of duplicated factor scores and the mean factor score per interviewer, item batterie work-life balance, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: The dashed line denotes the 95th percentile.

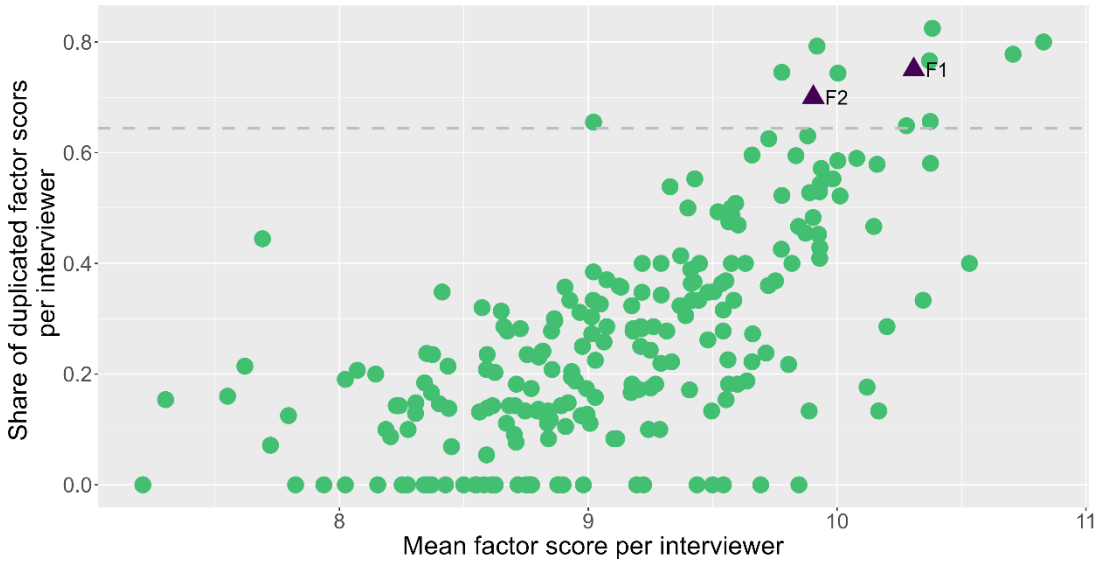

Figure 3.11: Scatter plots of the share of duplicated factor scores and the mean factor score per interviewer, item batterie leisure activities, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: The dashed line denotes the 95th percentile.

3.5.2 Longitudinal Identification Results

3.5.2.1 Correlations Between Falsification Indicators

We now shift the focus to longitudinal identification methods. Starting with the correlation between falsification indicators on the interviewer-level, we find that most interviewers repeatedly show similar response patterns between waves, indicated by high correlations. As **Figure 3.12** shows, most correlations between the falsification indicators lie above 0.5 for all waves. Only a few interviewers show correlations below this threshold.

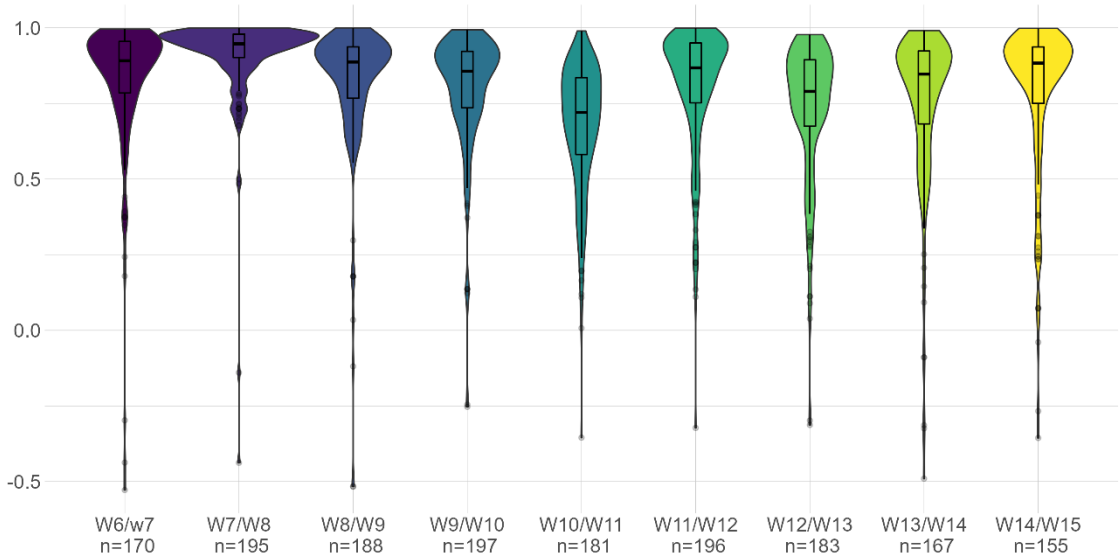

Figure 3.12: Violin plot including boxplot for correlations between falsification indicators, waves 6-15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Table 3.4 shows the interviewer IDs for the interviewers with the most outlying correlations between waves 14 and 15. However, F1 and F2 are not among them. This is also true for the other waves. Hence, correlations between falsification indicators were unable to identify falsifiers in the data. As we have seen for the cross-sectional analysis, the falsifiers seem to have adapted their behavior slowly over time. Therefore, no sudden changes—signified by low correlations—are observed.

Table 3.4: List of the most outlying interviewers with respect to correlations between falsification indicators, wave 14/15.

Interviewer ID	Falsifier	Correlation
I822	No	-0.36
I241	No	-0.27
I729	No	-0.04
I345	No	0.07
I408	No	0.07
I417	No	0.07
I220	No	0.24
I267	No	0.24
I259	No	0.24
I276	No	0.24
I594	No	0.26
I158	No	0.27
I174	No	0.31
I761	No	0.31
I172	No	0.38
I363	No	0.38
I381	No	0.38
I600	No	0.45
I636	No	0.48
I551	No	0.49
I620	No	0.49
I194	No	0.49

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

3.5.2.2 Correlation Between Response Patterns

We continue with the correlations between falsification indicators on the respondent-level—and therefore indicators of response patterns—and find similar results as for the correlations between falsification indicators on the interviewer-level. To summarize, we calculated the median correlation as well as the respective 10^{th} and 90^{th} percentiles for the respondents of a given interviewer. As **Figure 3.13** shows, F1 and F2 neither stand out due to exceptionally low average correlations nor due to exceptionally low percentiles (for further waves, see Appendix **Figure A 3.15**). Most interviewers show similar correlations ranging between 0.54 and 0.95. Only few interviewers show lower correlations[10]. Similar to the

[10] Note that correlations could be distorted for interviewers with a lot of respondent-switches, i.e., respondents that were interviewed by a different interviewer in the previous wave. Correlations might be lower for these interviewers, as a change of interviewer might also lead to some changes in response styles, i.e., indicators. We performed a robustness check regarding this issue. Out of the nine wave comparisons, four waves indeed showed significant differences between correlations of interviews with and without an interviewer switch. However, the main results did not change when excluding these interviews.

correlation between the interviewer-level indicators, slight changes in interviewer behavior of the falsifiers over time result in only very small impacts on correlations.

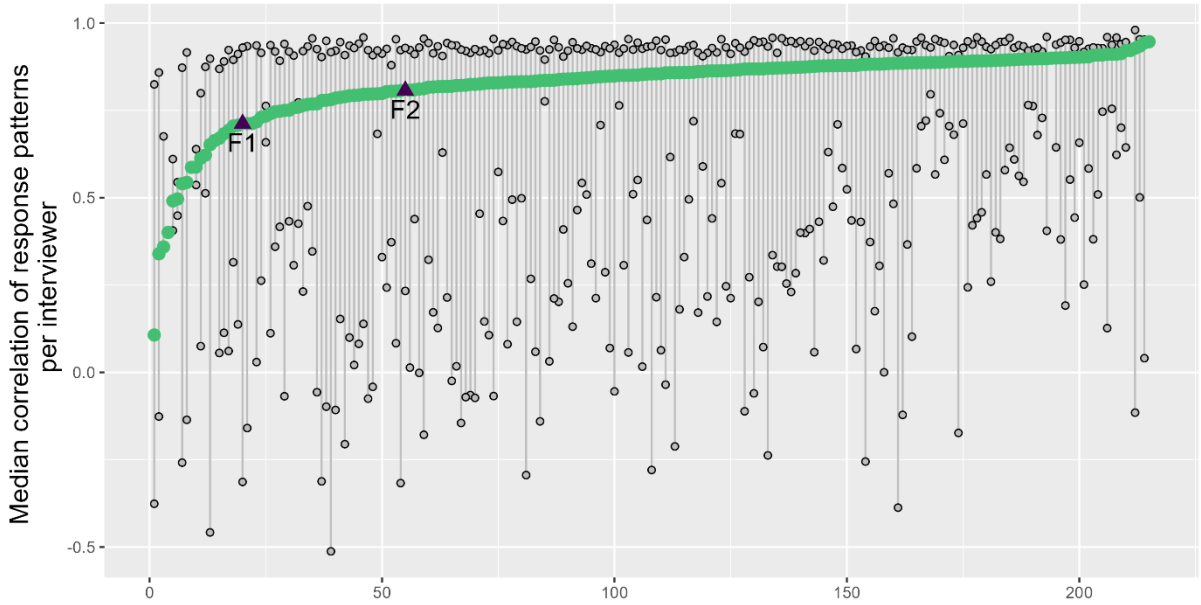

Figure 3.13: Aggregated results (median correlation between response patterns) per interviewer wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: Interviewers are ranked and ordered based on their resulting median, ranging from 1 (most suspicious) to 215 (least suspicious).

3.5.2.3 Correlations Between Items

Lastly, we turn to the analysis of correlations between relatively time-stable single items. In total, correlations were calculated for values of 20 items between waves 14 and 15 (a full overview of the 20 lowest correlations for each item can be found in Appendix **Table A 3.6**). These items include questions regarding satisfaction with health, living situation, and life in general, social participation, social standing, marital status, frequency of doctor visits, disability and health limitations, sports per week, type of health insurance, number of contacts outside the household, frequency of conflicts in the household, activities in unions and parties or other institutions, and trust in other people (for further information, see Appendix **Table A 3.7**).

We calculated the average overall correlation across all waves for the items to ensure that we only include time-stable items (see Appendix **Table A 3.8**). All correlations lie on average above 0.3, i.e., are moderate or strong. For the correlations between waves 14 and 15, 18 items have overall correlations above 0.5, i.e., are strong. For 8 of these items, either F1, F2, or both falsifiers were among the 20 interviewers with the lowest correlations. Hence,

correlations between items appear to be slightly more reliable to identify suspicious interviewers compared to correlations between indicators and response patterns—however, they also include a high number of false-positive cases. Further, one can hardly determine whether the falsifiers were labeled as unsuspicious for some items because they did not falsify data for these specific items or because the results of the detection method are unstable.

To summarize the results over all items, we calculated the mean correlation between the items. As shown in **Figure 3.14**, F1 (rank 23) and F2 (rank 25) generally show lower correlations compared to most interviewers, however, they are not clear outliers. Only five interviewers are clear outliers compared to the other interviewers, also lying below a threshold of 0.5 which is generally considered to be a low correlation. However, throughout all waves, F1 and F2 are repeatedly among the interviewers with lower overall correlations (see Appendix **Figure A 3.16**). Further robustness checks, only including items with strong correlations, give exactly the same results regarding F1 and F2 (Appendix **Figure A 3.17** and **Figure A 3.18**).

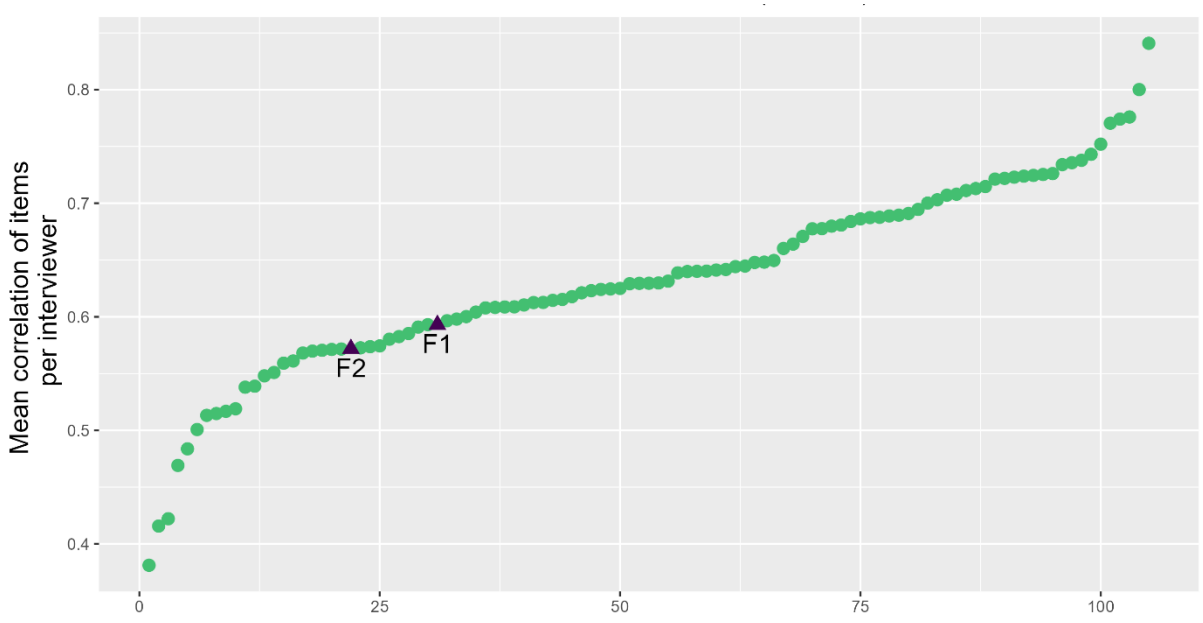

Figure 3.14: Mean correlations between items per interviewer, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: Interviewers are ranked and ordered based on their resulting median, ranging from 1 (most suspicious) to 105 (least suspicious).

3.5.3 Summary of Results

In summary, we find mixed results for wave 15 and rather small differences between waves 14 and 15 (see **Table 3.5**). In particular, the longitudinal analysis yielded fewer promising results. In comparison, the cross-sectional analyses performed well. Even though

some of the tested methods did not help in identifying the partial falsifiers (such as duplicate analysis or Principal Component Analysis for most item batteries), established tools like falsification indicators and cluster analysis performed well for single falsifiers. Particularly noteworthy were the results of the Isolation Forest algorithm (IsoForest), which was easy to implement and delivered good results. These results confirm earlier findings by Jebreel et al. (2020).

Table 3.5: Overview of performance for different analyses, wave 15 or between waves 14 and 15.

Method	**Level of Analysis**	**Identified**		**False-Positives**
		F1	**F2**	
Cross-Sectional Analysis (wave 15)				
Falsification Indicator: AVER	Interviewer-Level	Yes	No	12
Falsification Indicator: ACQ	Interviewer-Level	Yes	Yes	14
Falsification Indicator: MRS	Interviewer-Level	Yes	No	14
Falsification Indicator: INR	Interviewer-Level	No	No	0
Falsification Indicator: ROUND	Interviewer-Level	No	No	13
Falsification Indicator: DUR	Interviewer-Level	Yes	Yes	7
Average Linkage	Interviewer-Level	No	Yes	6
Complete Linkage	Interviewer-Level	No	Yes	2
Single Linkage	Interviewer-Level	No	Yes	6
Ward's Linkage	Interviewer-Level	No	Yes	26
IsoForest Respondent data	Respondent-Level	No	Yes	12
IsoForest Indicator Data	Respondent-Level	Yes	Yes	12
Duplicate Analysis	Respondent-Level	No	No	0
Near-Duplicates	Respondent-Level	Yes	Yes	12
PCA Life Satisfaction	Item-Level	No	No	12
PCA Work-Life Balance	Item-Level	No	No	12
PCA Leisure Activities	Item-Level	Yes	Yes	12
Longitudinal Analysis (waves 14-15)				
Correlations between Indicators	Interviewer-Level	No	No	22
Correlation between Resp-Patterns	Respondent-Level	No	No	11
Correlations between Items	Item-Level	No	No	0-36*

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).
Note: Number of false-positive cases differs for different items.

However, these results do not necessarily hold true for all other waves. Most analyses illustrate that F1 and F2 became increasingly suspicious over time. This could indicate a learning effect by the interviewers—the more experienced they became, the more they falsified and the riskier or bolder their deviant behavior became (see Olbrich et al. 2023). This slow expansion of deviant behavior might also explain the poor performance of the longitudinal

analysis. A more sudden switch in falsification behavior would have been more likely to be detected by these methods. Further, these results highlight the importance of applying a variety of different identification methods, also to avoid a large number of false-positive cases. Except for F1 and F2, very few interviewers were identified by multiple methods. Out of the twenty methods used, F1 was identified by seven methods, F2 by ten methods, and 29 additional interviewers were identified by three or more methods, but only four additional interviewers were identified by six or more methods (see Appendix **Table A 3.9** for a detailed overview).

3.6 Discussion

3.6.1 Main Findings

Although an increasing number of studies have recently addressed the issue of interviewer falsification and proposed a variety of data-driven methods for identifying such cases, two questions have so far been neglected. First, how do survey researchers detect interviewer falsification in panel studies and, second, how do they identify harder-to-detect forms of falsification such as partial falsification? We addressed these questions using data from a large-scale panel survey in Germany, which included verified cases of partial falsifications over multiple waves. In line with the literature (Bredl, Winker, and Kötschau 2012; Menold et al. 2013; de Haas and Winker 2016; Schwanhäuser, Sakshaug, and Kosyakova 2022), we found that falsification indicators as well as their multivariate combination—originally proposed for cross-sectional settings—are also useful for identifying partial falsifications in panel data. Furthermore, we found that (partial) falsifiers indeed produce a higher number of "near-duplicates" (Koczela et al. 2015; Kuriakose and Robbins 2016), but not complete duplicates. However, focusing on duplicated response patterns across item batteries using Categorical Principal Component Analysis (Blasius and Thiessen 2013) did not clearly reveal the falsifiers. In addition, we did not find evidence to corroborate the common assumption that falsifiers are straightforward to detect in a longitudinal setting due to differences in response behavior between waves (Schäfer et al. 2004b; Schräpler and Wagner 2005). Neither correlations between falsification indicators or response patterns, nor correlations between the same items across waves were successful in identifying the falsifiers.

3.6.2 Strengths, Limitations, and Future Work

Our study enriches the literature on interviewer falsification by focusing on partial interviewer falsifications in panel survey data. The verified falsifiers in the data showed different behavior in terms of how they falsified the data and regarding the strategies they used

to fabricate. Further, the availability of multiple waves allowed us to carry out the analyses on a large scale. Hence, we were able to evaluate various data-driven detection methods under different scenarios and exploit the panel data structure, which is rarely done in the literature. Lastly, we contribute to the literature by applying and evaluating a variety of data-driven methods where most other applications solely focus on one method.

However, the study does not come without limitations. The results may depend on the specific falsification behavior observed. Falsifiers who skip complete item batteries and fabricate all the respective answers might be easier to identify than interviewers that neglect other rules of standardized interviewing, e.g., by deviating from the scripted questions and answer categories, inappropriately probing, or including their own interpretations. This also applies to falsifiers who mostly follow the rules but falsify only a few interviews among their workloads and are therefore harder to detect than the confirmed falsifiers from this study. Thus, the results only give insights into a special falsification situation for one specific survey. Furthermore, as PASS interviewers often work in specific regional areas without an interpenetrated design, results could be distorted by this feature of the survey design.

We encourage the replication of these results using other survey data. As this is a case study using one specific dataset, we acknowledge that the methods need to be tested on other datasets to gain a better understanding about when and under which circumstances the demonstrated methods are likely to perform well. Experimental data which manipulates the share of falsified item batteries or the number of falsified interviews within an interviewer's workload might be particularly useful. Such an approach would also overcome our limitation of missing knowledge about which specific item batteries were falsified. In our setting, single falsified interviews or falsified parts of the interviews might have gone undetected, even though data quality was thoroughly tested in the PASS data. As the threshold between partial falsification and other relatively minor deviant behaviors or interviewer errors is often fluid, it is difficult to rule out this possibility. This could lead to biased results regarding the number of false-positive cases, as these cases could indeed include falsifications—therefore, making them true-positive cases.

Lastly, regarding the methods, we encourage the evaluation of new and innovative data-driven detection approaches. For example, since the stability of answers between adjacent waves has not been studied in much detail and the correlations did not perform well on the PASS data, we additionally suggest that alternative stability coefficients should be used and

tested. Besides the stability of answers, further methods able to identify illogical combinations of answers could also be developed to make additional use of the panel data structure.

3.6.3 Practical Implications

In summary, we encourage the use of data-driven falsification controls and data quality checks to enhance survey data quality. Data-driven methods like falsification indicators can aid in uncovering different forms of interviewer falsification (e.g., complete or partial falsification). Furthermore, these methods are not only useful in the context of cross-sectional data but also for longitudinal data. As we have shown, different methods showed different results for the two confirmed falsifiers. Hence, each method might be able to identify different types of falsification behavior, which is also in line with recent findings from the literature (Olbrich et al. 2023). The combination of different methods therefore creates a data quality safety net that flags as many suspicious cases as possible increasing the likelihood of flagging actual falsifiers, with the cases flagged by multiple methods being obvious candidates for further investigation. While this approach might increase the number of false-positive cases and lead to increased costs (e.g., for more re-interviews or checks of audio recordings), these added costs might still be relatively small compared to the potential consequences for data quality and reputation of the survey institute if the falsifications are discovered too late. Further, monitoring of interviewers over subsequent waves of a panel study could also provide further insights into potentially undesirable behavior of interviewers. However, as the results have shown, well-known and easily applied methods like falsification indicators were also successful in identifying the partial falsifications. Hence, these simpler tools might already be sufficient to detect different forms of interviewer falsification, independent of the data structure (cross-sectional or longitudinal). In the end, practitioners must decide on the methods and the effort they want to invest into their quality controls based on a cost-benefit calculation for their particular use case.

Appendix

Table A 3.1: Description of falsification indicators.

Indicator	Description
Acquiescent Responding (ACQ)	Proportion of content-independent, affirmative responses ("Agree or Strongly Agree") in rating scales
Middle Responding Style (MRS)	Proportion of middle responses in rating scales
Item Nonresponse (INR)	Proportion of item nonresponse in closed-ended questions
Rounding Tendency (ROUND)	Proportion of rounded numbers in numeric open-ended questions
Interview Duration (DUR)	Duration of an interview

Table A 3.2: Mean (standard deviation) of falsification indicators on the interviewer-level (N), waves 7 to 15.

	Wave 7	Wave 8	Wave 9	Wave 10	Wave 11	Wave 12	Wave 13	Wave 14	Wave 15
ACQ	0.46 (0.17)	0.46 (0.18)	0.47 (0.16)	0.42 (0.17)	0.66 (0.15)	0.60 (0.17)	0.56 (0.18)	0.42 (0.17)	0.49 (0.14)
MRS	0.49 (0.15)	0.53 (0.14)	0.45 (0.18)	0.50 (0.14)	0.48 (0.17)	0.59 (0.15)	0.64 (0.15)	0.51 (0.21)	0.52 (0.15)
INR	0.83 (0.16)	0.87 (0.14)	0.89 (0.13)	0.86 (0.17)	0.91 (0.11)	0.88 (0.14)	0.90 (0.13)	0.84 (0.20)	0.88 (0.15)
ROUND	0.25 (0.13)	0.27 (0.15)	0.45 (0.14)	0.28 (0.15)	0.29 (0.15)	0.44 (0.15)	0.30 (0.14)	0.37 (0.22)	0.29 (0.15)
DUR	0.74 (0.15)	0.69 (0.16)	0.73 (0.13)	0.94 (0.07)	0.61 (0.17)	0.67 (0.14)	0.90 (0.07)	0.69 (0.14)	0.56 (0.13)
AVER	0.55 (0.06)	0.57 (0.06)	0.60 (0.07)	0.60 (0.06)	0.59 (0.07)	0.64 (0.07)	0.66 (0.07)	0.57 (0.09)	0.55 (0.07)
N	237	253	259	237	246	258	229	129	226

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

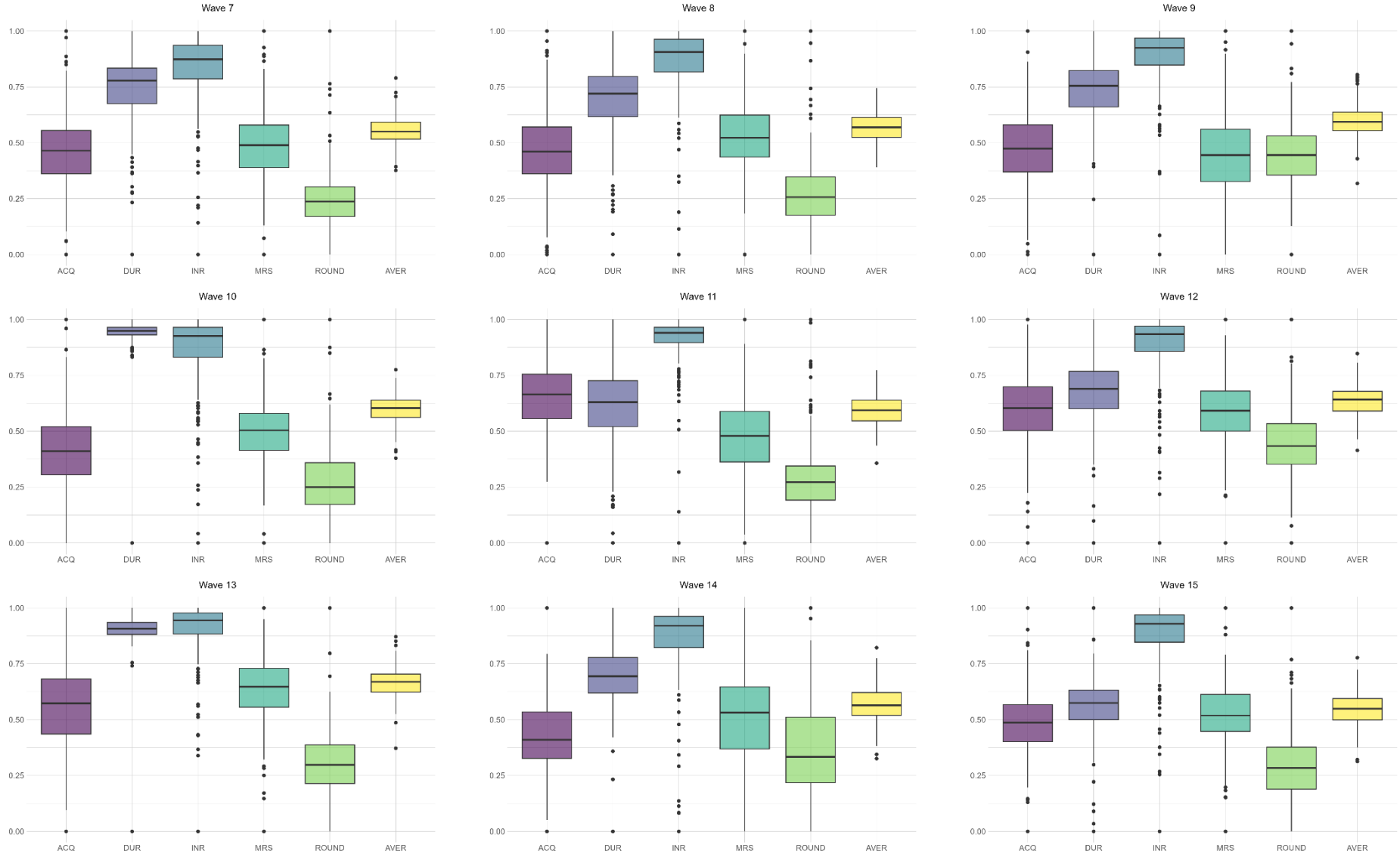

Figure A 3.1: Boxplot of falsification indicators per wave, waves 7 to 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Table A 3.3: Indicator values of the 10 highest/lowest ranked interviewers according to the average across all indicator values (AVER), wave 15.

Rank	Interviewer ID	ACQ	MRS	INR	ROUND	DUR	AVER
1	F1	0.84	0.78	1.00	0.41	0.86	0.78
2	I729	0.81	0.72	0.88	0.50	0.72	0.72
3	I625	0.79	0.79	0.97	0.53	0.51	0.72
4	I164	0.56	0.66	0.94	0.70	0.63	0.70
5	I735	0.65	0.79	1.00	0.51	0.52	0.69
6	I333	0.70	0.71	0.97	0.37	0.70	0.69
7	I381	0.83	1.00	0.95	0.08	0.52	0.68
8	I291	0.58	0.77	1.00	0.40	0.62	0.68
9	I420	0.71	0.64	0.90	0.40	0.72	0.67
10	I592	1.00	0.91	0.81	0.28	0.37	0.67
216	I527	0.00	0.24	0.99	0.26	0.62	0.42
217	I678	0.44	0.38	0.44	0.35	0.49	0.42
218	I536	0.20	0.47	0.79	0.08	0.54	0.42
219	I215	0.24	0.32	0.87	0.38	0.22	0.41
220	F2	0.69	0.00	0.00	0.47	0.86	0.40
221	I509	0.36	0.24	0.92	0.16	0.30	0.40
222	I780	0.28	0.37	0.83	0.05	0.45	0.39
223	I153	0.37	0.20	0.89	0.33	0.09	0.38
224	I788	0.14	0.18	0.79	0.06	0.43	0.32
225	I452	0.28	0.21	0.86	0.21	0.00	0.31

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

Table A 3.4: Anonymized interviewer IDs for the 25 highest average falsification indicator values (AVER), waves 7 to 15.

Rank	Wave 7	Wave 8	Wave 9	Wave 10	Wave 11	Wave 12	Wave 13	Wave 14	Wave 15
1	I614	I695	I876	I607	I621	I711	I65	I366	F1
2	I171	I625	I51	I625	I1181	I188	I226	I651	I729
3	I68	I856	I703	I171	I903	I265	I735	I634	I625
4	I15	I171	I1181	I331	I1169	I171	I625	F1	I164
5	I607	I705	I157	I226	I145	I607	I363	I420	I735
6	I876	I621	I827	I239	I165	I1180	I239	I441	I333
7	I148	I876	I171	I651	I226	I604	I592	I318	I381
8	I856	I1156	I115	I663	I625	I226	I290	I605	I291
9	I625	I641	I591	I41	I663	I291	I249	I501	I420
10	I164	I599	I735	I735	I735	I26	I76	I239	I592
11	I887	I169	I141	I188	I1180	I333	I607	I188	I187
12	I621	I649	I143	I143	I831	I441	I297	I356	I786
13	I135	I129	I604	I51	I604	I283	I348	I298	I301
14	I1169	I602	I625	I169	I235	I62	I188	I157	I571
15	I604	I215	I188	I605	I341	I866	I422	I220	I145
16	I62	I699	I621	I341	I866	I76	I558	I321	I634
17	I573	I135	I79	I621	I679	I625	I100	I155	I129
18	I902	I558	I147	I604	I171	I681	I1180	I431	I239
19	I827	I658	I62	I145	I62	I735	I41	I866	I209
20	I133	I62	I573	I349	I355	I612	I318	I905	I546
21	I188	I707	I146	I1169	I566	I100	I321	I592	I831
22	I831	I252	I291	I141	I370	I16	I634	I100	I1180
23	I154	I651	I592	I612	I76	I704	I268	I583	I605
24	I122	I188	I622	I321	I239	I32	I441	I129	I468
25	I566	I566	I831	I137	I298	I239	I594	I76	I259

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

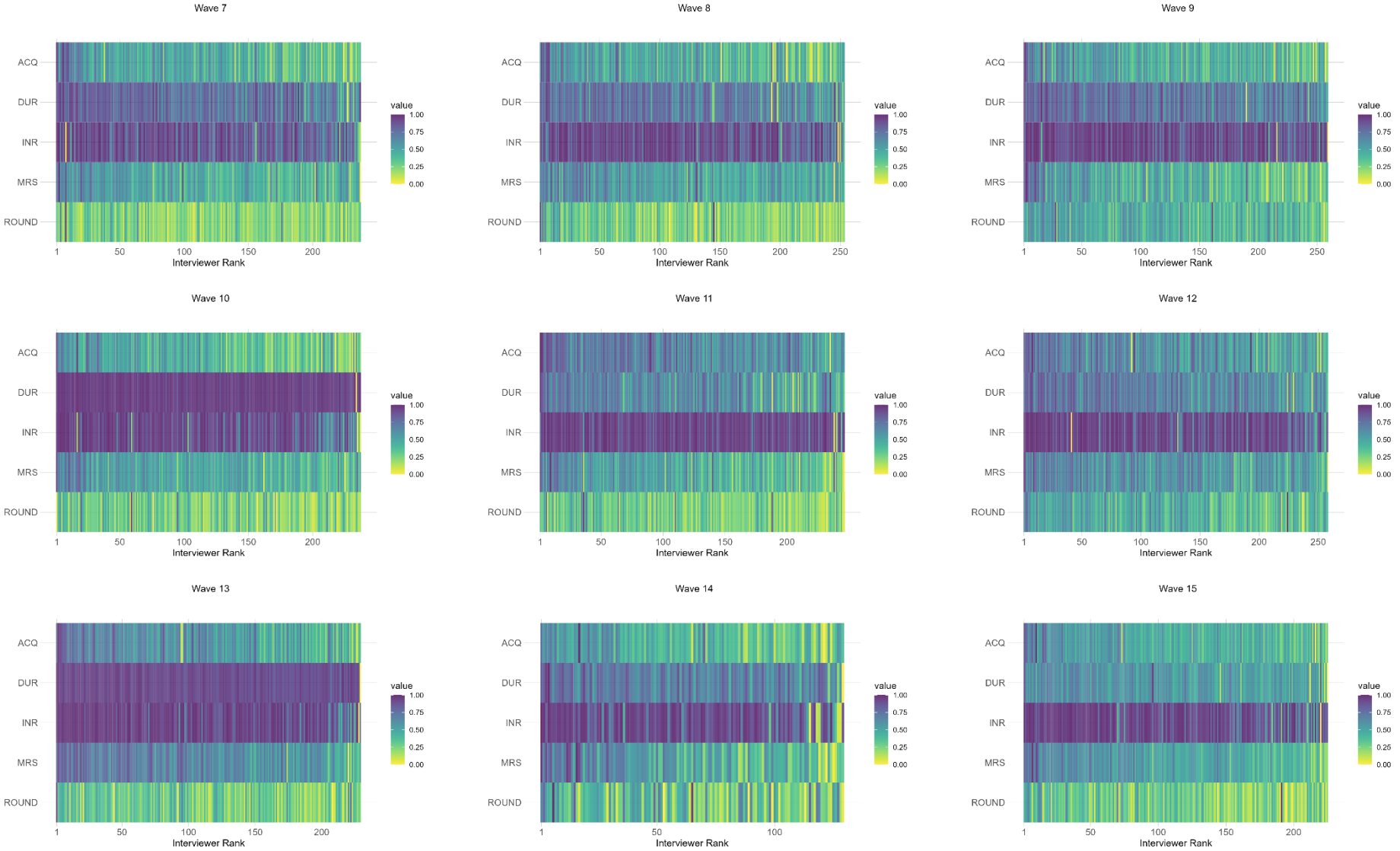

Figure A 3.2: Heat map of falsification indicators per wave ordered by average falsification indicator (AVER), waves 7 to 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

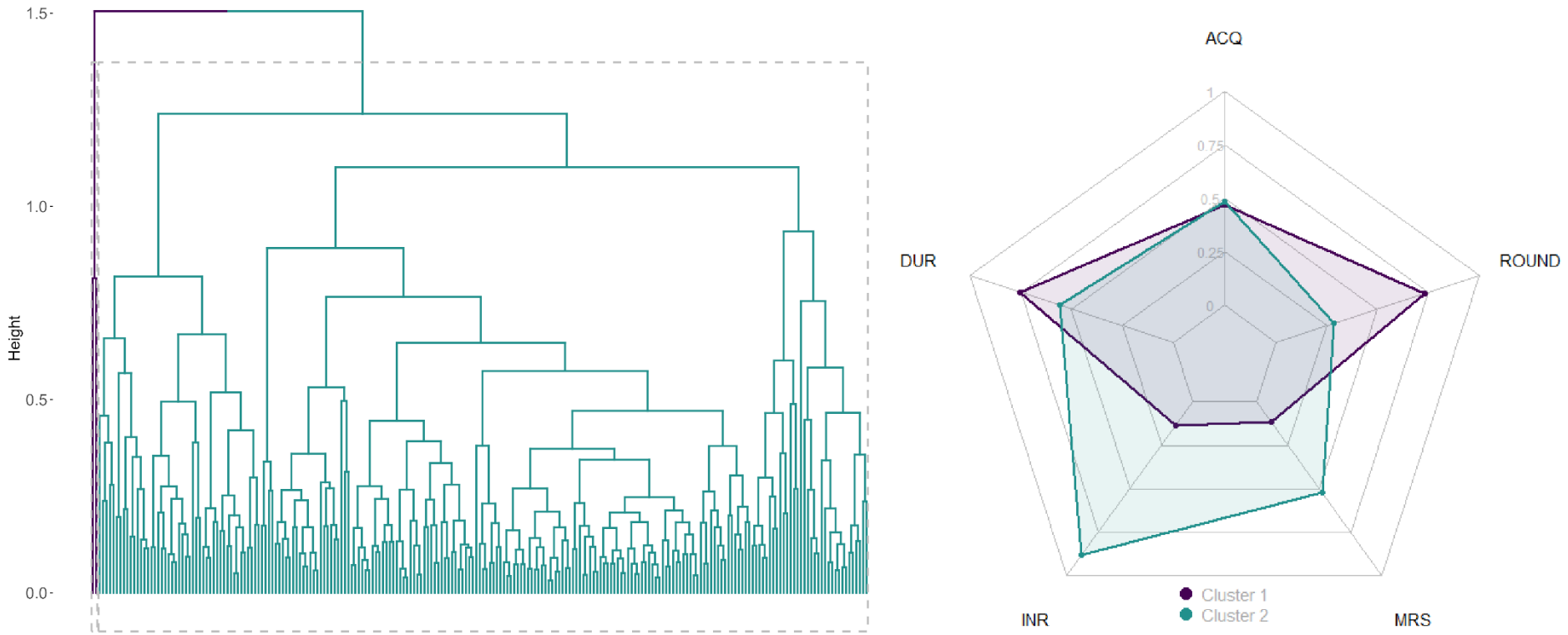

Figure A 3.3: Dendrogram for 2-cluster solution of Complete Linkage and radar plot for cluster, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

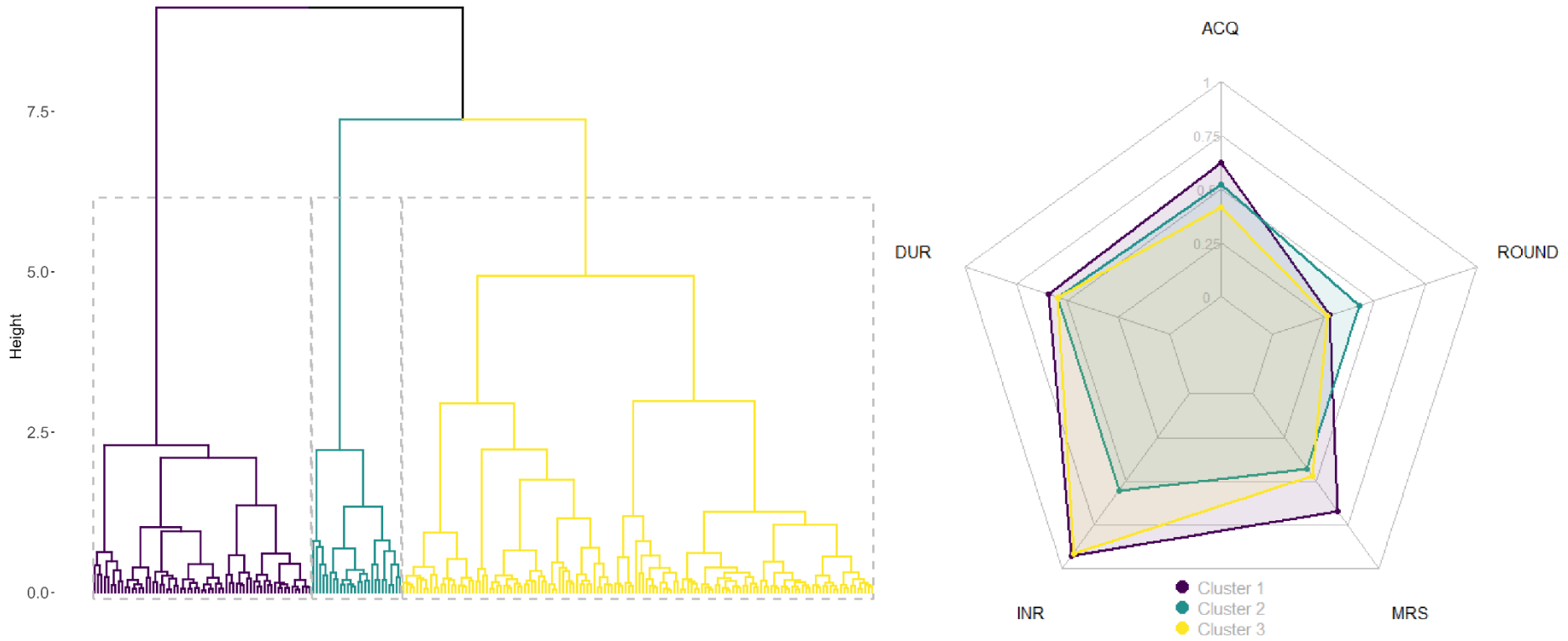

Figure A 3.4: Dendrogram for 3-cluster solution of Ward's Linkage and radar plot for cluster, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

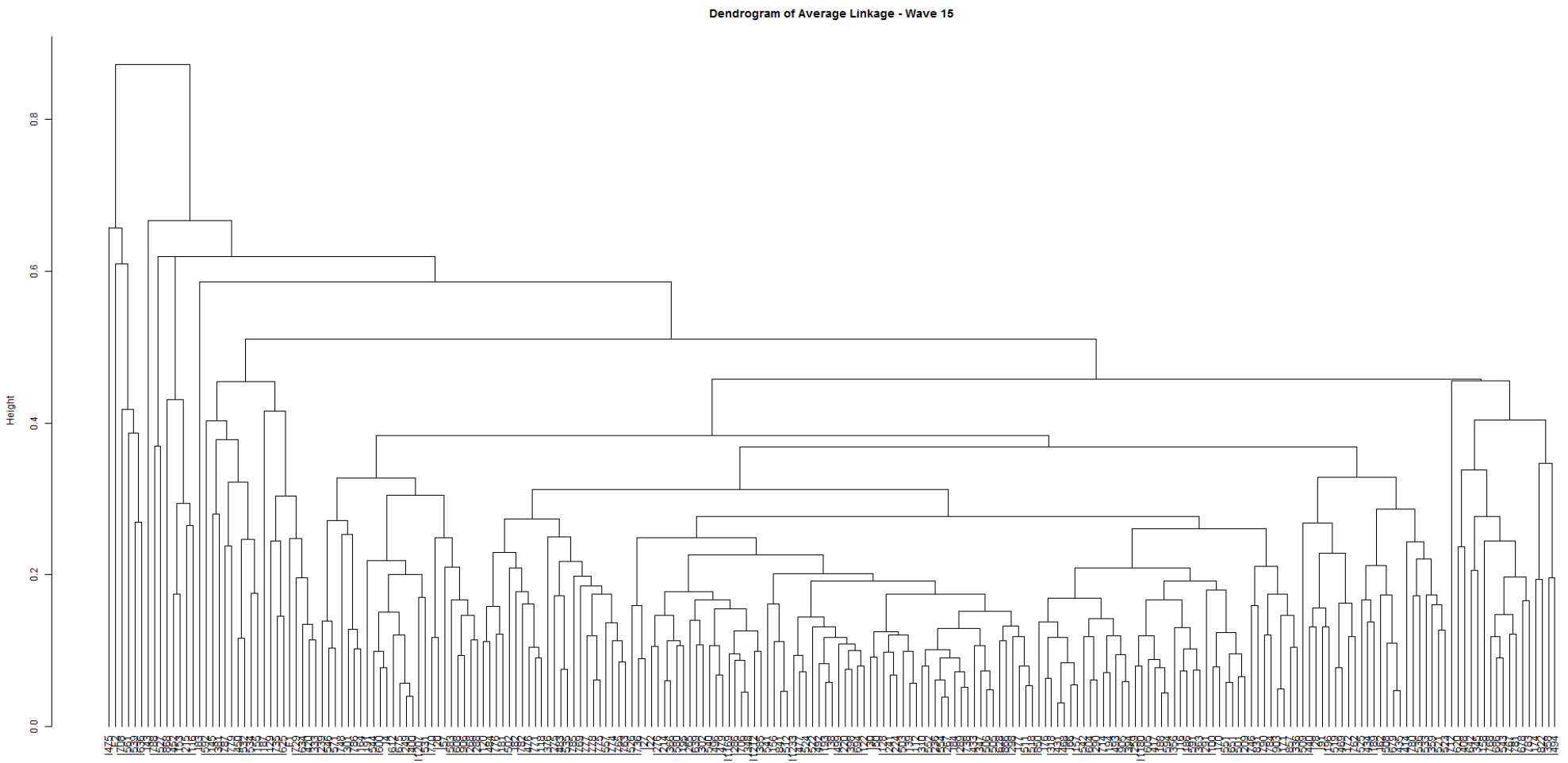

Figure A 3.5: Full dendrogram of Average Linkage, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

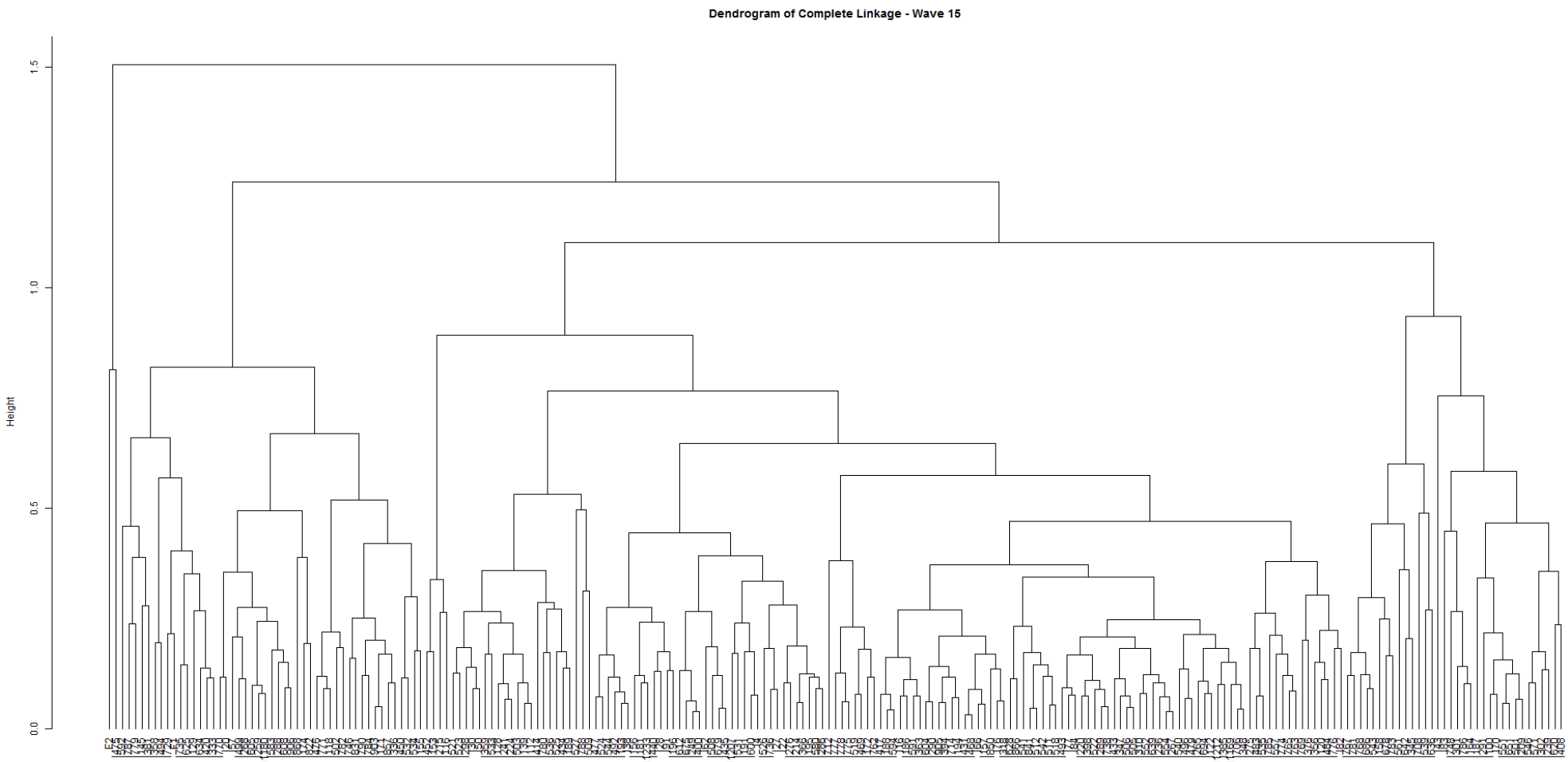

Figure A 3.6: Full dendrogram of Complete Linkage, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

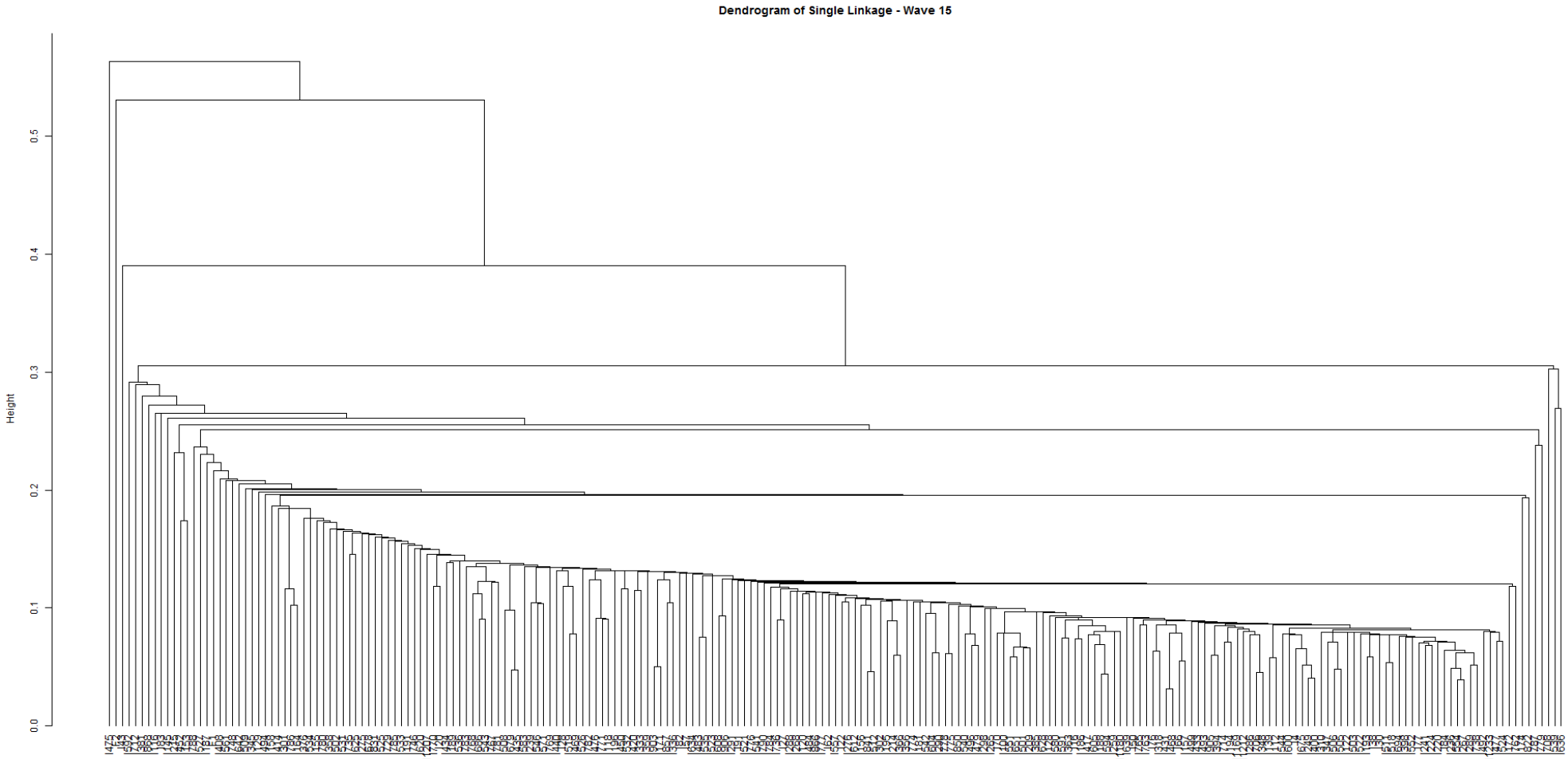

Figure A 3.7: Full dendrogram of Single Linkage, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

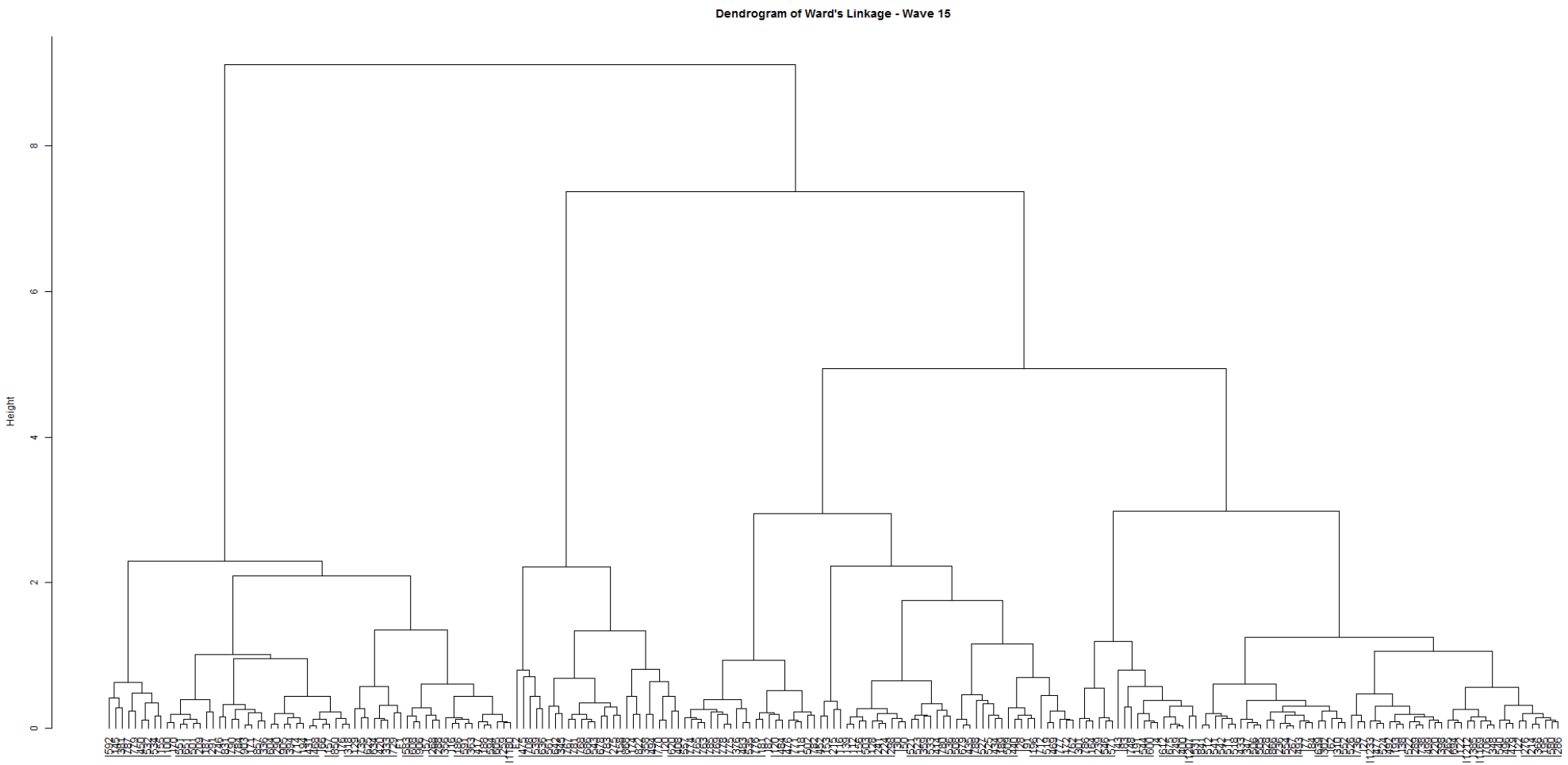

Figure A 3.8: Full dendrogram of Ward's Linkage, wave 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

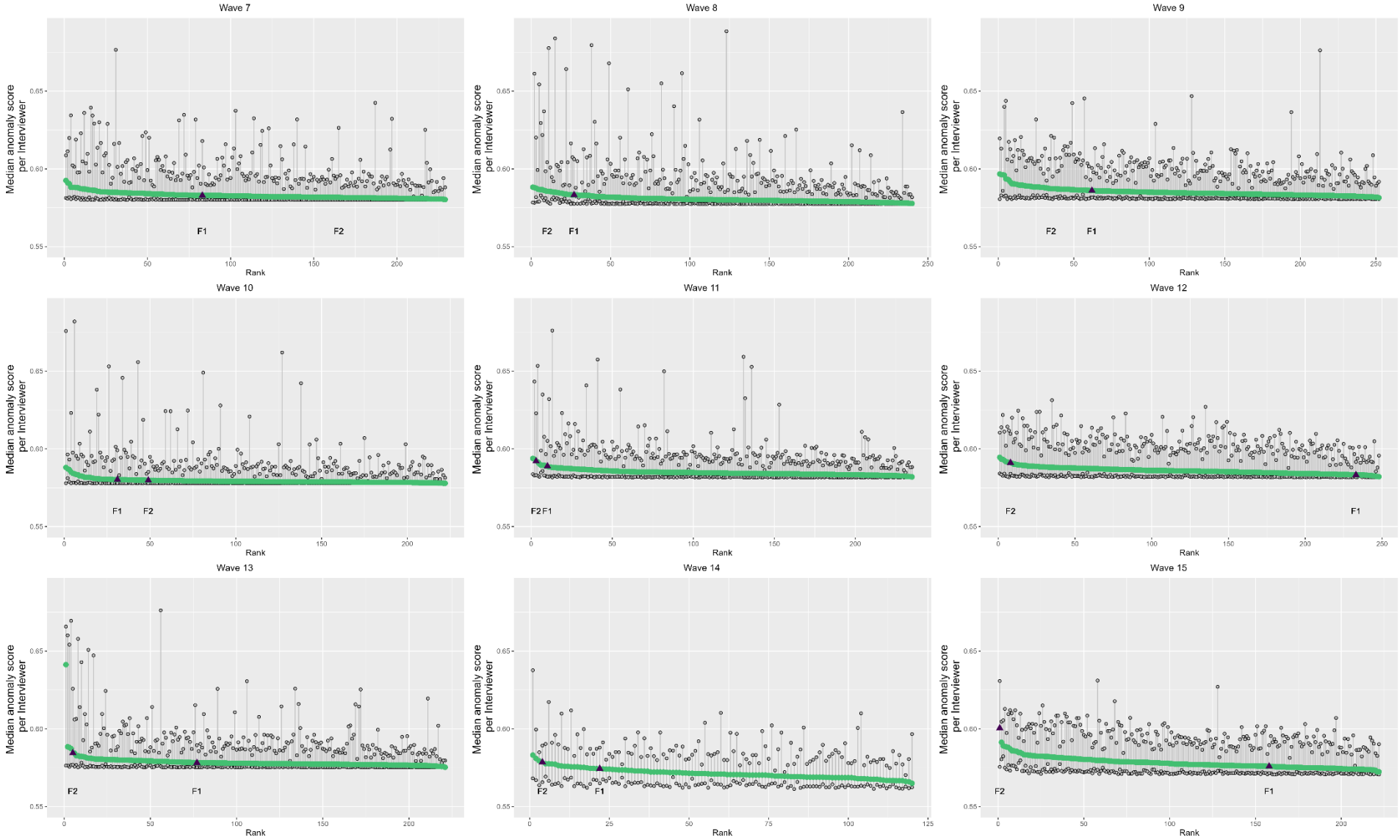

Figure A 3.9: Aggregated results (median anomaly score) per interviewer for response data, waves 7 to 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting median.

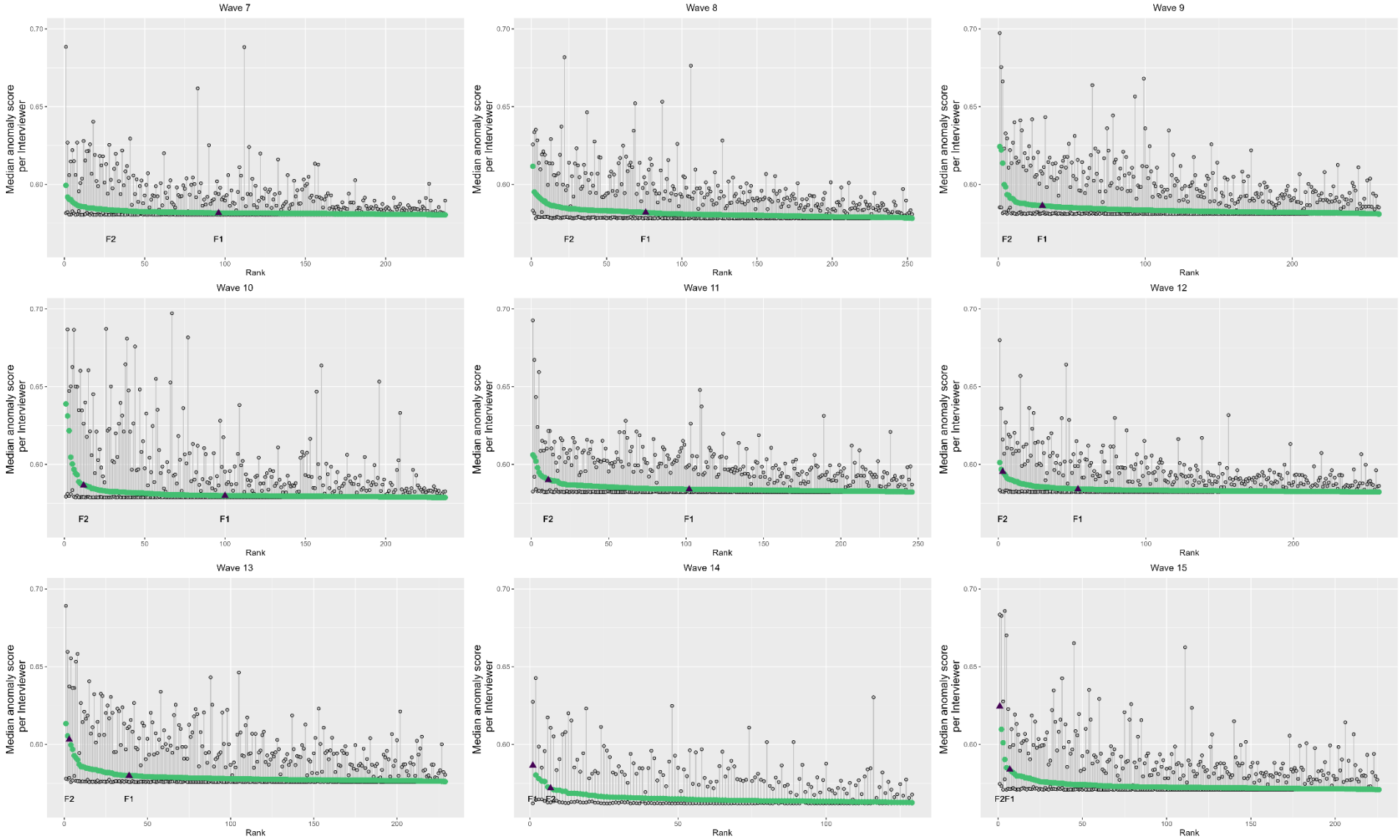

Figure A 3.10: Aggregated results (median anomaly score) per interviewer for indicator data, waves 7 to 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting median.

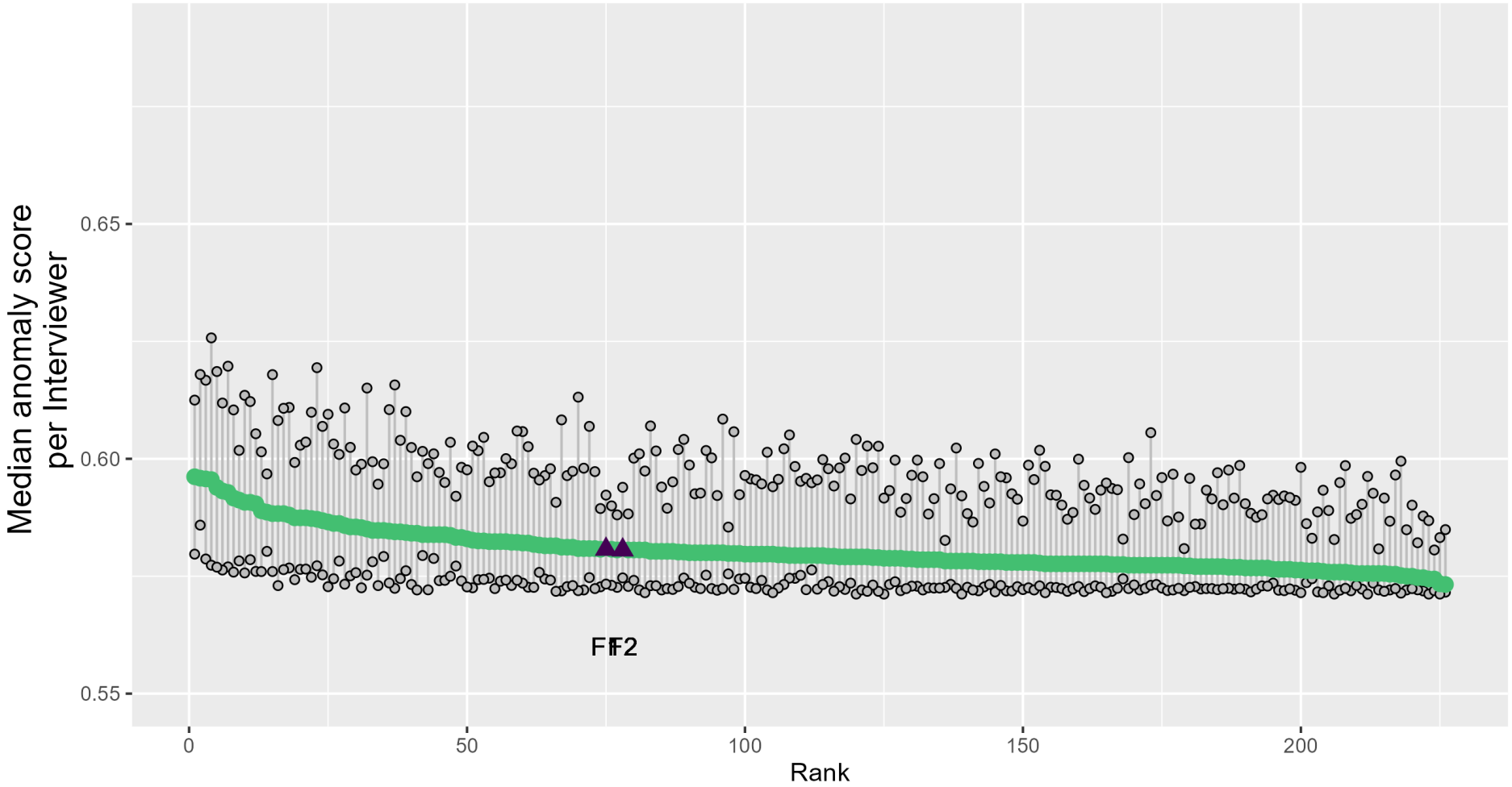

Figure A 3.11: Aggregated results (median anomaly score) per interviewer for response data, wave 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting median. Missing values in the response data were recorded as mean value.

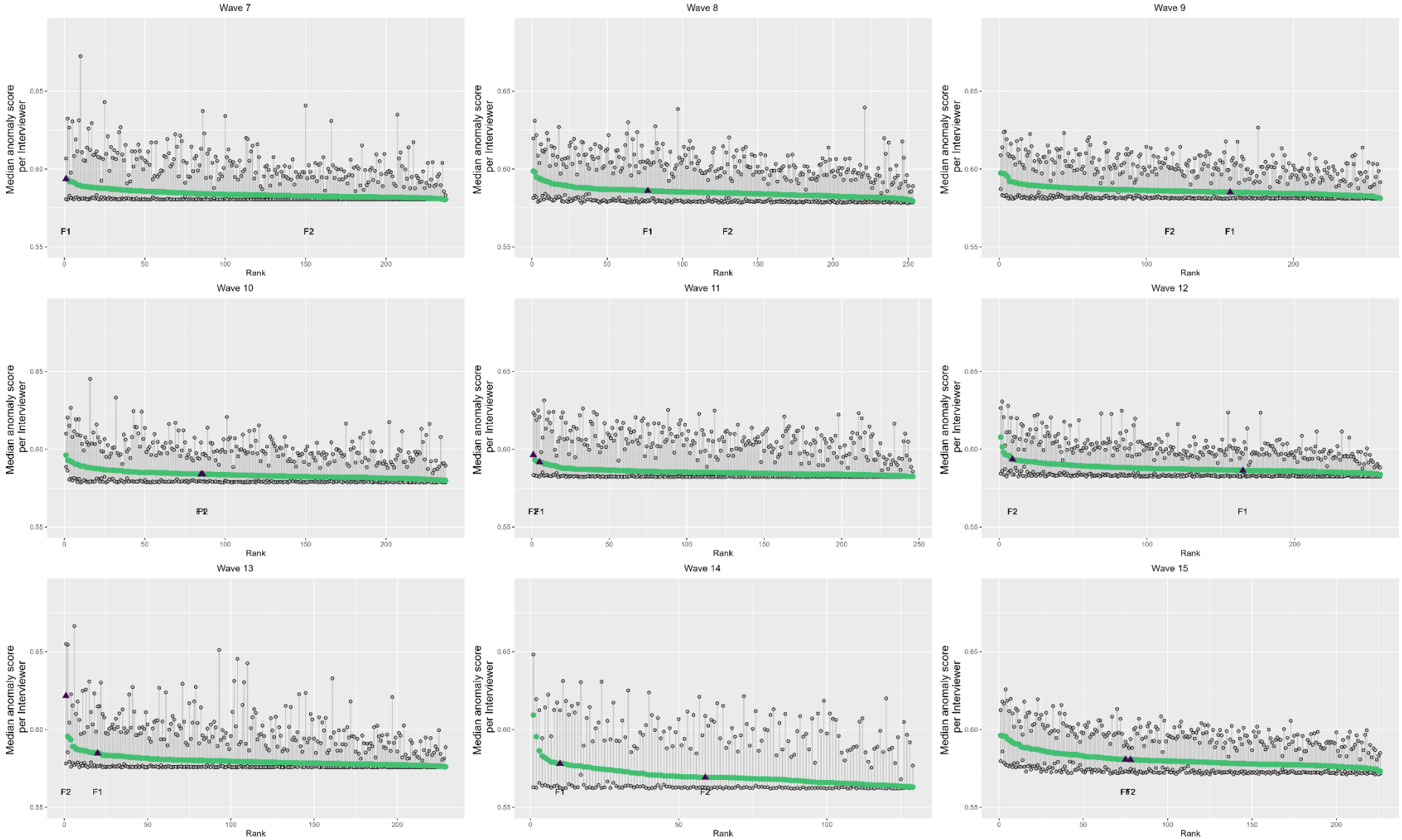

Figure A 3.12: Aggregated results (median anomaly score) per interviewer for response data, waves 7 to 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting median. Missing values in the response data were recorded as mean value.

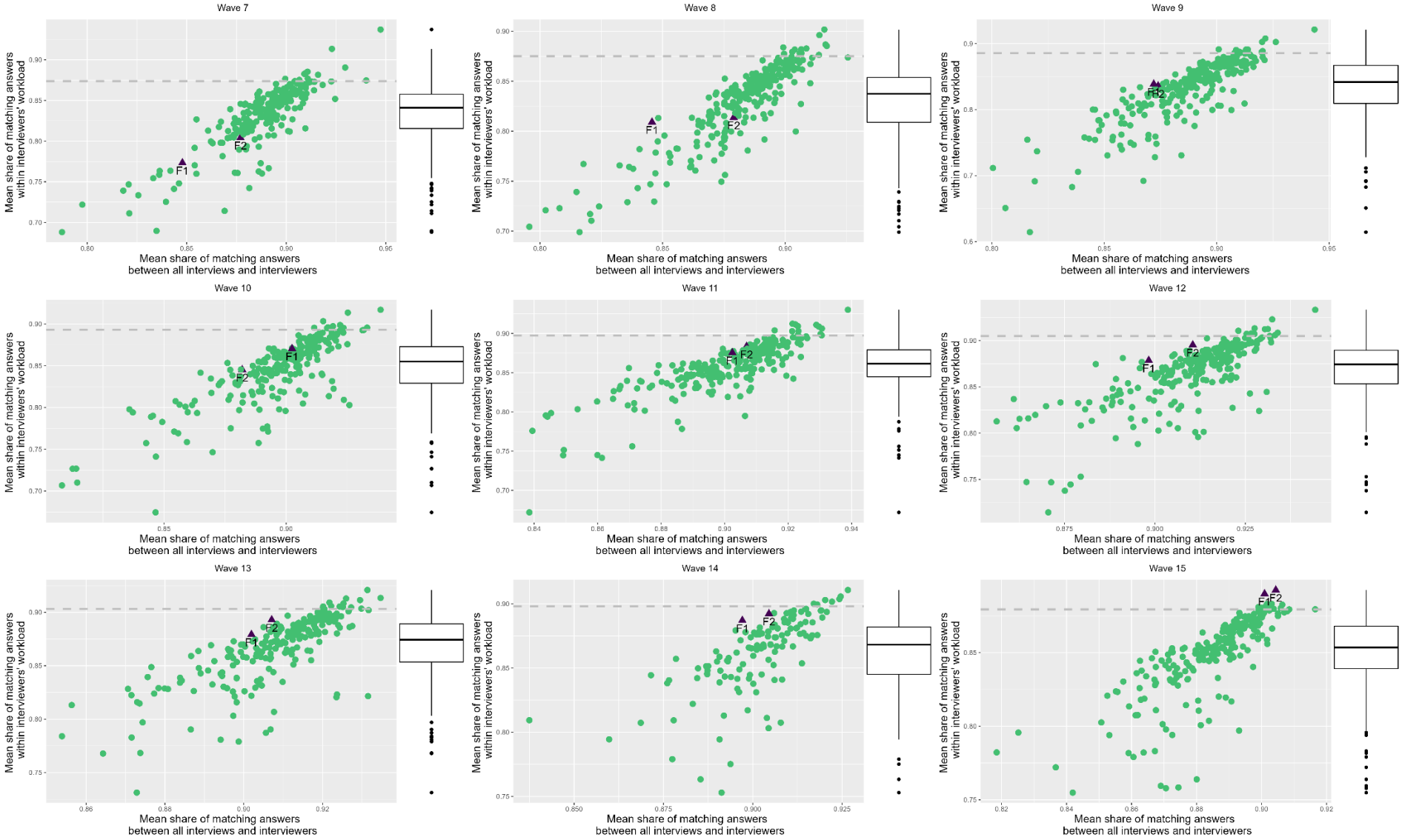

Figure A 3.13: Mean share of matching answers within interviewers' workload and between all interviews, waves 7 to 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: The dashed line denotes the 95th percentile of the mean share of matching answers within interviewers' workload.

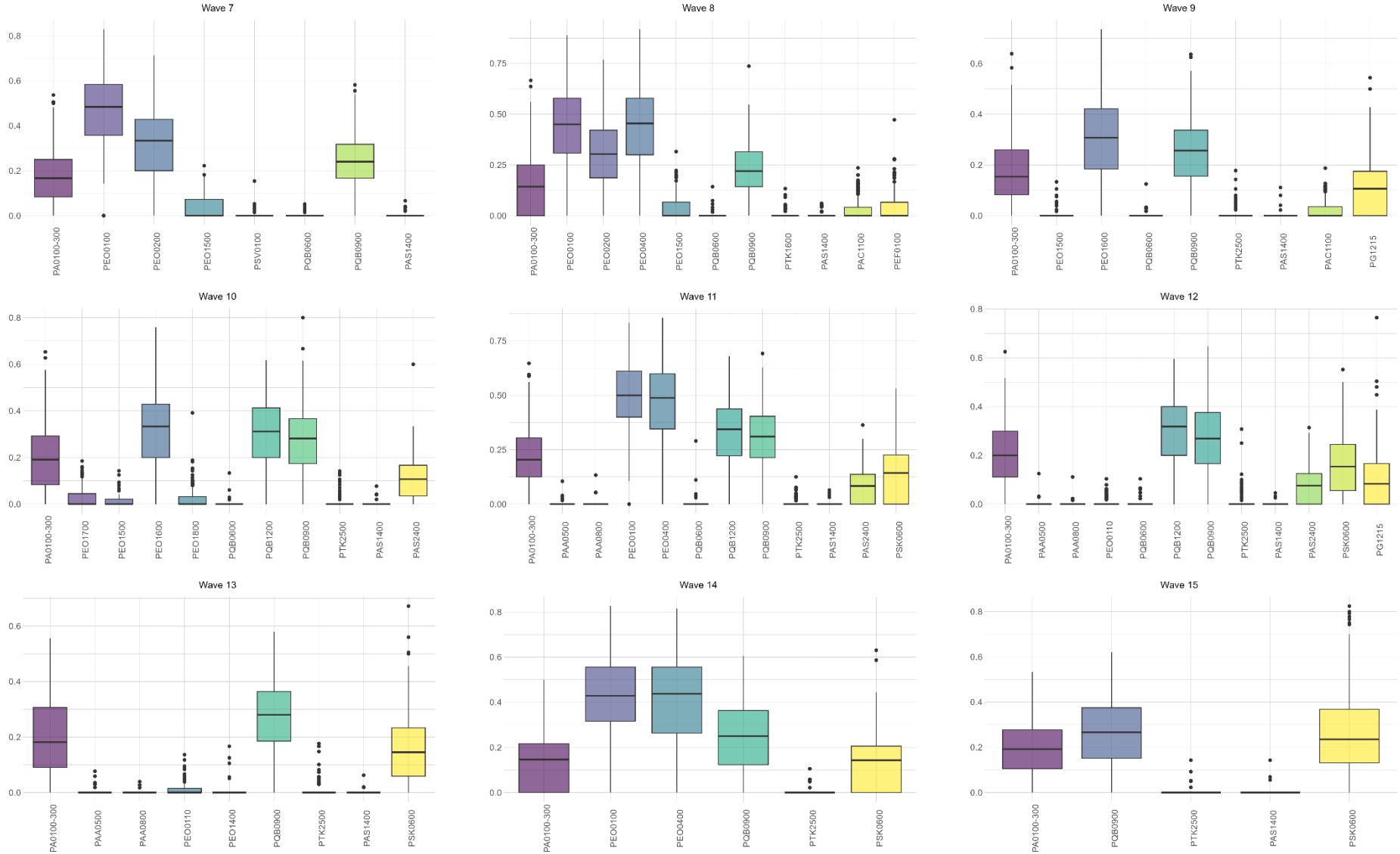

Figure A 3.14: Boxplots of the share of duplicated factor scores per interviewer for different item batteries, waves 7 to 15.
Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Table A 3.5: Information on item batteries used to calculate factor scores.

Item Battery	Content	Scale	Wave								
PA0100 to PA0300	Attitudes (satisfaction with areas of life)	1-10	7	8	9	10	11	12	13	14	15
PAA0500a-f	Educational aspiration (parental involvement in career choice)	1-4	-	-	-	-	11	12	13	-	-
PAA0800a-h	Educational aspiration (career choice)	1-4	-	-	-	-	11	12	13	-	-
PEO0100a-e	Attitudes (difficulties/problems)	1-4	7	8	-	-	11	-	-	14	-
PEO0110a-e	Educational aspiration (difficulties/problems)	1-4	-	-	-	-	-	12	13	-	-
PEO0200a-d	Attitudes work (gainful employment)	1-4	7	8	-	-	-	-	-	-	-
PEO0400a-d	Opinion (role/gender models)	1-4	-	8	-	-	11	-	-	14	-
PEO1700a-f	Attitudes (childcare)	1-4	-	-	-	10	-	-	-	-	-
PEO1400a-u	Big 5	1-5	-	-	-	-	-	-	13	-	-
PEO1500a-d	Attitudes (children's leisure time)	1-4	7	8	9	10	-	-	-	-	-
PEO1600a-f	Attitudes (reciprocity)	1-4	-	-	9	10	-	-	-	-	-
PEO1800a-h	Impulsiveness	1-5	-	-	-	10	-	-	-	-	-
PSV0100a-i	Stigma awareness	1-4	7	-	-	-	-	-	-	-	-
PQB0600a-l	Quality of employment (opportunities/pressures)	1-4	7	8	9	10	11	12	-	-	-
PQB1200a-i	Quality of employment (meaning of work)	1-4	-	-	-	10	11	12	-	-	-
PQB0900a-c	Quality of employment (work-life-balance)	1-4	7	8	9	10	11	12	13	14	15
PTK1600a-h	Contacts with unemployment agency (staff in general)	1-4	-	8	-	-	-	-	-	-	-
PTK2500a-h	Contacts with unemployment agency (job center)	1-4	-	-	9	10	11	12	13	14	15
PAS1400a-g	Job search (acceptance of disadvantages)	1-4	7	8	9	10	11	12	13	-	15
PAS2400a-i	Job search (meaning of work)	1-4	-	-	-	10	11	12	-	-	-
PAC1100a-d	Job search (skill assessment)	1-5	-	8	9	-	-	-	-	-	-
PSK0600a-f	Networks (leisure time)	1-5	-	-	-	-	11	12	13	14	15
PG1215a-i	Health (mental/physical health)	1-5	-	-	9	-	-	12	-	-	-
PEF0100a-h	Attitudes (finances)	1-4	-	8	-	-	-	-	-	-	-

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

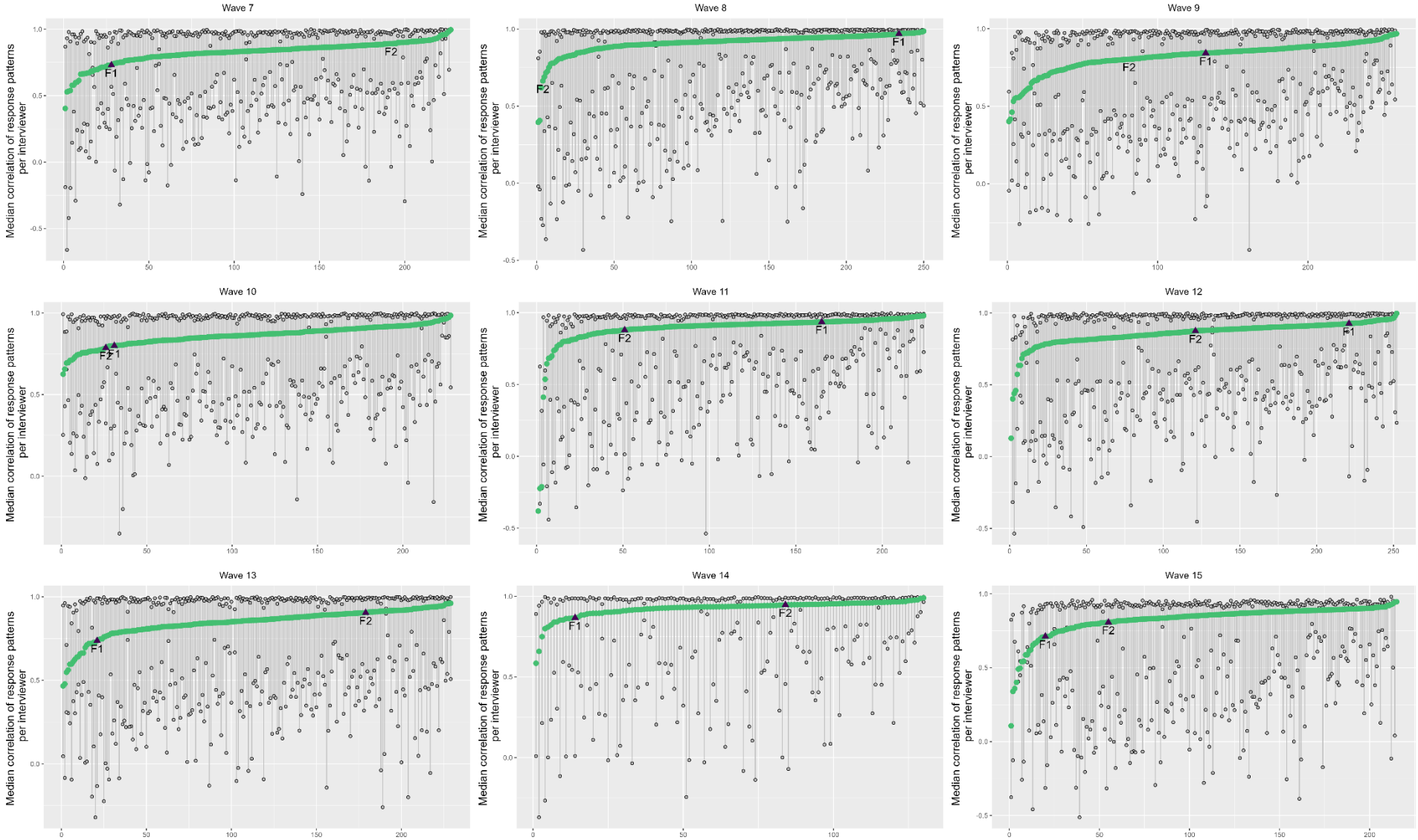

Figure A 3.15: Aggregated results (median correlation between response patterns) per interviewer, waves 7 to 15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting median.

Table A 3.6: Overview of all correlations for the 20 interviewers with the lowest correlation between items, waves 14/15.

Rank	PA0100		PA0200		PA0300		PA0800		PA0900		PA1000		PD0500		PG0100		PG0500	
	Corr	ID	Corr	ID	Corr	ID	Corr	ID	Corr	Corr	Corr	ID	Corr	ID	Corr	ID	Corr	ID
1	-0.02	I100	-0.24	I408	0.09	I831	-0.29	I831	-0.73	-0.73	-0.27	I120	0.10	I440	-0.17	I1169	-0.73	I310
2	0.03	I708	0.09	I345	0.14	I708	-0.15	I345	-0.11	-0.11	0.20	I850	0.53	I634	-0.11	I259	-0.11	I708
3	0.19	I636	0.15	I1207	0.16	I196	-0.13	I358	0.00	0.00	0.20	I868	0.64	I552	-0.10	I729	0.00	I139
4	0.19	I841	0.21	I554	0.19	I408	0.01	I484	0.01	0.01	0.21	I286	0.66	I28	-0.09	I503	0.01	I174
5	0.21	I171	0.22	I905	0.21	I499	0.06	I301	0.10	0.10	0.22	I738	0.70	I290	-0.07	I905	0.10	I394
6	0.22	I310	0.26	I831	0.23	I16	0.07	I188	0.11	0.11	0.23	I220	0.76	I554	-0.06	I57	0.11	I431
7	0.30	I503	0.27	I139	0.27	I158	0.07	I122	0.23	0.23	0.24	I129	0.84	I76	-0.03	I16	0.23	F1
8	0.33	I706	0.28	I1169	0.31	F2	0.11	I16	0.28	0.28	0.25	I484	0.84	I841	-0.01	I100	0.28	I224
9	0.34	I398	0.31	F2	0.31	I286	0.14	I554	0.30	0.30	0.25	I171	0.84	I174	0.01	I822	0.30	I499
10	0.36	I174	0.32	I736	0.32	I122	0.17	I499	0.30	0.30	0.28	I84	0.87	I122	0.02	I1233	0.30	I66
11	0.38	I580	0.33	I286	0.42	I736	0.24	I521	0.31	0.31	0.28	I591	0.89	I286	0.02	I651	0.31	I116
12	0.38	I679	0.33	I748	0.44	I554	0.29	I174	0.31	0.31	0.30	I679	0.89	F1	0.04	I431	0.31	I220
13	0.41	I639	0.34	I358	0.44	I738	0.30	I224	0.36	0.36	0.31	I431	0.90	I301	0.06	I336	0.36	I554
14	0.42	I736	0.35	I521	0.44	I1207	0.30	I729	0.41	0.41	0.34	I831	0.90	I4	0.07	I484	0.41	I345
15	0.42	I408	0.37	I499	0.46	I551	0.30	I158	0.43	0.43	0.34	I905	0.91	I153	0.07	I552	0.43	I28
16	0.44	I591	0.39	I82	0.47	I30	0.31	I551	0.43	0.43	0.36	I408	0.92	I224	0.08	I82	0.43	I679
17	0.44	I158	0.40	I16	0.49	F1	0.32	I129	0.44	0.44	0.36	I196	0.93	I738	0.11	I736	0.44	I82
18	0.44	I431	0.40	I552	0.51	I394	0.32	I290	0.46	0.46	0.37	I554	0.94	I546	0.11	I356	0.46	I1180
19	0.45	I499	0.41	I84	0.51	I310	0.32	I1207	0.47	0.47	0.38	I540	0.95	I748	0.12	I4	0.47	I290
20	0.45	I831	0.41	I196	0.53	I850	0.37	I196	0.50	0.50	0.38	I468	0.96	I605	0.12	F2	0.50	I546

Note: All correlations below 0.3 are highlighted in light gray, negative correlations are highlighted in dark gray.

Table A 3.6 (continued)

Rank	PG0800		PG1235		PG1300		PSK0200		PSK0300		PSK0400a		PSK0400b		PSK0400c		PSK0400d	
	Corr	ID	Corr	ID	Corr	ID	Corr	ID	Corr	ID	Corr	ID	Corr	ID	Corr	ID	Corr	ID
1	-0.15	I594	-0.11	I532	0.14	I16	0.14	I634	-0.23	I174	-0.08	I420	0.65	I301	0.21	I16	-0.13	I186
2	-0.11	I414	-0.05	I129	0.14	I492	0.15	I118	0.00	I551	-0.08	I822	0.67	I499	0.33	I66	-0.06	I174
3	-0.04	I636	-0.02	I612	0.37	I1212	0.17	I174	0.08	I592	-0.04	I706	0.67	I414	0.35	I440	-0.04	I748
4	0.08	I385	0.09	I1180	0.41	I66	0.28	I28	0.15	I209	-0.04	I440	0.68	I736	0.40	I499	0.25	I394
5	0.17	I706	0.09	I499	0.47	I905	0.33	I571	0.17	I188	0.33	I708	0.69	I318	0.44	I76	0.27	I358
6	0.19	I171	0.17	I414	0.54	I554	0.36	I546	0.18	I591	0.35	I224	0.69	I122	0.45	I82	0.31	I1207
7	0.21	I518	0.19	I822	0.60	I259	0.36	I84	0.22	I634	0.41	I158	0.78	I1207	0.46	I651	0.33	I608
8	0.26	I484	0.19	I503	0.67	I503	0.40	I408	0.25	I30	0.41	I499	0.79	I220	0.46	I209	0.34	I706
9	0.28	I679	0.19	I336	0.67	I394	0.40	I394	0.33	I431	0.52	I286	1.00	I139	0.50	I583	0.35	I302
10	0.29	I356	0.20	I84	0.68	I822	0.42	I625	0.33	I608	0.53	I122	1.00	I158	0.53	I239	0.36	I345
11	0.33	I868	0.20	I174	0.69	I82	0.42	I866	0.34	I239	0.61	I28	1.00	I492	0.53	I675	0.41	I116
12	0.35	I736	0.20	I639	0.70	I552	0.44	I116	0.34	I850	0.67	I345	1.00	I153	0.54	I116	0.47	I76
13	0.35	I492	0.20	I841	0.72	I301	0.44	I358	0.36	I414	0.67	I521	1.00	I286	0.54	I868	0.48	F1
14	0.36	I850	0.21	I850	0.78	I592	0.45	I57	0.37	I552	0.67	I518	1.00	I129	0.56	I552	0.48	I868
15	0.39	I841	0.22	I196	0.85	I484	0.47	I675	0.39	I259	0.67	I385	1.00	I679	0.56	I554	0.52	I4
16	0.39	I431	0.22	I30	0.86	I694	0.47	I431	0.43	F1	0.68	I414	1.00	I188	0.60	I122	0.52	I636
17	0.41	I1180	0.26	I420	0.93	I27	0.47	I591	0.44	I157	0.68	I540	1.00	I310	0.65	I729	0.52	I532
18	0.42	I822	0.26	I385	0.94	I551	0.48	I706	0.45	I492	0.68	I608	1.00	I518	0.67	I521	0.53	I552
19	0.43	I301	0.27	I591	0.94	I116	0.48	I850	0.51	I4	0.69	I196	1.00	I484	0.67	I220	0.54	I290
20	0.45	I1207	0.28	I356	0.94	I483	0.48	I532	0.51	I546	0.69	I605	1.00	I1233	0.67	I139	0.54	I714

Note: All correlations below 0.3 are highlighted in light gray, negative correlations are highlighted in dark gray.

Table A 3.6 (continued)

Rank	PSK0400e		PA2000	
	Corr	ID	Corr	ID
1	-0.22	I120	0.01	I503
2	-0.15	I822	0.14	I822
3	-0.11	I532	0.22	I499
4	-0.10	I116	0.23	I188
5	-0.10	I636	0.23	I604
6	-0.09	I714	0.26	I612
7	-0.08	I158	0.27	I708
8	-0.08	I499	0.27	I594
9	-0.08	I220	0.28	F2
10	-0.08	I679	0.32	I224
11	-0.08	I468	0.32	I608
12	-0.08	I420	0.32	I518
13	-0.08	I196	0.33	I414
14	-0.05	I583	0.33	I239
15	-0.05	I841	0.36	I1207
16	-0.04	I239	0.40	I866
17	-0.04	I552	0.40	I82
18	-0.04	I224	0.41	I1180
19	0.07	I122	0.42	I551
20	0.15	I30	0.42	I591

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).
Note: All correlations below 0.3 are highlighted in light gray, negative correlations are highlighted in dark gray.

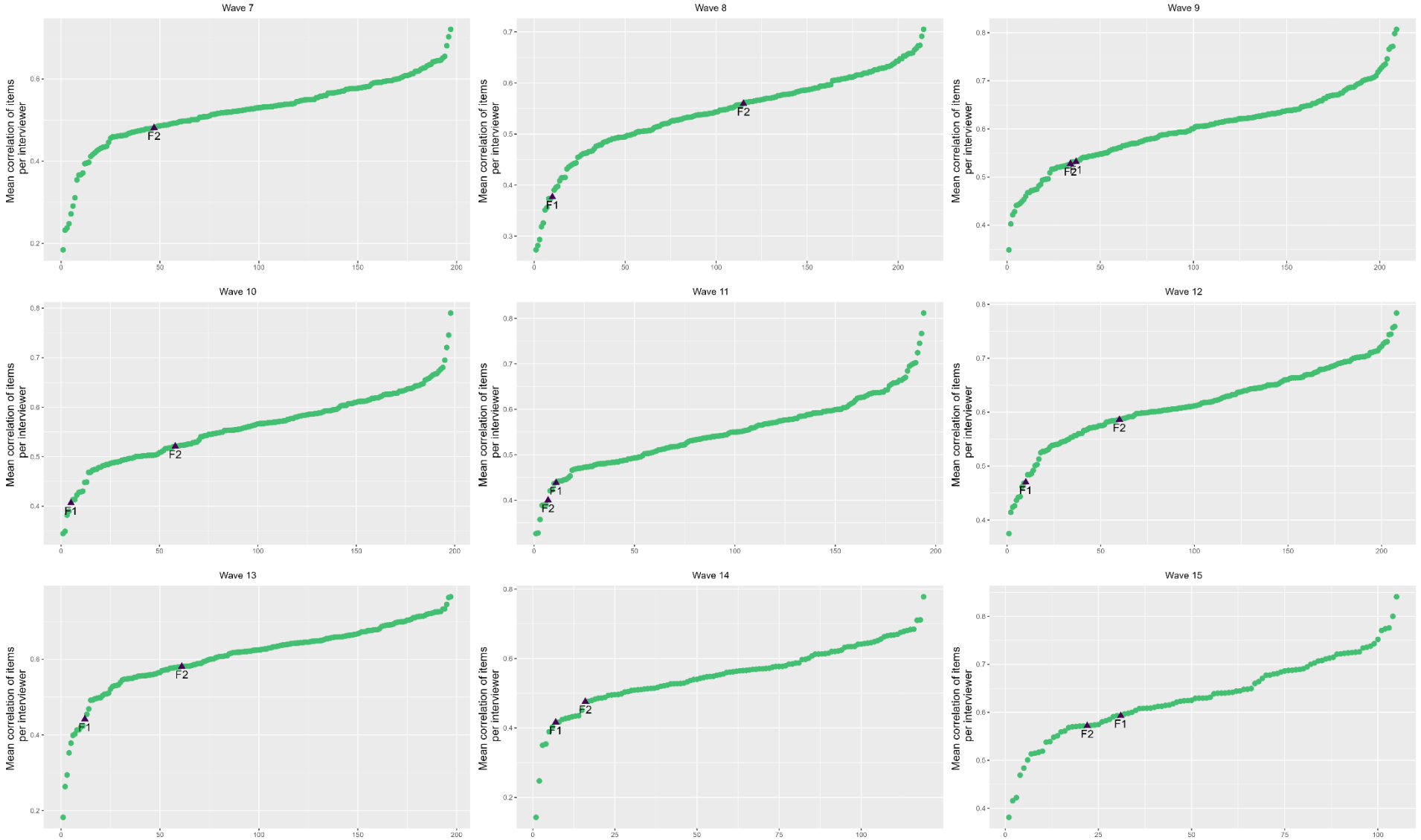

Figure A 3.16: Mean correlations between items per interviewer, waves 6/7 to 14/15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting mean.

Table A 3.7: Information on item batteries used to calculate mean correlations.

Item Batterie	Content	Scale/categories	Wave									
PA0100	Attitudes (satisfaction with areas of life)	1-10	6	7	8	9	10	11	12	13	14	15
PA0200	Attitudes (satisfaction with areas of life)	1-10	6	7	8	9	10	11	12	13	14	15
PA0300	Attitudes (satisfaction with areas of life)	1-10	6	7	8	9	10	11	12	13	14	15
PA0800	Social participation	1-10	6	7	8	9	10	11	12	13	14	15
PA0900	Social participation	1-10	6	7	8	9	10	11	12	13	14	15
PA1000	Attitudes (general situation)	1-10	6	7	8	9	10	11	12	13	14	15
PA2000	Social Trust	1-10	-	-	-	-	-	11	12	13	14	15
PD0500	Marital status	7	6	7	8	9	10	11	12	13	14	15
PEO0100a-e	Attitudes (Self Efficacy)	1-4	6	7	8	-	-	11	-	-	14	-
PEO0200a-d	Attitudes (Employment)	1-4	6	7	8	-	-	-	-	-	-	-
PEO0400a-d	Opinions (role/gender models)	1-4	-	-	8	-	-	11	-	-	14	-
PEO1600b-f	Attitudes (reciprocity)	1-4	-	-	-	9	10	-	-	-	-	-
PG0100	Health (visits to doctor)	open	6	7	8	9	10	11	12	13	14	15
PG0500	Health (recognized disabilities)	3	6	7	8	9	10	11	12	13	14	15
PG0800	Health (health restrictions)	2	6	7	8	9	10	11	12	13	14	15
PG1235	Health (activities)	1-5	6	7	8	9	10	11	12	13	14	15
PG1300	Health (health insurance)	6	6	7	8	9	10	11	12	13	14	15
PSK0200	Networks (close friends/family members)	open	6	7	8	9	10	11	12	13	14	15
PSK0300	Networks (conflicts in household)	1-5	6	7	8	9	10	11	12	13	14	15
PSK0400a-e	Networks (engagement in organizations/associations)	2	6	7	8	9	10	11	12	13	14	15

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

Table A 3.8: Correlations of the item batteries and further summery statistics on correlations between waves, waves 6/7 to 14/15.

	Correlations between Waves											
Item	6/7	7/8	8/9	9/10	10/11	11/12	12/13	13/14	14/15	Mean	Variance	Classification
PA0100	0.62	0.66	0.65	0.64	0.64	0.65	0.66	0.68	0.68	0.65	0.0003	Strong
PA0200	0.53	0.58	0.56	0.56	0.59	0.59	0.60	0.59	0.61	0.58	0.0006	Strong
PA0300	0.65	0.64	0.64	0.62	0.66	0.69	0.65	0.67	0.68	0.65	0.0004	Strong
PA0800	0.60	0.60	0.61	0.60	0.62	0.64	0.62	0.64	0.54	0.61	0.0008	Strong
PA0900	0.58	0.58	0.59	0.59	0.60	0.60	0.61	0.60	0.64	0.60	0.0004	Strong
PA1000	0.61	0.62	0.63	0.63	0.63	0.63	0.61	0.62	0.60	0.62	0.0001	Strong
PA2000	-	-	-	-	-	0.58	0.60	0.62	0.62	0.60	0.0003	Strong
PD0500	0.88	0.90	0.90	0.92	0.93	0.94	0.90	0.91	0.92	0.91	0.0003	Strong
PEO0200a	0.49	0.51	-	-	-	-	-	-	-	0.50	0.0002	Strong
PEO0200b	0.51	0.59	-	-	-	-	-	-	-	0.55	0.0016	Strong
PEO0200c	0.45	0.48	-	-	-	-	-	-	-	0.47	0.0002	Moderate
PEO0200d	0.49	0.53	-	-	-	-	-	-	-	0.51	0.0003	Strong
PEO1600b	-	-	-	0.46	-	-	-	-	-	0.46	-	Moderate
PEO1600c	-	-	-	0.46	-	-	-	-	-	0.46	-	Moderate
PEO1600d	-	-	-	0.30	-	-	-	-	-	0.30	-	Moderate
PEO1600e	-	-	-	0.47	-	-	-	-	-	0.47	-	Moderate
PEO1600f	-	-	-	0.41	-	-	-	-	-	0.41	-	Moderate
PG0100	0.38	0.33	0.37	0.37	0.36	0.43	0.32	0.39	0.31	0.36	0.0013	Moderate
PG0500	0.70	0.70	0.70	0.72	0.73	0.74	0.75	0.73	0.72	0.72	0.0002	Strong
PG0800	0.53	0.53	0.54	0.55	0.56	0.56	0.58	0.58	0.61	0.56	0.0006	Strong
PG1235	0.58	0.62	0.55	0.54	0.56	0.57	0.57	0.57	0.51	0.56	0.0008	Strong
PG1300	0.52	0.63	0.67	0.62	0.65	0.73	0.67	0.82	0.75	0.67	0.0069	Strong
PSK0200	0.41	0.48	0.47	0.53	0.54	0.50	0.56	0.62	0.63	0.53	0.0044	Strong
PSK0300	0.52	0.54	0.55	0.56	0.59	0.58	0.58	0.56	0.62	0.57	0.0007	Strong
PSK0400a	0.66	0.66	0.68	0.71	0.67	0.68	0.67	0.72	0.71	0.68	0.0005	Strong
PSK0400b	0.77	0.76	0.75	0.77	0.78	0.78	0.80	0.82	0.76	0.78	0.0004	Strong
PSK0400c	0.57	0.64	0.58	0.60	0.58	0.62	0.61	0.64	0.65	0.61	0.0007	Strong
PSK0400d	0.63	0.65	0.64	0.63	0.64	0.64	0.65	0.68	0.69	0.65	0.0004	Strong
PSK0400e	0.40	0.39	0.39	0.40	0.38	0.38	0.41	0.42	0.46	0.40	0.0006	Moderate

Table A 3.8 (continued)

Item	Correlations between Waves				Mean	SD	Classification
	6/7	7/8	8/11	11/14			
PEO0100a	0.44	0.48	0.44	0.40	0.44	0.0009	Moderate
PEO0100b	0.43	0.44	0.37	0.37	0.40	0.0010	Moderate
PEO0100c	0.44	0.44	0.42	0.40	0.42	0.0004	Moderate
PEO0100d	0.40	0.42	0.37	0.38	0.39	0.0003	Moderate
PEO0100e	0.41	0.41	0.38	0.32	0.38	0.0012	Moderate
PEO0400a	-	-	0.44	0.45	0.44	0.0000	Moderate
PEO0400b	-	-	0.47	0.46	0.46	0.0000	Moderate
PEO0400c	-	-	0.35	0.35	0.35	0.0000	Moderate
PEO0400d	-	-	0.52	0.52	0.52	0.0000	Strong

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

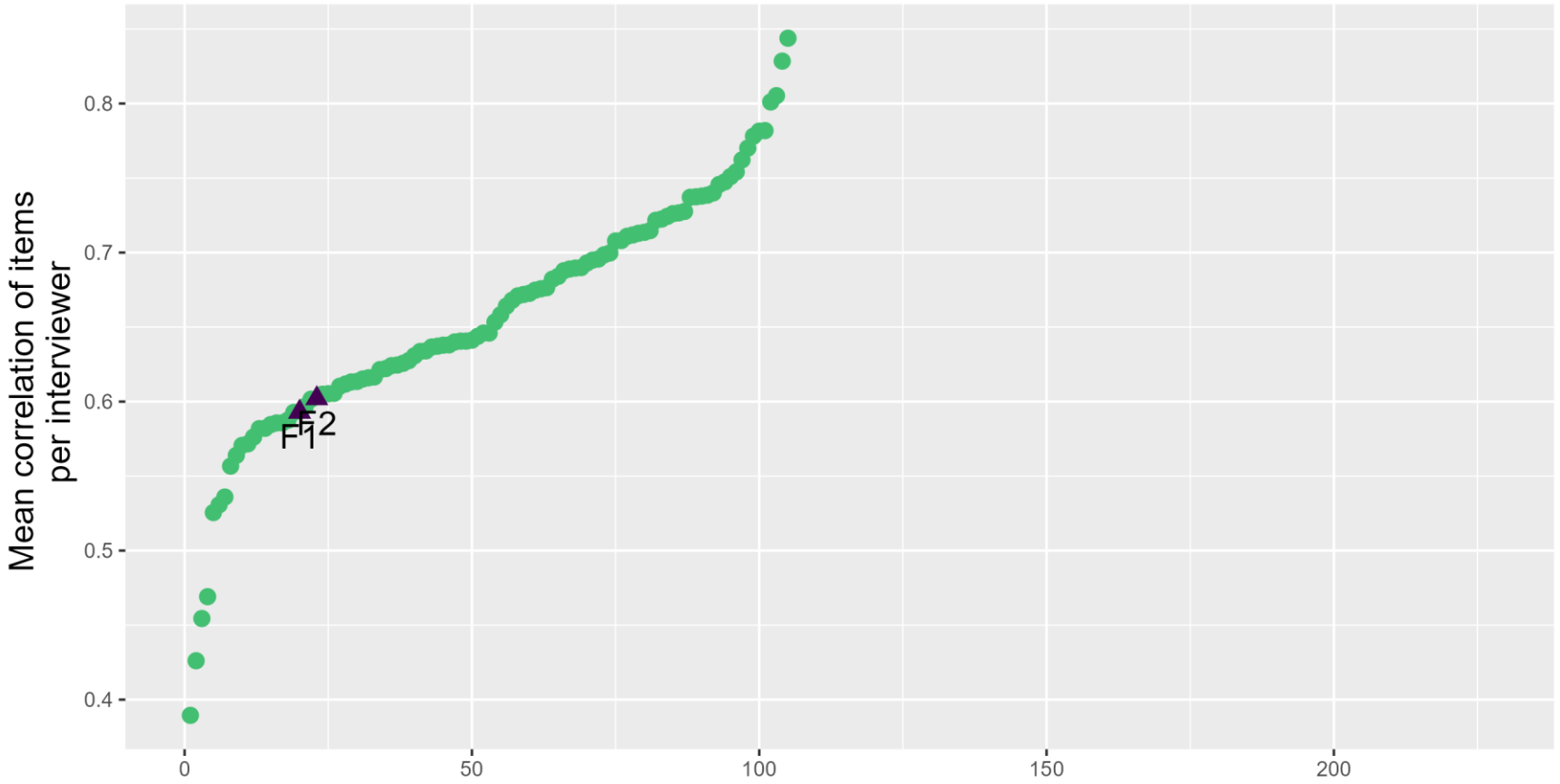

Figure A 3.17: Mean correlations between items per interviewer, waves 14/15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting mean. Figure only includes items with an overall correlation above 0.5.

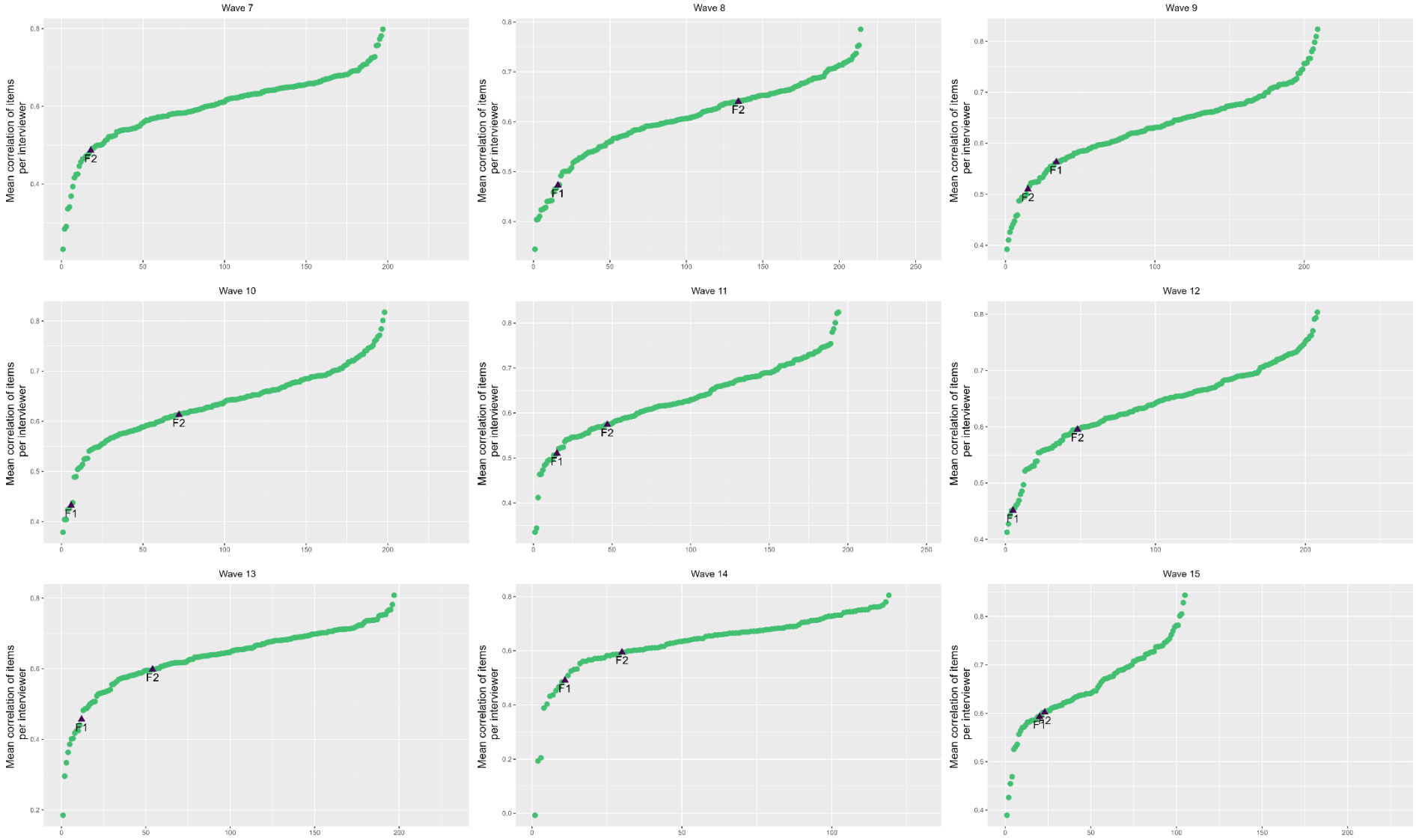

Figure A 3.18: Mean correlations between items per interviewer, waves 6/7 to 14/15.

Source: Panel study "Labour Market and Social Security" (PASS SUF W15).

Note: Interviewers are ranked and ordered based on their resulting mean. Figure only includes items with an overall correlation above 0.5.

Table A 3.9: List of interviewers, identified by multiple identification methods (at least three methods), wave 15.

Falsification Indicators						Cluster-Analysis				IsoForest		Near-	PCA			Correlations			
AVER	ACQ	MRS	INR	ROUND	DUR	AL	CL	SL	WL	Resp.	Indi.	Dup.	Life	Balance	Leisure	Indi.	Resp.	Item	Total
F1	F1	F1	-	-	F1	-	-	-	-	-	F1	F1	-	-	F1	-	-	-	7
-	F2	-	-	-	F2	F2	F2	F2	F2	F2	F2	F2		-	F2	-	-	-	10
I145	I145	I145	-	-	-	-	-	-		-	-	-	-	-	-	-	I145	-	4
-	I174	-	-	-	-	-	-	-	I174	-	-	-	-	-	-	I174	-	-	3
I187	-	I187	-	I187	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
I301	-	-	-	I301	-	-	-	-	-	-	-	-	-	I301	I301	-	-	-	4
I333	I333	I333	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
-	-	I336	-	-	-	-	-	-	-	-	-	I336	I336	-	I336	-	-	-	4
-	-	-	-	-	-	-	-	-	I345	-	-	I345	-	-	I345	I345	-	-	4
I381	I381	I381	-	-	-	-	-	-	-	-	-		-	-	-	I381	I381	-	5
-	-	-	-	I408	-	-	-	-	I408	-	-	I408	-	-	-	I408	-	-	4
I420	I420	-	-	-	I420	-	-	-	-	-	-	-	-	-	-	-	-	-	3
-	-	-	-	I43	I43	-	-	I43	-	-	-	-	-	-	-	-	-	-	3
-	-	-	-	I475	-	I475	I475	I475	I475	-	I475	-	-	-	-	-	-	-	6
-	I494	I494	-	-	-	-	-	-	I494	-	-	-	-	-	-	-	-	-	3
-	-	I532	-	-	-	-	-	-	-	-	-	I532	-	I532	-	-	-	-	3
-	-	-	-	I539	-	I539	-	I539	I539	-	I539	-	-	-	-	-	-	-	5
-	-	-	-	-	-	I561	-	-	I561	I561	I561	-	-	-	-	-	-	-	4
I571	-	-	-	I571	I571	-	-	-	-	-	-	-	-	-	-	-	-	-	3
I592	I592	I592	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3

Table A 3.9 (continued)

Falsification Indicators						Cluster-Analysis				IsoForest		Near-Dup.	PCA			Correlations			
AVER	ACQ	MRS	INR	ROUND	DUR	AL	CL	SL	WL	Resp.	Indi.		Life	Balance	Leisure	Indi.	Resp.	Item	Total
-	-	-	-	I608	-	-	-	-	-	-	-	-	I608	I608	-	-	-	-	3
-	-	-	-	-	I620	-	-	-	I620	-	I620	-	-	-	-	I620	-	-	4
I625	I625	I625	-	I625	-	-	-	-	-	-	-	-	-	-	I625	-	-	-	5
I634	I634	-	-	-	I634	-	-	-	-	-	-	-	-	-	-	-	-	-	3
-	-	-	-	I636	-	I636	-	I636	I636	-	I636	-	-	-	-	I636	-	-	6
-	I708	-	-	I708	-	I708	-	I708	I708	-	I708	-	-	-	-	-	-	-	6
I729	I729	I729	-	I729	I729	-	-	-	-	-	-	-	-	-	I729	I729	-	-	7
I735	-	I735	-	I735	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
-	I779	I779	-	-	-	-	-	-	-	-	I779	-	-	-	-	-	-	-	3
-	I787	I787	-	-	-	-	-	-	-	I787	-	-	-	-	-	-	-	-	3

Source: Panel study “Labour Market and Social Security” (PASS SUF W15).

References

Altschul, Sophie, Sebastian Bähr, Jonas Beste, Matthias Collischon, Mustafa Coban, Sandra Dummert, Corrinna Frodermann, Patrick Gleiser, Stefanie Gundert, Benjamin Küfner, Jan Mackeben, Sonja Malich, Bettina Müller, Stefan Schwarz, Jens Stegmaier, Nils Teichler, Mark Trappmann, Stefanie Unger, Claudia Wenzig, Marco Berg, Ralph Cramer, Christian Dickmann, Reiner Gilberg, Birgit Jesske, and Martin Kleudgen. 2022. "Panel Arbeitsmarkt und Soziale Sicherung (PASS) – Version 0621 V1." Forschungsdatenzentrum der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB). DOI: 10.5164/IAB.PASS-SUF0621.de.en.v2

Bergmann, Michael, Karin Schuller, and Frederic Malter. 2019. "Preventing Interview Falsifications During Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE)." Longitudinal and Life Course Studies 10(4):513–30.

Beste, Jonas, Lukas Olbrich, and Silvia Schwanhäuser. 2021. "Interviewer: innenkontrolle im Panel Arbeitsmarkt und soziale Sicherung (PASS)." Institut für Arbeitsmarkt- und Berufsforschung. Available at https://doku.iab.de/fdz/reporte/2021/MR_04-21.pdf.

Biemer, Paul P., and S. Lynne Stokes 1989. "The Optimal Design of Quality Control Samples to Detect Interviewer Cheating." Journal of Official Statistics 5(1):23–39.

Billiet, Jaak B., and Eldad Davidov. 2008. "Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design." Sociological Methods & Research 36(4):542–562.

Birnbaum, Benjamin, Gaetano Borriello, Abraham D. Flaxman, Brian DeRenzi, and Anna R. Karlin. 2013. "Using behavioral data to identify interviewer fabrication in surveys." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Available at https://bbirnbaum.com/assets/publications/chi13.pdf.

Blasius, Jörg, and Victor Thiessen. 2012. Assessing the Quality of Survey Data. SAGE Publications.

Blasius, Jörg, and Victor Thiessen. 2013. "Detecting Poorly Conducted Interviews." In Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 67–88. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Blasius, Jörg, and Victor Thiessen. 2015. "Should we trust survey data? Assessing response simplification and data fabrication." Social Science Research 52:479–93.

Blasius, Jörg, and Victor Thiessen. 2021. "Perceived corruption, trust, and interviewer behavior in 26 European Countries." Sociological Methods & Research 50(2):740–77.

Bredl, Sebastian, Peter Winker, and Kerstin Kötschau. 2012. "A Statistical Approach to Detect Interviewer Falsification of Survey Data." Survey Methodology Journal 38(1):1–10.

Available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012001/article/11680-eng.pdf.

Brüderl, Josef, Bernadette Huyer-May, and Claudia Schmiedeberg. 2013. "Interviewer behavior and the quality of social network data." In Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 147–60. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Bushery, John M., Jennifer W. Reichert, Keith A. Albright, and John C. Rossiter. 1999. "Using Date and Time Stamps to Detect Interviewer Falsification." Proceedings of the Survey Research Method Section, American Statistical Association, 316–20. Available at http://www.asasrms.org/Proceedings/papers/1999_053.pdf.

Castorena, Oscar, Mollie J. Cohen, Noam Lupu, and Elizabeth J. Zechmeister. 2023. "How worried should we be? The implications of fabricated survey data for political science." International Journal of Public Opinion Research 35(2):1–9.

Cohen, Mollie J., and Zach Warner. 2021. "How to Get Better Survey Data More Efficiently." Political Analysis 29(2):121–38.

Crespi, Leo P. 1945. "The cheater problem in polling." Public Opinion Quarterly 9(4):431–45.

de Haas, Samuel, and Peter Winker. 2014. "Identification of partial falsifications in survey data." Statistical Journal of the IAOS 30(3):271–281.

de Haas, Samuel, and Peter Winker. 2016. "Detecting Fraudulent Interviewers by Improved Clustering Methods—The Case of Falsifications of Answers to Parts of a Questionnaire." Journal of Official Statistics 32(3):643–60.

DeMatteis, Jill M., Linda J. Young, James Dahlhamer, Ronald E. Langley, Joe Murphy, Kristen Olson, and Sharan Sharma. 2020. "Falsification in Surveys: Task Force Final Report." Washington, DC: American Association for Public Opinion Research. Available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Data_Falsification_Task_Force_Report.pdf.

Edwards, Brad, Aaron Maitland, and Sue Connor. 2017. "Measurement error in survey operations management: detection, quantification, visualization, and reduction." In Total Survey Error in Practice, edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, Brady T. West, 253–77. John Wiley & Sons.

Edwards, Brad, Hanyu Sun, and Ryan Hubbard. 2020. "Behavior Change Techniques for Reducing Interviewer Contributions to Total Survey Error." In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 77–89. Boca Raton, FL: Taylor & Francis Group.

Finn, Arden, and Vimal Ranchhod. 2017. “Genuine fakes: The prevalence and implications of data fabrication in a large South African survey.” The World Bank Economic Review 31(1):129–57.

Fujita, Frank, and Ed Diener. 2005. “Life satisfaction set point: stability and change.” Journal of personality and social psychology 88(1):158–64.

Good, Marie, Teena Willoughby, and Michael A. Busseri. 2011. “Stability and change in adolescent spirituality/religiosity: a person-centered approach.” Developmental Psychology 47(2):538–50

Groves, Robert M. 2004. “Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects.” Survey Research 35(1):1–5.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. Survey Methodology. Hoboken, NJ: Wiley.

Hauck, Mathew. 1969. “Is Survey Postcard Verification Effective?” Public Opinion Quarterly 33(1):117–20.

Hood, Catherine C., and John M. Bushery. 1997. “Getting More Bang from the Reinterview Buck: Identifying ‘At Risk’ Interviewers.” Proceedings of the Survey Research Method Section, American Statistical Association, 820–24. Available at http://www.asasrms.org/proceedings/papers/1997_141.pdf.

IAB. 2017. “Revidierter Datensatz Der IAB-BAMF-SOEP-Befragung Von Geflüchteten.” Institut für Arbeitsmarkt und Berufsforschung. Available at http://doku.iab.de/grauepap/2017/Revidierter_Datensatz_der_IAB-BAMF-SOEP-Befragung.pdf.

Jebreel, Najeeb Moharram, Rami Haffar, Ashneet Khandpur Singh, David Sánchez, Josep Domingo-Ferrer, and Alberto Blanco-Justicia. 2020. “Detecting bad answers in survey data through unsupervised machine learning.” In Privacy in Statistical Databases, UNESCO Chair in Data Privacy, International Conference Proceedings, edited by Josep Domingo-Ferrer and Krishnamurty Muralidhar, 309–20. Springer International Publishing.

Jesske, Birgit. 2013. “Concepts and Practices in Interviewer Qualification and Monitoring.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 91–102. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Josten, Michael, and Mark Trappmann. 2016. “Interviewer effects on a network-size filter question.” Journal of Official Statistics 32(2):349–73.

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley.

Keating, Michael, Charles Loftis, Joseph McMichael, and Jamie Ridenhour. 2014. "New dimensions of mobile data quality." Federal CASIC Workshops. Available at https://www.census.gov/fedcasic/fc2014/ppt/05_keating.pdf.

Koch, Achim. 1995. "Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994." ZUMA Nachrichten 19(36):89–105.

Koczela, Steve, Cathy Furlong, Jaki McCarthy, and Ali Mushtaq. 2015. "Curbstoning and Beyond: Confronting Data Fabrication in Survey Research." Statistical Journal of the IAOS 31(3):413–22.

Kosyakova, Yuliya, Jan Skopek, and Stephanie Eckman. 2015. "Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach." International Journal of Public Opinion Research 27(3):417–31.

Kosyakova, Yuliya, Lukas Olbrich, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2019. "Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany." Institut für Arbeitsmarkt- und Berufsforschung. Available at https://fdz.iab.de/187/section.aspx/Publikation/k190404302.

Krosnick, Jon A. 1999. "Survey research." Annual review of psychology 50 (1): 537-567.

Kuriakose, Noble, and Michael Robbins. 2016. "Don't Get Duped: Fraud Through Duplication in Public Opinion Surveys." Statistical Journal of the IAOS 32(3):283–91.

Landrock, Uta. 2017. "Explaining Political Participation: A Comparison of Real and Falsified Survey Data." Statistical Journal of the IAOS 33(2):447–58.

Li, Jianzhu, J. Michael Brick, Back Tran, and Phyllis Singer. 2011. "Using Statistical Models for Sample Design of a Reinterview Program." Journal of Official Statistics 27(3):433–50.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. 2008. "Isolation forest." 2008 Eighth IEEE International Conference on Data Mining. Available at https://feitonyliu.wordpress.com/wp-content/uploads/2009/07/liu-iforest.pdf.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. 2012. "Isolation-based anomaly detection." ACM Transactions on Knowledge Discovery from Data (TKDD) 6(1):1–39.

Menold, Natalja, Peter Winker, Nina Storfinger, and Christoph J. Kemper. 2013. "A Method for Ex-Post Identification of Falsification in Survey Data." In Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 25–47. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Murphy, Joe, Joe Eyerman, Colleen McCue, Christy Hottinger, and Joel Kennet. 2005. "Interviewer Falsification detection using data mining." Proceedings of Statistics Canada Symposium 2005, Methodological Challenges for Future Information Needs. Available at https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X20050019445.

Murphy, Joe, Paul Biemer, Chris Stringer, Rita Thissen, Orin Day and Y. Patrick Hsieh. 2016. "Interviewer falsification: Current and best practices for prevention, detection, and mitigation." Statistical Journal of the IAOS 32(3):313–26.

Murphy, Joe, Rodney Baxter, Joe Eyerman, David Cunningham, and Joel Kennet. 2004. "A System for Detecting Interviewer Falsification." Proceedings of the American Statistical Association and the American Association for Public Opinion Research. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000517.pdf.

Olbrich, Lukas, Elisabeth Beckmann, and Joseph W. Sakshaug. 2024. "Multivariate assessment of interviewer-related errors in a cross-national economic survey." Working Paper No. 253. Österreichische Nationalbank (OeNB). Available at https://www.econstor.eu/handle/10419/286404.

Olbrich, Lukas, Yuliya Kosyakova, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2023. "Detecting Interviewer Fraud Using Multilevel Models." Journal of Survey Statistics and Methodology 12(1):14–35.

Porras, Javier, and Ned English. 2004. "Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys." Proceedings of the Survey Research Method Section, American Statistical Association, 4223–28. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000879.pdf.

Robbins, Michael. 2018. "New frontiers in detecting data fabrication." In Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC), edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop, and Brita Dorer, 771–805 Wiley & Sons, Inc.

Schäfer, Christin, Jörg-Peter Schräpler, Klaus-Robert Müller, and Gert G. Wagner. 2004a. "Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methodes." DIW Discussion Paper No. 441. Berlin: DIW–German Institute for Economic Research. Available at http://hdl.handle.net/10419/18293.

Schäfer, Christin, Jörg-Peter Schräpler, and Klaus-Robert Müller. 2004b. "Identification, Characteristics and Impact of Faked and Fraudulent Interviews in Surveys." European Conference on Quality and Methodology in Official Statistics. Available at https://www.diw.de/documents/dokumentenarchiv/17/41963/paper2004_schaeferetal.pdf.

Schräpler, Jörg-Peter. 2011. "Benford's Law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP)." Jahrbücher für Nationalökonomie und Statistik 231(5-6):685–718.

Schräpler, Jörg-Peter, and Gert G. Wagner. 2005. "Characteristics and Impact of Faked Interviews in Surveys: An Analysis of Genuine Fakes in the Raw Data of SOEP." Allgemeines Statistisches Archiv 89:7–20.

Schreiner, Irwin D., Jennifer Newbrough, and Karen Pennie. 1988. “Interviewer falsification in Census Bureau surveys.” Proceedings from Section on Survey Research Methods, 491–96. Available at http://www.asasrms.org/Proceedings/papers/1988_090.pdf.

Schwanhäuser, Silvia, Joseph W Sakshaug, Yuliya Kosyakova, and Frauke Kreuter. 2020. “Statistical identification of fraudulent interviews in surveys: improving interviewer controls.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 91–106. Boca Raton, FL: Taylor & Francis Group.

Schwanhäuser, Silvia, Joseph W. Sakshaug, and Yuliya Kosyakova. 2022. “How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification.” Public Opinion Quarterly 86(1):51–81.

Sharma, Sharan, and Michael R. Elliott. 2020. “Detecting Falsifications in a Television Audience Measurement Panel Survey.” International Journal of Market Research 62(4):432–48.

Simmons, Katie, Andrew Mercer, Steve Schwarzer and Courtney Kennedy. 2016. “Evaluating a new proposal for detecting data falsification in surveys.” Statistical Journal of the IAOS 32(3): 327–38.

Slomczynski, Kazimierz Maciek, Przemek Powalko, and Tadeusz Krauze. 2017. “Non-Unique Records in International Survey Projects: The Need for Extending Data Quality Control.” Survey Research Methods 11(1):1–16.

Storfinger, Nina, and Peter Winker. 2013. “Assessing the Performance of Clustering Methods in Falsification Using Bootstrap.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 46–65. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Thissen, M. Rita, and Susan K. Myers. 2016. “Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality.” Statistical Journal of the IAOS 32(3):339–47.

Trappmann, Mark, Sebastian Bähr, Jonas Beste, Andreas Eberl, Corinna Frodermann, Stefanie Gundert, Stefan Schwarz, Nils Teichler, Stefanie Unger, and Claudia Wenzig. 2019. “Data Resource Profile: Panel Study Labour Market and Social Security (PASS).” International Journal of Epidemiology 48(5):1411–1411g.

Van Vaerenbergh, Yves, and Troy D. Thomas. 2013. “Response styles in survey research: A literature review of antecedents, consequences, and remedies.” International Journal of Public Opinion Research 25(2): 195–217.

Wagner, James, Kristen Olson, and Minako Edgar. 2017. “The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata.” Survey Research Methods 11(3):218–33.

Weijters, Bert, Maggie Geuens, and Niels Schillewaert. 2010. “The stability of individual response styles.” Psychological methods 15(1): 96–110.

Weinauer, Marlene. 2019. “Be a Detective for a Day: How to Detect Falsified Interviews with Statistics.” Statistical Journal of the IAOS 35(4):569–75.

4. Leaving No Data Unturned: Evaluating Machine Learning Algorithms to Detect Interviewer Falsification

Abstract

Interviewer-administered surveys are inherently susceptible to the influence of deviant or fraudulent behavior on the part of interviewers. Even small amounts of data, fabricated by interviewers, can severely bias estimation results. Consequently, identifying falsified interviews is an important part of the quality control process. In addition to established quality control methods, like re-interviews or monitoring, statistical i.e., data-based detection methods can help identify potential falsifications by flagging suspicious patterns in the data. One understudied statistical detection approach in this context is the use of supervised machine learning algorithms that is algorithms trained on existing falsification data. This study explores the application of these algorithms for detecting falsifications, employing both experimental data and real survey data: The experimental data were collected specifically to study falsifications and the behavior of falsifiers. The survey data come from a large nationally representative survey of refugees in Germany with known fabricated interviews. We investigate how effective different supervised algorithms, such as regression models, decision trees, support vector machines, and neural networks, are at identifying patterns caused by falsifiers. Simulating different scenarios, we evaluate the effectiveness of these algorithms 1) when training them on falsifications within the same survey, 2) when training them on falsifications induced by different falsifiers within the same survey, and 3) when training them on falsifications from a completely different survey. Our results show that supervised algorithms very precisely detected falsifications within the same survey, especially algorithms based on decision trees. However, performance of all algorithms strongly decreases in the between-survey scenario. No algorithm was able to precisely identify falsifications in another survey.

4.1 Introduction

Interviewers wield a crucial role in the collection of survey data. They identify and convince respondents to participate in the survey or clarify questions and inquiries, facilitated due to the direct communication (Groves et al. 2009). However, the interviewers' involvement is twofold: While they can encourage respondent engagement, they can also encourage any person willing to participate—which is considered a deviation from the selection rules; while interviewers can answer questions and provide clarifying details, they can also manipulate the

wording of questions or add false explanation—thus deviating from standardized interview protocols. In the worst case, interviewers can fabricate (parts of) the questions without any involvement of the respondent, also known as complete or partial interview falsification. The range of possible forms of misbehavior is wide: The American Association for Public Opinion Research (AAPOR) considers any interviewer behavior that represents an intentional and unreported deviation from the guidelines or instructions to be interview falsification (Groves 2004). Such deviations can severely bias results and estimates (Schräpler and Wagner 2005), making the prevention and identification of these interviews a critical goal for ensuring data quality (DeMatteis et al. 2020).

Common strategies for detecting falsifications include monitoring (e.g., evaluation of audio recordings) and re-interviewing procedures (e.g., re-contacting respondents) (Groves 2004; Robbins 2018). In addition, data-based detection methods are gaining popularity as a cost-effective complement to monitoring and re-interviewing (e.g., Blasius and Thiessen 2013; Menold et al. 2013; Thissen and Myers 2016; Slomczynski, Powalko, and Krauze 2017; Bergmann, Schuller, and Malter 2019; Schwanhäuser, Sakshaug, and Kosyakova 2022). By identifying interviews with a high falsification probability or interviewers with suspicious patterns, data-based detection methods allow for more targeted and therefore effective controls. Consequently, the number of proposed detection methods has increased in recent years. At the same time, machine learning algorithms are gaining popularity in various fields. They hold significant potential in the context of survey research, particularly for the identification of falsified interviews (Buskirk et al. 2018). Compared to other data-based detection methods, machine learning methods are relatively straightforward to implement and automate, and resulting outcomes are less ambiguous (Schwanhäuser, Sakshaug, and Kosyakova 2022). In theory, they hold the potential of identifying new falsification patterns or different falsification types and can learn from real-world falsification behavior (Shah et al. 2020). Yet, the current literature lacks evaluations regarding the effectiveness of (supervised) machine learning in identifying interviewer falsification and its implementation in the quality control process.

This study aims to address the aforementioned gap, by investigating the potential of supervised machine learning, i.e., algorithms that are trained on existing falsification data. More specific, our objective is to ascertain the effectiveness of different supervised algorithms, including regression models, decision trees, support vector machines, and neural networks, in detecting falsifications. We make use of two distinct data sources, including falsification data caused by interviewers. First, we rely on an experimental dataset that was specifically collected

to study the behavior of falsifiers and patterns of falsifiers. Second, we utilise data coming from a large-scale survey of refugees in Germany, including documented cases of falsifications. In order to evaluate the performance of supervised machine learning algorithms in different settings, we simulate three distinct scenarios. In the first scenario, the datasets were randomly divided into training and test sets. The algorithms were trained on the falsifications in the training data to predict the status of the interviews in the test data (falsification versus real interview). This was done separately for the experimental data and the real-world refugee data. In the second scenario, we adopted a similar approach, ensuring that all interviews conducted by one interviewer were either assigned to the training or the test data. In the final scenario, the machine learning algorithm was trained based on the falsifications in the experimental data, and its ability to detect falsifications was tested based on the real-world refugee data.

The finding of our study indicate that supervised machine learning algorithms are effective in detecting falsified interviews in the first two scenarios. In particular algorithms that employ tree-based methods demonstrated robust performance, independent of the data source (either experimental or real-world data). The results of the last scenario were less encouraging. Although we were able to identify some falsifications using the different algorithms, most interviews and falsifications were wrongly classified by them. Consequently, the use of falsification data from one survey to train machine learning algorithms with the objective of detecting falsifications in another survey did not result in more targeted falsification detection than a random control of interviewers would have.

4.2 Falsification Detection and the Usage of Machine Learning

Interview falsification represents a significant threat to survey data quality (Schräpler and Wagner 2005; DeMatteis et al. 2020). Consequently, methods targeting the identification of falsifications represent a crucial part of data quality controls. In practice, standard quality controls often include a variety of different observational monitoring approaches, including silent monitoring of interviews, computer assisted audio recordings (CARI), the use of GPS locations, the collection of digital validation material such as screenshots or photos of the interview location, as well as re-contact methods verifying the proper conduction of the interview or its content (e.g., Thissen et al. 2008; Jesske 2013; Thissen 2014; Finn and Ranchhod 2017; Thissen and Myers 2016; Wagner, Olson, and Edgar 2017).

In addition to these observational methods, data-based detection methods have become increasingly used. These statistical methods, aim to identify outlying, repetitive, or otherwise

suspicious patterns using outlier analysis, duplicate analysis, or applying models to assess the falsification likelihood (e.g., Murphy et al. 2005; Li et al. 2011; Menold et al. 2013; Slomczynski, Powalko, and Krauze 2017). Other statistical methods focus on systematic differences between the real respondent data and the falsified data, e.g., by using falsification indicators in different ways. Falsification indicators measure patterns that are indicative of deviant interviewer behavior. Often, they are based on assumptions regarding possible motives for falsifying: falsifiers endeavor to maximize their monetary benefit and minimize their time expenditure and effort, while trying to remain undetected (Menold et al. 2013; Kosyakova, Skopek, Eckman 2015; Winker 2016). Indicators can be generated from all available survey data. Despite the shared idea behind falsification indicators, studies widely vary in the concrete falsification indicators they use. Studies range from using quality or performance indicators—like the number of item-nonresponse or the variation within the data—(e.g., Bredl, Winker, and Kötschau 2012; Menold et al. 2013; Bergmann, Schuller, and Malter 2019; Schwanhäuser, Sakshaug, and Kosyakova 2022), to raw paradata or metadata—for example scroll and click patterns or time stamps—(e.g., Birnbaum et al. 2013), and the already mentioned monitoring results such as GPS-data, CARI, or captured photos (e.g., Finn and Ranchhod 2015; Thissen and Myers 2016; Wagner, Olson, and Edgar 2017). For a detailed overview of indicators, see Schwanhäuser, Sakshaug and Kosyakova (2022).

Recently, the usage of machine learning algorithms has also increased in the context of falsification identification, even though it is still rather limited compared to other quality control methods. This is despite the potential advantages that machine learning holds in this area of application. One major advantage of machine learning is the possibility of automating decision-making processes, resulting in significant time savings. Compared to time-consuming interview monitoring via audio recordings or often complicated analysis of GPS data (Thissen 2014, Wagner, Olson, and Edgar 2017, Schwanhäuser, Sakshaug, and Kosyakova 2022), interpretation of machine learning results is straight forward, as binary classification algorithms categorize each case either as falsification or no falsification. Once established, an analysis pipeline—including data import, processing, and analysis by the algorithms—may require minimal adjustments. Especially supervised machine learning, provides the opportunity for a continuous and highly adaptable learning process. It allows to identify future instances of falsification based on real patterns instead of assumptions about falsification behavior (Walzenbach 2021), continuous improvement of model performance based on new falsification data, and allows for the identification of different types of falsification behavior at the same

time (Olbrich et al. 2023). In summary, machine learning can simplify and accelerate data quality controls, enabling more targeted and cost-effective follow-up controls of interviews.

4.2.1 Unsupervised Machine Learning

In practice, due to the lack of appropriate test and training data, most studies apply unsupervised machine learning algorithms to detect interview falsification in survey data[11]. These studies commonly focus on demonstrating single algorithms, rather than evaluating or comparing different algorithms. Although some studies early on used prior knowledge about falsifiers or falsifications to improve re-interview samples (Biemer and Stokes 1989; Stokes and Jones 1989), Murphy et al. (2004) were the first to use unsupervised machine learning tools: They use scoring models and anomaly detection to identify suspicious patterns in response and paradata of the US survey on substance use and abuse (NSDUH). Rather than evaluating these algorithms, they use the techniques to identify new falsification indicators specific to their data which might therefore not be applicable to other surveys. A commonly used unsupervised method for detecting interview falsifications is cluster analysis—a multivariate method grouping similar objects based on their (dis-)similarity (Bredl, Winker, and Kötschau 2012). Common practice involves utilizing falsification indicators as inputs for cluster analysis. Studies relying on this approach use a variety of clustering algorithms like Average Linkage, Ward's Linkage, Single Linkage, k-Means, or a heuristic optimization approach called threshold acceptance algorithm (e.g., Bredl, Winker, and Kötschau 2012; Menold et al. 2013; Bergmann, Schuller, and Malter 2019; Schwanhäuser, Sakshaug, and Kosyakova 2022). Other studies use (categorical) Principal Component Analysis (PCA)—a statistical tool for reducing dimensionality in the data while preserving its variability—to detect interviewers with duplicated response patterns or low variance between responses across different interviews (Blasius and Thiessen 2012, 2013, 2015, 2021; Schwanhäuser et al. 2020). Even though the method is a valuable tool for identifying data quality problems of various kinds, the method can be sensitive regarding survey characteristics (e.g., population, number of items, scale length) (Schwanhäuser et al. 2020).

4.2.2 Supervised Machine Learning

Only few studies focus primarily on supervised algorithms. One early contribution uses regression models on data including falsifications from the current population survey (CPS) (Li

[11] Note that we will focus on literature that uses methods based on survey response data, rather than, for example, observational paradata such as audio recordings from Computer Assisted Recorded Interviewing (CARI). There is a growing body of literature on automated analysis of CARI data (see, for example, Sun and Yan 2023).

et al. 2011), to predict the likelihood of falsification for each interview. They show that a focused re-interview sample based on these predictions detects more falsified interviews than a random sample. Other applications of supervised machine learning can be found in the work of Cohen and Warner (2020). Rather than evaluating the performance of an algorithm, they use classification algorithms to assess the effectiveness of different quality control procedures and indicators, using labeled data (published vs. deleted from the data release) from the Americas Barometer survey. To the best of our knowledge, only one study primarily focuses on different supervised machine learning algorithms (Shah et al. 2020), using logistic regression, discriminant analysis, support vector machines, classification trees, k-Means clustering, naïve Bayes, and neural networks on data from an Indian mobile phone survey, combined with experimentally produced falsifications from the interviewer training. However, they neglect a performance evaluation of the algorithms, only describing their control system.

4.2.3 Combined Use of Supervised and Unsupervised Methods

Another body of literature rather focuses on the comparisons and demonstration of different machine learning methods. In line with this notion, one study—using cluster analysis, PCA, and duplicate analysis—found that cluster analysis exceptionally outperformed PCA (Schwanhäuser et al. 2020). This study uses survey data including confirmed falsifications. Similarly, Jebreel et al. (2020) used a density-based clustering method, PCA, and ensemble-based regression trees, simulating "low quality" data instead of focusing on verified falsifications. They find that a combination of the different methods provided better performance than single methods. Another study applies k-Means clustering and logistic regression to experimental data from Indonesia, including a similar proportion of real and falsified data (Rosmansyah et al. 2019). However, it does not provide information on the sensitivity or performance of the methods, stating that the methods worked and will be used as a tool for advanced monitoring. Lastly, Weinauer (2019) suggests specific algorithms for outlier analysis (namely the binomial test, and a measure of deviations from the median) and cluster analysis, however, without providing an evaluation of the performance. Further, Birnbaum (2012) use different supervised and unsupervised algorithms on a mobile phone survey in Africa, namely multinomial modeling techniques and S-value techniques as unsupervised algorithms, and logistic regression, a K2 algorithm for learning a Bayesian network, and random forest as supervised algorithms. The study shows that all algorithms are able to detect falsifications with high sensitivity and specificity. However, they see the combined use of supervised and unsupervised algorithms along with repeated feedback loops as the ideal

solution. A related paper focuses on the application of a supervised classification algorithm, namely Random Forest (Birnbaum et al. 2013). They use experimental data from a mobile phone survey: simulating different levels of motivation and informedness of falsifiers by increasing incentives for "good" falsifications and the information provided on the detection methods. This approach achieves an excellent performance, even when interviewers are fully aware of the controls.

4.3 Research Question and Motivation

This study contributes to the body of literature comparing and evaluating supervised machine learning algorithms for identifying interviewer falsifications. More specifically, we investigate how effective different regression models, decision trees, support vector machines, and neural networks, are at identifying patterns caused by falsifiers. We investigate the potentials of these supervised algorithms, by simulating different scenarios: 1) training algorithms on falsifications within the same survey, 2) training algorithms on falsifications induced by different falsifiers within the same survey, 3) training algorithms on falsifications of a different survey. More concrete, we answer the following three research questions (RQ):

RQ1: How effective are supervised machine learning algorithms in detecting falsifications, when training these algorithms on other falsifications detected in the same survey?

To test this, we train and test the algorithms based on two distinct datasets: An experimental dataset, including an equal amount of falsifications and real interviews, and survey data including fabricated interviews. Both datasets are separately and randomly divided into training and test data; The algorithms are trained based on the known falsifications in the training data in order to predict the status of the interviews in the test data. By using both datasets separately, we are able to identify, whether different supervised algorithms are able to identify falsifications within the same survey. Further we are able to evaluate, if certain algorithms perform similar well in different settings i.e., surveys.

RQ2: How effective are supervised machine learning algorithms in detecting falsifications, when training these algorithms on other falsifications detected in the same survey but caused by different falsifiers?

Next, we aim to evaluate whether we can identify falsifications caused by different falsifiers within the same survey. Using a random split of half the interviewers, we are able to

identify, whether different supervised algorithms are able to identify falsifications coming from different falsifiers. By holding out a set of interviewers in the training data and predicting the status of their interviews, we are able to evaluate whether we are able to identify different falsifiers. For practitioners, the results could indicate whether it is worth the effort of collecting some artificial falsifications, produced before the launch of a survey.

RQ3: How effective are supervised machine learning algorithms in detecting falsifications, when training these algorithms on previous falsifications detected in a different survey?

Last, we test whether the different supervised methods are able to identify falsifications based on two different datasets, by training the machine learning algorithm on the falsifications in the experimental data, and evaluating the ability to detect falsifications in the survey data. Answering this question, hence, gives us the insight into whether we can compile existing falsification data from other surveys in order to better control more recent surveys. Since different surveys likely produce distinct patterns, this comes with the risk of dataset shifting and hence lower performance.

4.4 Data

In order to answer the three research questions, we utilize data from two different sources: First, we use data from an experimental study designed to examine falsifications and falsifiers behavior. The data includes an equal amount of real and falsified data (total = 1,420), both collected within an experimental setting. Second, we use real-world data from a refugee survey in Germany, including 351 verified falsifications (7.3 %) on the person level. In both datasets, falsifications are labeled as 1 and real non-falsified data are labeled as 0. Hence, the task at hand is a binary classification problem. For the experimental data classes are balanced, whereas classes for the real-world data are imbalanced. Hence, we are also able to evaluate the supervised algorithms under different scenarios.

4.4.1 Experimental Data

The experimental data come from an experiment that was conducted in 2011 at the University of Giessen, Germany (see Menold et al. 2013, de Haas and Winker 2014 for further information). The experiment includes real survey interviews as well as falsified interviews. In a first step, 78 trained students from University of Giessen, conducted approximately ten interviews with randomly selected fellow students. This resulted in 710 real face-to-face

interviews. For all interviews, audio recordings were used to ensure the quality of these interviews. In a second step, the student-interviewers randomly received a socio-demographic profile based on the collected survey data from the other student-interviewers. They were asked to falsify another ten interviews (resulting in a total of 710 falsified interviews) based on this profile—including for example information like sex, age, subject of study, number of semesters enrolled—in a laboratory setting (see **Table 4.1** for an overview). On average, these student interviewers completed between 15 and 20 (mean = 18, median = 19) real interviews and fabrications. The real interviews had a mean length of 33 minutes (median = 32 minutes). The student interviewers were aware of the purpose of the study. Additionally, they were incentivized to deliver a "realistic" falsification that tries to mimic a real interview as closely as possible, by paying a high incentive to the three hardest-to-detect falsifiers.

Table 4.1: Overview of falsified and real data, experimental data.

	Person-level interviews	
	N	Percentage
Total	1,420	100.0
thereof falsified	710	50.0
thereof real	710	50.0

Source: Experimental data, University of Giessen, 2011.

The questionnaire included 62 questions, specifically including questions that allow the analysis of falsification and evaluation of detection tools, but that are also close to real world data. This applies, for example, to item batteries which allow for the calculation of response patterns. The questionnaire included the following main topics: Attitudes toward political issues, attitudes toward women's labor force participation, the economic situation, social justice, political participation, personality traits. Hence, most questions stemmed from the 2008 round of the General Population Survey of the Social Sciences (ALLBUS) in Germany.

4.4.2 Survey Data

We further utilize data from the IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33; Brücker, Rother, and Schupp 2017). The survey is an annually conducted panel household survey, which started in 2016. The data is integrated into the data structure of the German Socio-Economic Panel (GSOEP). In the first wave, the target population of IAB-BAMF-SOEP included refugees who arrived between 2013 and 2016 as well as their adult household members. The initial sample was based on the German Central Register of Foreigners (Ausländerzentralregister; AZR) (Kroh et al. 2017). The first wave initially resulted

in an sample of 3,554 responding households (household-level interviews), which amounts to a household-level response rate (Response Rate 2; AAPOR 2023) of 48.7 percent. Additionally, 4,816 person-level interviews were conducted. Fieldwork was carried out by 98 trained interviewers. These interviewers completed between 1 and 289 computer-assisted personal interviews (CAPI) on the person-level. These interviews had a mean length of 90 minutes (median = 81 minutes). To account for the different language prophecies, multilingual interviewers were doing the work and questionnaires were provided in various languages (Arabic, English, Farsi/Dari, German, Kurmanji, Pashtu, and Urdu). Additionally, audio files containing recordings of the questions and access to an interpreter hotline was available (Jacobsen 2018). The person-level questionnaire included around 450 possible questions, depending on the filter. It included the following main topics: migration and escape history, migration biographies on education, language acquisition and employment, as well as satisfaction in different life domains, health, and attitudes (Brücker, Rother, and Schupp 2017). Different quality controls applied after the first wave revealed cases of interviewer fraud and misbehavior conducted by a total of three falsifiers (Kosyakova et al. 2019). These falsifiers in total conducted 351 person interviews (see **Table 4.2**), which were subsequently excluded from the officially released data (version SOEP.v34).

Table 4.2: Overview of falsified and real data, real-world data.

	Household-level interviews		Person-level interviews	
	N	Percentage	N	Percentage
Total	3,554	100.0	4,816	100.0
thereof falsified	265	7.5	351	7.3
thereof real	3,289	92.5	4,465	92.7

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

4.5 Algorithms and Evaluation Strategy

4.5.1 Analysis Strategy

In order to assess the classification precision of the different supervised algorithms, we proceed in the following way: First, we prepare appropriate features, available for both experimental and real-world data. As we use two different datasets including different questionnaires we rely on falsification indicators, which are commonly used in the literature and available for most surveys. To guaranty comparability of indicators between datasets, all indicators are standardized. Second, we prepare different training and testing data from the experimental and real-world data, also dependent on the research question. **Figure 4.1** shows

an overview of the different training and testing data, according to each research question (RQ). To answer RQ1 we split both experimental data (RQ1a) and real-world data (RQ1b) in separate training and testing datasets. Based on conventional practice, both original datasets are randomly split 80/20, meaning that 80% of the data are used as training data and the remaining 20% of the data as testing data. To answer RQ2 we only rely on the experimental data, as the real-world data only includes three falsifiers. The original dataset is split into two random samples of interviewers including around 50% of the original data each. To answer RQ3 we utilize both datasets together, using the experimental data as training data and the real-world data as testing data.

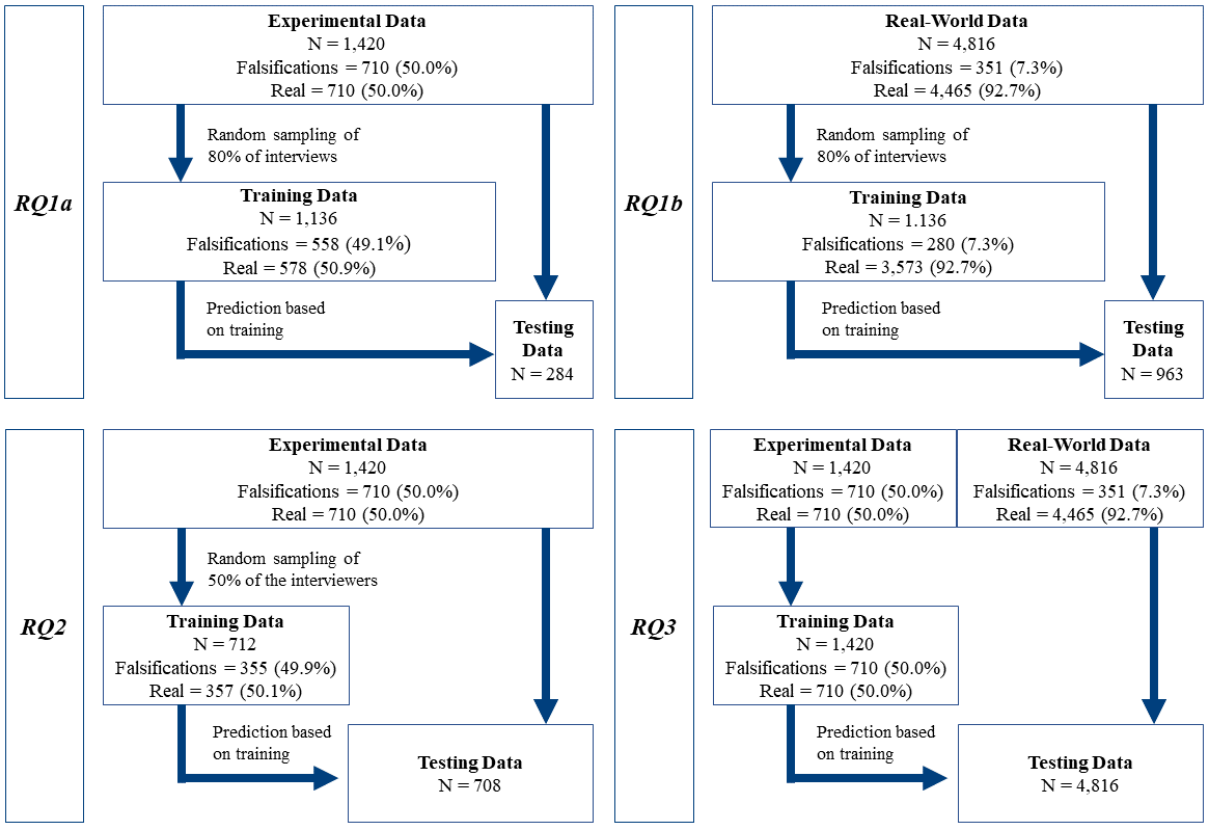

Figure 4.1: Overview of training and testing procedure, separate for each Research Question. *Source*: Own illustration.

Using the different training data versions, third, we perform parameter tuning in order to select the best model, in each setting. The evaluation of the various tuning parameters (listed in Appendix **Table A 4.1**) and, hence, evaluation of each model is based on two different performance metrics: F1-Score and ROC/AUC. This is done to take various performance criteria into account and address the imbalancedness of the classification problem in case of the real-world data. Lastly, selected models are applied to the different testing datasets. Again, final performance evaluation is done using the two metrics F1 and ROC/AUC. Training, testing, and

evaluation of the models are done using the program R, under version 4.2.1., using the package "caret", which combines various different machine learning packages (Kuhn et al. 2020). For the model evaluation we additionally rely on repeated 5-fold cross-validations (3 repeats).

4.5.2 Features: Falsification Indicators

Machine learning algorithms require features as input variables to train and test the models. In the context of falsification identification, all data produced in the survey could potentially be used as such features. The problem with information like e.g., paradata or GPS data is that they are not available to all surveys. What is, in fact, available for all surveys is the respondents survey data themselves. Therefore, we will rely on this type of data for our analysis. However, as every survey has its own unique topics on items, we will use a common concept of falsification identification methods: falsification indicators. Each indicator represents a feature for our analysis. In total, we use 11 different indicators: Acquiescent-Responding-Style, Extreme-Responding-Style, Benford's Law, Filter questions, Item Nonresponse, Middle-Responding-Style, Non-Differentiation, Primacy and Recency Effects, Rounding Tendency, and Semi-Open responses. Aligning with Schwanhäuser, Sakshaug and Kosyakova (2022), **Table 4.3** shows an overview of these indicators and their respective definitions.

According to the literature, the rationale behind these indicators is as follows: First, falsifiers tend to produce lower response variance within interviews compared to honest interviewers (Schäfer et al. 2004; Menold et al. 2013), measurable as *Non-Differentiation* (ND) within item batteries. This is caused by a variety of strategies or behaviors, which also serve as possible indicators: Falsifiers have a tendency for choosing answers in the middle of ordinal response scales (*Middle-Responding-Style*; MRS) rather than extreme values (Extreme-Responding-Style; ERS) to avoid suspicious inconsistencies (Porras and English 2004; Storfinger and Winker 2013). Further, they tend to avoid *Item-Nonresponse* (INR) by providing answers to all closed-ended questions (Bredl, Winker, and Kötschau 2012). To reduce implausible answer combinations, which could raise suspicion, falsifiers rarely show *Acquiescent-Response-Behavior* (ACQ) i.e., the tendency to agree or answer "yes" to opinion items. To decrease their effort, falsifiers often choose answers which trigger fewer follow-up questions due to *filtering* (FILTER) (Hood and Bushery 1997; Eckman et al. 2014; Kosyakova, Skopek, Eckman 2015). Furthermore, real respondents hear the questions, whereas falsifiers read and answer the questions as in a self-administered mode, which may lead to different *Primacy* (choosing the first options of answer lists; PRIM) and *Recency Effects* (choosing the last options of answer lists; RECE) (Menold et al. 2013). Respondents also show a higher

Rounding Tendency (ROUND) in open numeric questions (e.g., income, working hours) compared to falsifiers (Menold et al. 2013). Additionally, falsifiers tend to avoid answering *Semi-Open-Ended Items* (SEMIOP) leading to higher rates of nonresponse and less frequent selection of the "Other, specify"-option (Bredl, Winker, and Kötschau 2012). Last, falsifiers struggle to replicate some answer distribution like *Benford's Law* (BFL) which approximates the distribution of the first digit in naturally occurring numbers (Schäfer et al. 2004).

Table 4.3: Overview of indicators, acronyms, and their definition.

Indicator	**Acronym**	**Definition**
Acquiescent-Responding-Style	ACQ	Share of positive connotation ("Agree/ Strongly Agree") independent of content
Extreme-Responding-Style	ERS	Share of extreme responses to rating scales
Benford's Law	BFL	Deviations from the decreasing distribution of leading digit for numeric quantities
Filter questions	FILTER	Share of responses leading to follow-up questions
Item Nonresponse	INR	Share of item nonresponse within an interview
Middle-Responding-Style	MRS	Share of middle responses to rating scales
Non-Differentiation	ND	Mean standard deviation within different item scales
Primacy Effects	PRIM	Share of first two categories in non-ordered answer option lists
Recency Effects	RECE	Share of last two categories in non-ordered answer option lists
Rounding Tendency	ROUND	Share of rounded numbers in numerical open-ended questions
Semi-Open responses	SEMIOP	Share of responses to "other" in semi-open-ended question

Source: The table was adapted from Schwanhäuser, Sakshaug and Kosyakova (2022).

We calculate each of these indicators separate for each interview. Since the two surveys include different questions, indicators values could vary based on the dataset used, making results less comparable. Hence, we calculate z-standardized indicator values $z_{i,j}$, following equation 4.1:

$$z_{i,j} = \frac{x_{i,j} - \bar{x}_j}{S_j} \tag{4.1}$$

Here, $x_{i,j}$ denotes the raw indicator values for interview *i* of indicator *j*, with $\bar{x}_j$ and S_j denoting the mean indicator value and the respective standard deviation. In order to make results for single indicators more comparable, we further code all indicators in a way, that positive values indicate the suspicious direction, according to the findings in the literature.

4.5.3 Machine Learning Algorithms

To provide a comprehensive overview of the performance of different supervised machine learning tools, we draw on a wide variety of commonly used algorithms. We start with a simple and widely used method—also known outside of machine learning—namely (1) regression models. Since our classification problem is binary—falsification vs. real interview––we first rely on logistic regression. We extend the framework by applying more advanced regression models, namely boosted logistic regression as well as two different regularized logistic regressions (Lasso and Ridge). The second set of algorithms we use are (2) decision trees. Again, we start with the basic version, namely simple decision Trees. We also use different ensemble methods (bagging and boosting), to improve the performance of the tree-based learner: Bagged decision Tress, Bagged AdaBoost, Boosted Classification Trees, and XGBoost. Besides, we rely on (Conditional Inference) Random Forest. The third set of algorithms we use are (3) Support Vector Machines, for which we rely on different kernels (Linear, Polynomial, and Radial Basic Function). Finally, we use (4) Artificial Neural Networks: Simple Neural Networks and Monotone Multi-Layer Perceptron Networks. **Table 4.4** shows a summary of the mentioned algorithm families, the respective algorithms as well as the methods used in the R-package 'caret' and an overview of the used tuning parameters of each method. An overview of all used parameter values within the named tuning parameters can be found in the Appendix (**Table A 4.1**).

Table 4.4: List of used Algorithms, methods in the 'caret' R-package, and model tuning parameters.

Family	Algorithm	Method	Model Parameters
(1) Regression Models	Logistic regression (LR)	'glmnet'	alpha lambda
	Boosted logistic regression (BLR)	'LogitBoost'	nIter
(2) Decision Trees	Simple decision Tree (DT)	'rpart'	cp
	Bagged decision Tree (BDT)	'treebag'	-
	Bagged AdaBoost (ABDT)	'AdaBag'	mfinal maxdepth
	Boosted Classification Trees (ADT)	'ada'	iter maxdepth nu nrounds
	eXtreme Gradient Boosting (XBDT)	'xgbDART'	max_depth eta gamma subsample colsample_bytree rate_drop skip_drop
	Random Forest (RFDT)	'rf'	mtry
	Conditional Inference Random Forest (CFDT)	'cforest'	mtry
(3) Support Vector Machines	Linear Kernel (LSMV)	'svmLinear'	C
	Polynomial Kernel (PSMV)	'svmPoly'	degree scale C
	Radial Basis Function Kernel (RSMV)	'svmRadial'	sigma C
(4) Artificial Neural Networks	Neural Networks (ANN)	'nnet'	size decay
	Monotone Multi-Layer Perceptron Network (MMNN)	'monmlp'	hidden1 n.ensemble

4.5.3.1 Regression Models

Regression models are often considered as the most basic type of supervised machine learning algorithm, predicting the relationship between a dependent variable and the input matrix of independent variables (Bishop 2006; Rebala, Ravi, and Churiwala 2019). The most basic regression model—linear regression—uses a linear combination between the independent variables to explain the outcome of the dependent variable (thus, assuming a linear relationship) (Witten and Frank 2005). However, as our variable of interest is binary—it is either a falsification (1) or not (0)—we use logistic regression models instead. This classification method relies on a logistic (Sigmoid) function, which predicts a probability score between 0 and 1 (Rebala, Ravi, and Churiwala 2019). Model training is done by adjusting the coefficients to maximize the likelihood of explaining the observed data (Hastie, Tibshirani, and Friedman 2009).

This learning principle can be extended using boosting, an ensemble method that combines many weaker learners, i.e., simple methods with a low accuracy, into a stronger classifier (Schapire and Freund 2012). In the case of boosted logistic regression, several simple logistic regression models are applied sequentially to reweight the training data and adjust the prediction by weighted majority vote. As a result, the residual error of the model is reduced and the resulting model will flexibly fit non-linear relationships (Friedman, Hastie and Tibshirani 2000). To avoid overfitting and thus increase the generalizability of the models, we will also rely on Ridge (L2) and Lasso (L1) regularized regressions. In this case, the variance is reduced by shrinking the regression coefficients. This is done by adding a penalty to the cost function (Hoerl and Kennard 1970; Tibshirani 1996).

4.5.3.2 Decision Trees

In general, decision trees are algorithms which build the model based on a series of deterministic decisions partitioning i.e., splitting the entire dataset (root node) into different branches (sub-datasets), which therefore form a tree structure. Each decision (node) about the best split is made based on a mathematical criterion selecting a feature for the split (Lantz 2019; Rebala, Ravi, and Churiwala 2019). The best split is defined as the point at which the data yields the highest information gain, e.g., the most homogeneous grouping of the data (Kern et al. 2019). In our case the best classification of falsifications and nonfalsifications. In classification problems the most common criteria are the cross-entropy or the Gini index (Bishop 2006). The algorithm continues creating branches until a certain stopping criterion is met, for example, a maximum depth or a minimum number of samples in a node. The last decision results in the

leaf nodes, which denote the resulting classification based on the combination of all decisions (Lantz 2019).

The simplest version of this concept, the CART (Classification and Regression Trees) algorithm (Breiman et al. 1984), relies on the Gini index. At each node of the tree, CART iterates through all potential splits based on the features, aiming to maximize the reduction of the Gini index, effectively partitioning the data into subsets that exhibit higher homogeneity of classes (Kern et al. 2019). This concept of CART can be extended by the ensemble method bagging (Breiman 1996). Bagged CART uses bootstrap aggregation to enhance the predictive performance, i.e., creating multiple CART models by resampling the original dataset. Each tree is built independently using different subsets of the data, allowing for variability in the training process. The trees are ensembled through majority voting, effectively mitigating overfitting, yielding a more stable and accurate classification model. Similar to the regression models, we also employ the ensemble method boosting to decision trees (Friedman, Hastie, and Tibshirani 2000). Boosted Trees are sequentially trained decision trees. Each tree in the sequence aims to correct the errors made by its predecessors by focusing on observations with larger residuals. Hence, trees are weighted and combined, enhancing overall model performance.

An algorithm that combines both, the idea of bagging and boosting, is bagged AdaBoost (extension of AdaBoost; Freund and Schapire 1997). Again, the idea is to create a stronger learner by combining multiple models: Bagged AdaBoost involves training multiple models on bootstrapped samples of the dataset, leveraging both AdaBoost's sequential learning and bagging's resampling to enhance accuracy, reduce variance and improve overall robustness of the classfication (Alfaro, Gamez, and Garcia 2013). Another iterative ensemble learning technique—eXtreme Gradient Boosting or XGBoost—leverages the gradient boosting framework introduced by Friedman, Hastie, and Tibshirani (2000) and Friedman (2001): It sequentially builds a series of decision trees, each focusing on minimizing classification errors made by the preceding trees. This is done using the gradient descent optimization algorithm, assigning weights to correct misclassifications (Chen and Guestrin 2016). Additionally, it integrates regularization techniques to control model complexity and reduce overfitting, contributing to enhanced predictive accuracy. Hence, XGBoost provides a robust and efficient solution for classification problems, demonstrating superior performance across various domains.

Additional to these algorithms, we also use Random Forest and its extension Conditional Inference Random Forest. Unlike XGBoost and AdaBoost, which focus on iteratively building

a stronger learner, Random Forest relies on the combination of multiple decision trees by bootstrapping the dataset (similar to bagging) and considering random subsets of features at each split (Lantz 2019). Each tree independently learns patterns and collectively contributes to the final prediction through majority voting (Breiman 2001). Due to this, Random Forest mitigates overfitting, enhances robustness, and is able to provide feature importance metrics. Conditional Inference Random Forest is an extension of the Random Forest algorithm that includes statistical tests for variable selection (Hothorn, Kornik, and Zeileis 2015).

4.5.3.3 Support Vector Machines

The learning principle of Support Vector Machines (SVMs) is very different from the aforementioned concept of decision trees, but somewhat related to regression principles. Generally speaking, SVMs are (often highly complex) algorithms that use high-dimensional spaces in order to describe the relationship between input features and the outcome (Lantz 2019). Within an N-dimensional space it aims to identify an optimal hyperplane i.e., a flat boundary, which best partitioners the data based on the maximum distance or margin between the data points (Cortes and Vapnik 1995; Lantz 2019). However, as classification problems are often non-linear, SVMs further rely on geometric properties, using non-linear mapping to map the input into a high-dimensional space (Stitson et al. 1996). This is done using kernel functions, which define the type of SVM. We rely on three different kernel functions: A simple linear kernel, a polynomial kernel, and a radial kernel.

4.5.3.4 Artificial Neural Networks

Similar to concepts of bagging or boosting, the basic idea of Artificial Neural Networks (ANNs) is to combine many weaker learning elements to solve complex problems by learning patterns within the data. However, ANNs are especially closely related to SVMs (see Aggarwal 2018 for more details). Inspired by the neural structure of the human brain, ANNs link nodes (similar to neurons) to process information and build models through (mostly multiple) layers (Lantz 2019). In this process, input signals are weighted according to their importance and accordingly connect the nodes within layers via weighted edges. Further, the weighted and summarized input is processed according to an activation function (Aggarwal 2018).

We apply two different implementations of neural networks: first, a standard artificial neural network and second, a monotone multi-layer perception network. The former uses the sigmoid function as activation function whereas the later uses an adapted activation function which is specifically adapted for monotone input-output relationships in the network (Lang 2005).

4.5.4 Parameter Tuning and Result Evaluation

Beside the decision on algorithms and respective features, parameter tuning as well as performance evaluation metrics are another crucial part in machine learning (Hoffmann et al. 2019). Both are necessary in order to evaluate the resulting classifications, choosing the best model, applying them to the testing data, and finally evaluate which algorithm worked best in detecting falsifications. We will rely on two widely used metrics: The ROC curve (Receiver Operating Characteristic) together with AUC (Area under the ROC curve) and the F1-score (see Tharwat 2020 for an overview). These metrics will aid in evaluating each model and its respective tuning parameters in the training data as well as in the final evaluation of models in the test data. Values of the used Tuning Parameters (see **Table 4.4** for the list of tuned parameters and **Table A 4.1** for the respective values) were chosen arbitrarily based on their possible value range to find the best performing model. Further, we also calculate and interpret additional performance metrics, namely False-Positive and False-Negative Rates, and the Accuracy. **Table 4.5** presents an overview of how to calculate the required metrics. See Fawcett (2005) for a detailed discussion of the calculated metrics.

Table 4.5: Overview of formulas and principles for evaluation metrics.

Evaluation metric	**Formula/Principle**	
False-Positive Rate	$FP_{rate} = \frac{FP}{TN + FP} = 1 - Q_{spec}$	(4.2)
False-Negative Rate	$FN_{rate} = \frac{FN}{TP + FN} = 1 - Q_{sens}$	(4.3)
Specificity	$Q_{spec} = \frac{TN}{TN + FP} = 1 - FP_{rate}$	(4.4)
Sensitivity/Recall	$Q_{sens} = \frac{TP}{TP + FN} = 1 - FN_{rate}$	(4.5)
Precision	$Q_{prec} = \frac{TP}{TP + FP}$	(4.6)
Accuracy	$Q_{acc} = \frac{TP + TN}{TP + TN + FP + FN}$	(4.7)
F1-Score	$F1_{score} = 2 * Q_{prec} * Q_{sens} / (Q_{prec} + Q_{sens})$	(4.8)
Receiver Operating Characteristic (ROC)	*two-dimensional graph, depicting:* $Q_{sens}\ vs.\ 1 - Q_{spec}$	
Area Under the Curve (AUC)	$\int_0^1 ROC$	

Note: FP = False-Positive Cases, FN = False-Negative Cases,
TP = True-Positive Cases, TN = True-Negative Cases

There are some important differences between the F1-Score and the ROC/AUC metric. The F1-Score is calculated based on the outcomes of the resulting confusion matrix, i.e., the final binary classification. This binary classification depends on the chosen probability threshold. In our data, we rely on a threshold of 0.5[12], meaning that every case with a predicted probability of at least 0.5 will be classified as "falsification", whereas every case with a lower predicted probability will be classified as "real interview". In contrast, the ROC metric is instead based on the raw estimated class probability of the respective models. It is created before the binary classification takes place and hence also takes the trade-off between possible thresholds into account. Based on the individual class probability of each case, it depicts the cases impact on model sensitivity (or recall) and specificity. The AUC uses this result to quantify the model's overall discriminatory power (Hoffmann et al. 2019). Therefore, the plot of the ROC curve (sometimes called sensitivity/specificity plot) allows for assessing the interplay of false-positive and true-positive classifications based on the probability estimation (Lantz 2019).

As this metric is mainly based on the positive class, it is usually insensitive to skewed class distributions and changes in the distribution (Fawcett 2005; Tharwat 2020). In comparison, the F1-Score considers both, false-positive cases as well as false-negative cases. Hence, it captures the precision-recall trade off within one metric (Lantz 2019). But as it focusses on the positive as well as the negative class, it is sensitive to imbalanced data and changes in class distribution (see Tharwat 2020 for a detailed discussion of the different relationships between the metrics). As our data includes balanced classes for the experimental data and imbalanced data for the real-world data, we use both metrics, in order to take these possible differences into account.

Lastly, we will consider the False-Positive and False-Negative Rates of the models. In the case of falsification classification, it could be argued that false-negative cases, i.e., falsifications that have been overlooked, may be more problematic than a properly conducted interviewer being classified as falsification. This is because these results should be verified through non-statistical methods, false-positive cases can still be accurately classified retrospectively. Error-costs of false-negatives in terms of data quality are hence higher. To consider this factor, further Error-Rates are used beside the ROC/AUC and F1-Score.

[12] For simplicity and comparability between F1-Score and ROC, we rely on a threshold of 0.5 for both metrics, even though ROC would allow to adapt the threshold according to its results. As robustness check we also compared model performance after adjusting the threshold for ROC. However, we did only find insignificant differences between results which were unsystematic in nature.

4.6 Results

4.6.1 Descriptive Results and Feature Importance

To get a first impression about our features, their discriminatory power, and possible similarities and dissimilarities between real-world and experimental data we first compare the respective falsification indicators. In order to get models with high classification performance, falsification indicators should ideally be able to differentiate between real interviews and falsifications. Hence, **Figure 4.2** shows the mean indicator value for falsifications (indicated with the triangle) and real interviews (indicated with the square), separate for the real-world data and the experimental data.

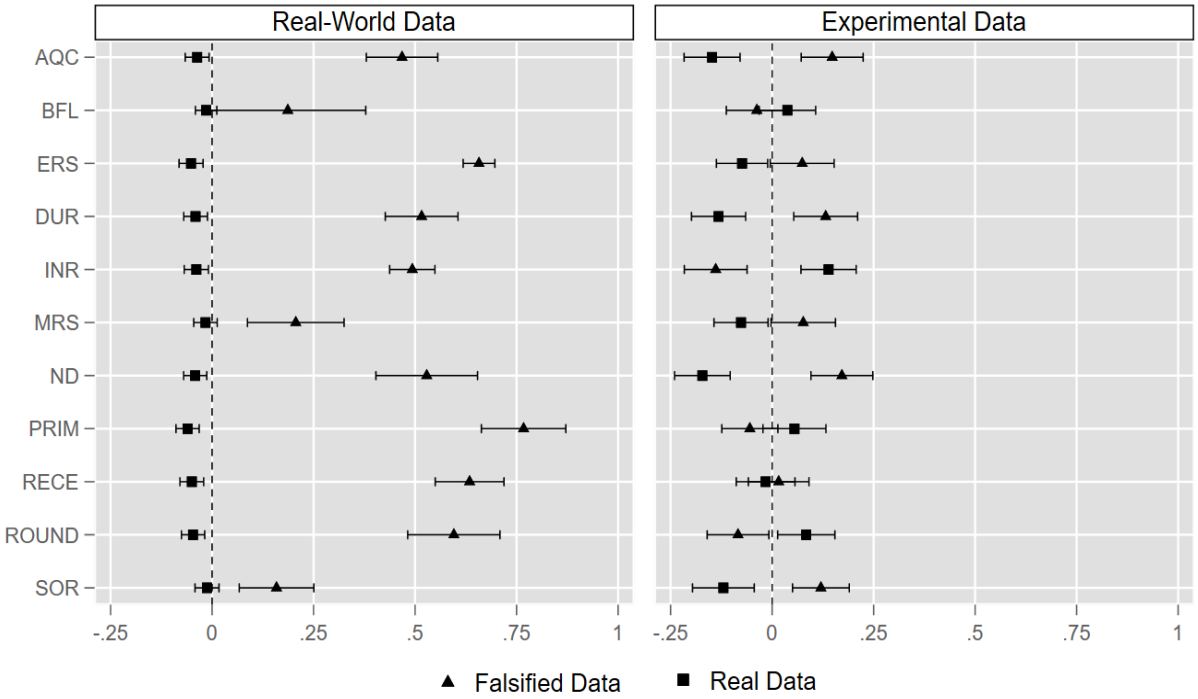

Figure 4.2: Mean indicator values of all falsified and real interviews, separate for real-world and experimental data.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Interestingly, the falsification indicators do much clearer discriminate between the falsifications in the real-world data compared to the experimental data. For the real-world data, all indicators values coming from falsified interviews (except for BFL) are significantly different from zero and all hint into the positive direction. As indicators were coded in a way, that positive values show higher suspicion according to literature, we find that the real-world falsifiers confirm the assumptions of the literature. In contrast, non-fabricated interviews of the real-world data are characterized by values around zero with a slight negative tendency, hence

do not indicate fraudulent behavior. In the case of the real-world data, indicator PRIM shows the largest differences, compared to BFL and SOR showing nearly no differences. For the experimental data, differences are much smaller, with some indicators (BFL, INR, PRIM, and ROUND) even hinting into the other direction compared to the real-world data. Only the indicators ACQ, DUR, INR, ND, ROUND and SOR show significant differences between falsified and real interviews. Importantly, this graphical examination only depicts a limited level of possible correlations. Dependencies between the indicators could be far more complex, which is why they might only be revealed by machine learning algorithms. Hence, we further evaluate the feature importance of the falsification indicators using Boruta.

Boruta is an algorithm which manipulates the input to random forest in order to determine the importance of the models features. The process is done by comparing the importance of each input variable against a shadow set of randomized variables. In other words, it creates a set of randomly generated variables, in order to assess whether real variables/features are less relevant than a random probe and evaluates the ranking and importance of each respective feature (Kursa and Rudnick 2010). To examine possible differences between our datasets, we used Boruta on 1) the real-world data and 2) the experimental data.

Figure 4.3 and **Figure 4.4** show the respective feature importance for the real-world survey data and the experimental data. In summary, none of the falsification indicators was deemed unimportant (below the min value). However, for the experimental data the importance of four indicators was very close to the randomly generated variables (MRS, PRIM, RECE, SOR), visible through the max value. As they, on the one hand, lie above the min value but are not significantly different from the max value, this means their explanatory contribution was only very minor. Further, the evaluation of the relative feature importance varies between the experimental data and the real-world data. The only indicators that seem to be of high importance in both data sources are: ACQ, BFL, DUR, ERS, INR, and ND. Since no variable was classified as unimportant, our results will still include all features.

4.6.2 Research Question 1a (RQ1a)

In order to answer RQ1a, i.e., how effective the algorithms are in detecting falsifications in the experimental survey data, we first train and afterwards test the 14 different algorithms based on a random split of the experimental data, as presented in **Figure 4.1**. The results are based on the final selected models, selected during the tuning phase.

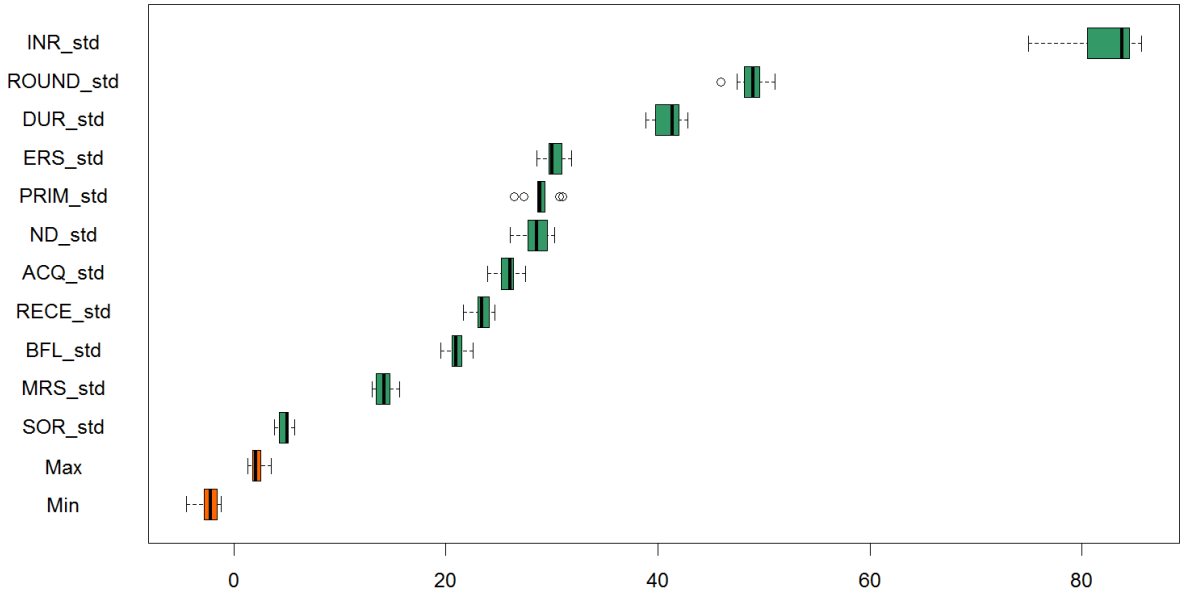

Figure 4.3: Feature importance of indicators according to Boruta algorithm.
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

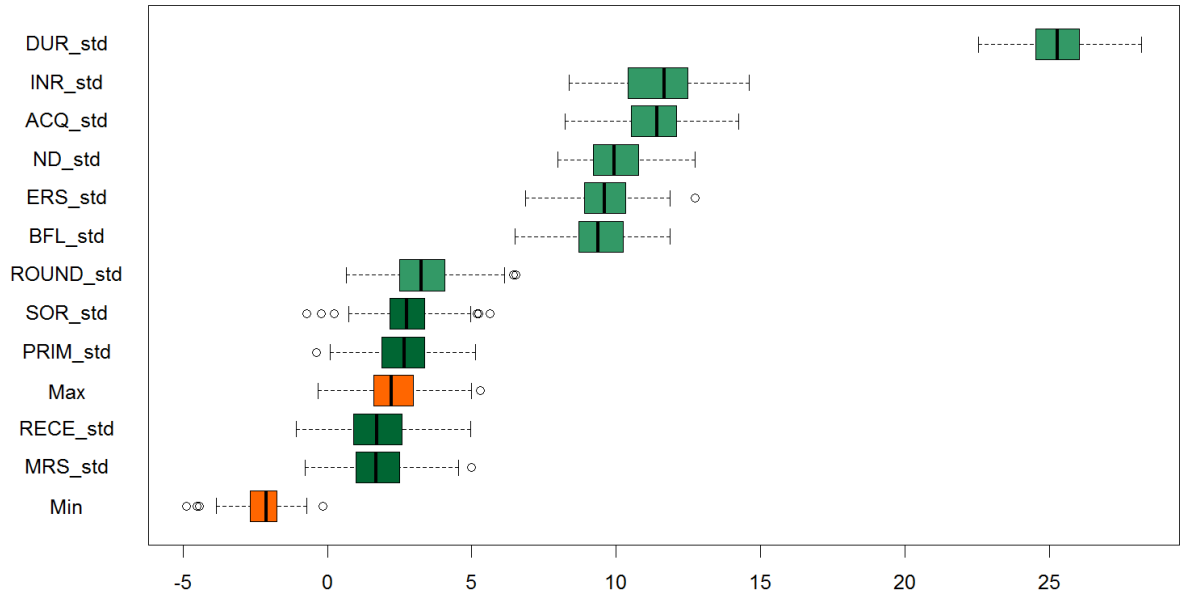

Figure 4.4: Feature importance of indicators according to Boruta algorithm.
Source: Experimental data, University of Giessen, 2011.

4.6.2.1 Training of Algorithms

Starting with the results of models based on the F1-Score, we find strong variation between the different respective algorithms. **Figure 4.5** shows that simple Decision Trees (DT) and the Support Vector Machines with a Polynomial Kernel (PSMV) and a Radial Basis Function Kernel (RSMV) show the lowest performance. Because all three classified all falsifications as real interviews, the F1-Score could not be calculated (see Appendix **Table A 4.2** and **Table A 4.10**for more details on the confusion matrix and the different performance

measures). Hence, performance of these algorithms is low. On the other hand, we also find three algorithms with very strong performance outcomes: Bagged Decision Tree (BDT; F1-Score of 1.00), Bagged AdaBoost (ABDT; F1-Score of 1.00), and Random Forest (RFDT; F1-Score of 1.00). BDT and ABDT were able to reach perfect classification, meaning that every falsification and real interview was predicted correctly. RFDT predicted one falsification incorrect as a real interview (false-negative). Most other models show very similar, moderate performance outcomes around an F1-Score of 0.58 to 0.79. Logistic Regression (LR) performed a little bit lower, with a F1-Score of 0.32, whereas Monotone Multi-Layer Perceptron Network (MMNN) showed a stronger performance with an F1-Score of 0.90.

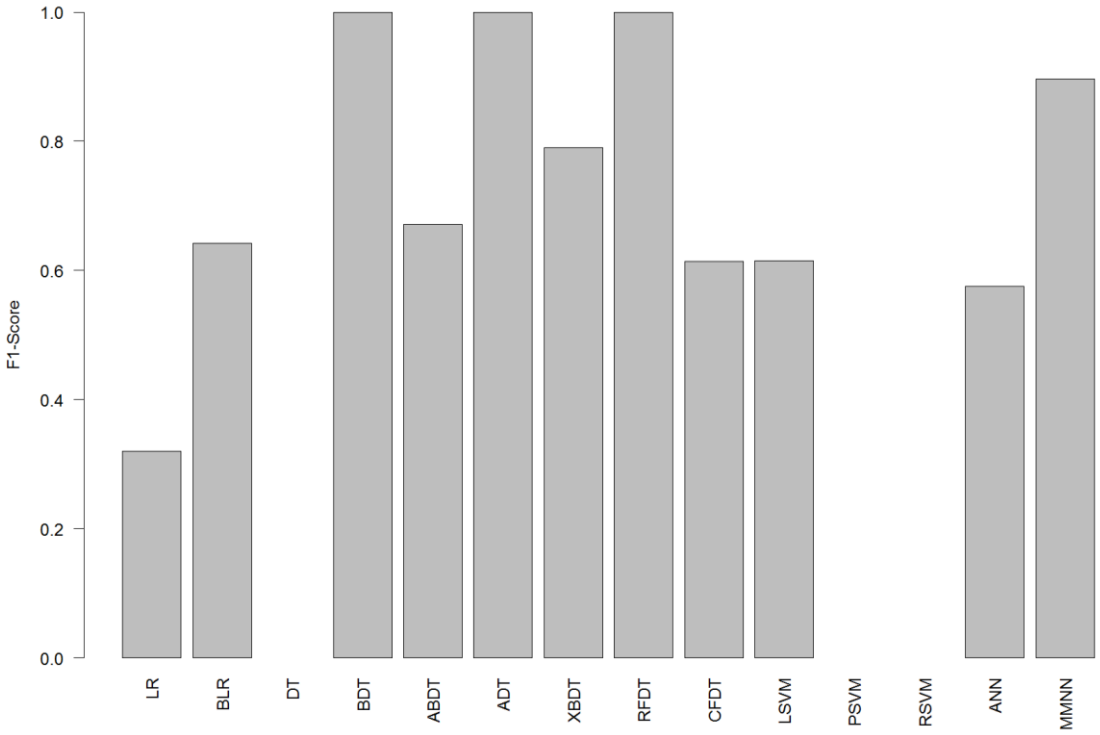

Figure 4.5: F1-Scores of selected models, training data (RQ1a).
Source: Experimental data, University of Giessen, 2011.

Turning to the results from the ROC/AUC metric (**Figure 4.6**), we find that the different models all show moderate performance, with some slight differences (see Appendix **Table A 4.3** and **Table A 4.10** for more details). As for the F1 models, BDT (AUC of 0.68) and RFDT (AUC of 0.70) reach a strong performance. With a probability threshold of 0.5 (i.e., every case with a probability of at least 0.5 is classified as falsification), both BDT results in a perfect classification of all interviews and falsifications and RFDT only classified one falsification incorrect (false-negative). Similarly, ABDT showed a low false-negative rate of only 5% (i.e., only 26 false-negative cases), resulting in an AUC of 0.66. However, the results also

demonstrate, why the AUC needs to be treated with caution in some settings. As the ROC is based on the interplay of Sensitivity and Specificity, there are also some models which show a high-performance evaluation in terms of AUC but not in terms of their False-Positive and False-Negative Rate. For example, Conditional Random Forest (CFDT) reached an AUC of 0.68 but had a False-Positive Rate of 15% and a False-Negative Rate of 18%. eXtreme Gradient Boosting (XBDT) also reached an AUC of 0.68 but had a False-Positive Rate of 25% and a False-Negative Rate of 33%.

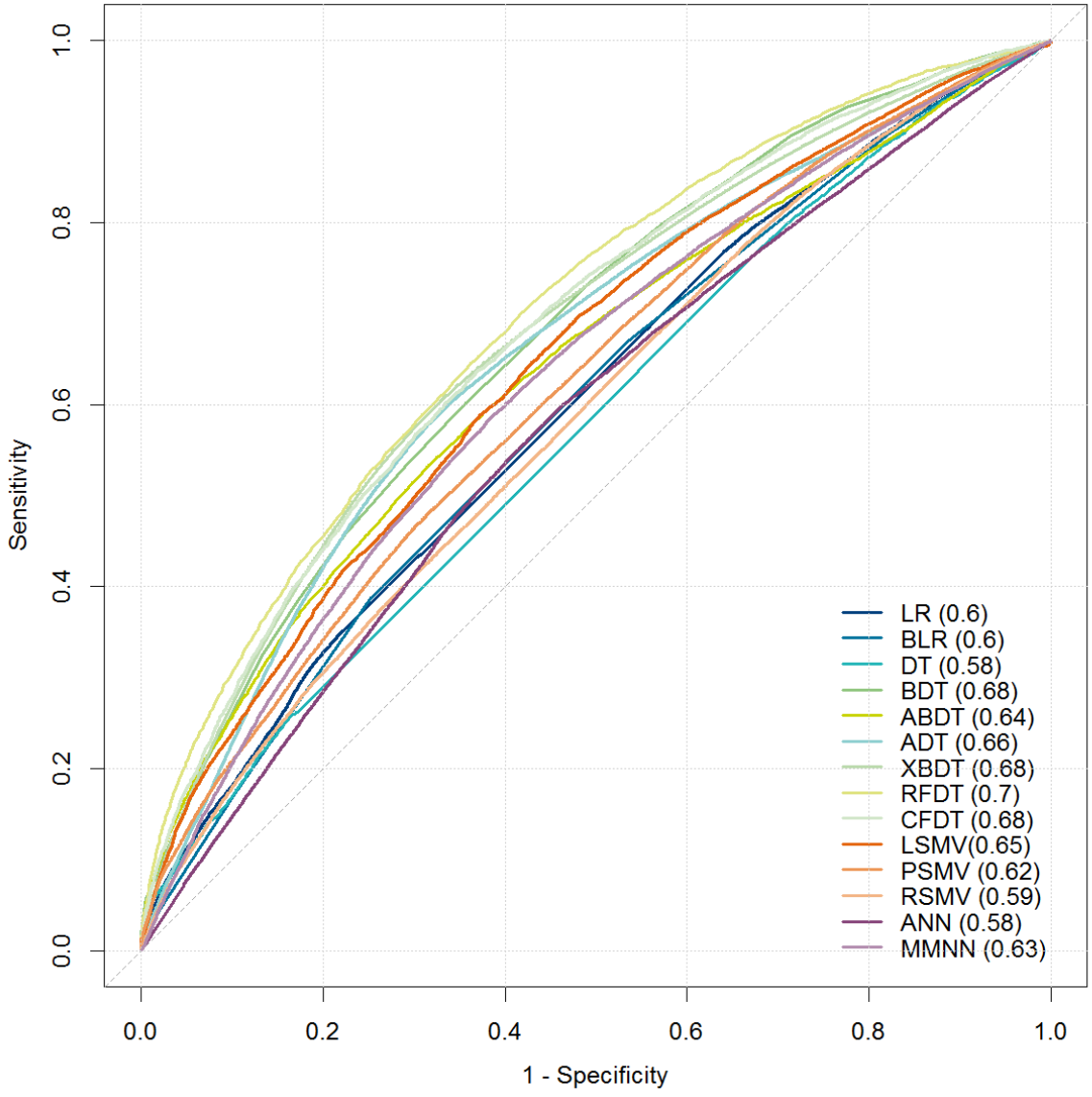

Figure 4.6: ROC curve of selected models, training data (RQ1a).
Source: Experimental data, University of Giessen, 2011.

4.6.2.2 Testing of Algorithms

Turning to the final evaluation based on the F1-Score, we find very similar results to the ones based on the initial training stage (see **Figure 4.7** and **Table 4.6**). Again, DT, PSMV, and RSMV could not be calculated, as they classified all cases as real interviews. Similarly, BDT and RFDT showed the best performance with an F1-Score of 0.71 and 0.70, respectively. However, BDT performed a bit better in terms of the false-positive rate, as it only misclassified 28% of the falsifications whereas RFDT misclassified 36% of falsifications (see Appendix **Table A 4.14** for the confusion matrix). At the same time, BDT classified 33% of real interviews as falsifications, compared to only 23% with RFDT. Contrary to the training results, the Boosted Logistic Regression (BLR; F1-Score of 0.69) performed third best, because it only overlooked 21% of falsifications. At the same time, the False-Positive Rate is quite high with 58%. All other algorithms—except for LR—show moderate performance, varying between a F1-Score from 0.58 to 0.65. Taken together, the group of models based on decision tress seemed to reach the highest performance in comparison to the other groups.

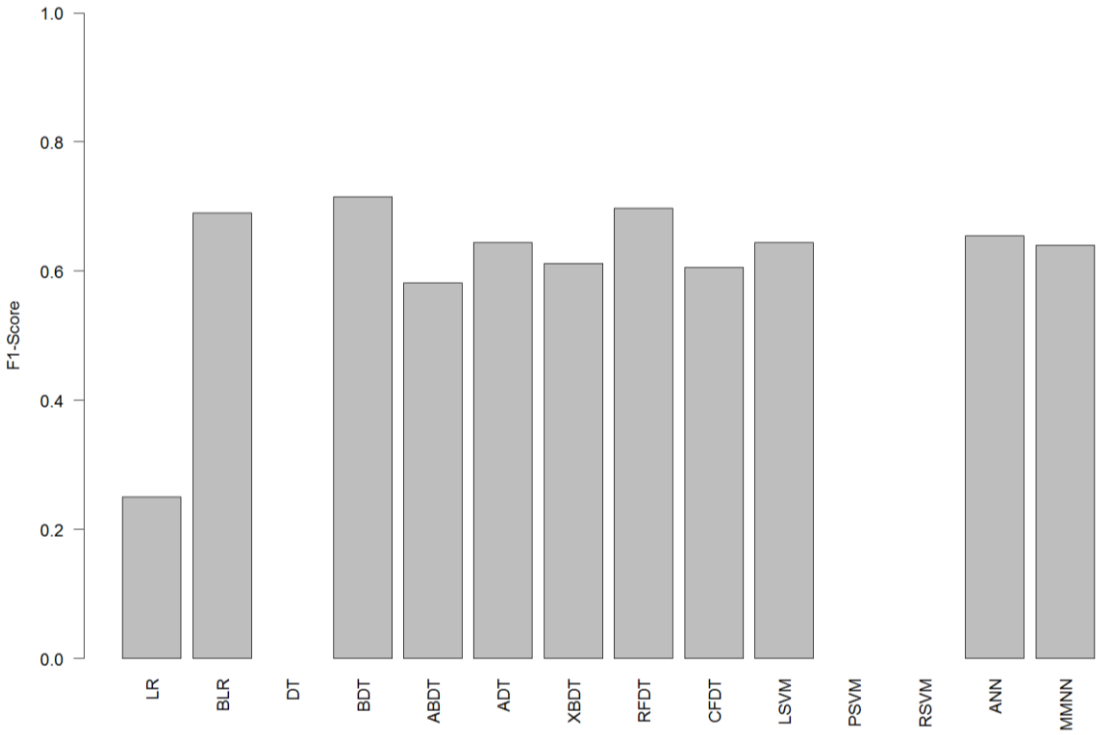

Figure 4.7: F1-Scores of selected models, test data (RQ1a).
Source: Experimental data, University of Giessen, 2011.

Results for the ROC/AUC metric (**Figure 4.8**) closely mirror the findings of the training stage and the findings using the F1-Score on the testing data. The highest AUC is reached by RFDT (AUC of 0.71), BDT (AUC of 0.69), and ABDT (AUC of 0.69), closely followed by

ADT, CFDT, and XBDT. Again, models based on decision trees reached the best performance. However, all other algorithms reached very similar performance, as all algorithms had False-Negative and False-Positive Rates around 30%, except for BLR with a lower False-Negative Rate but a much higher False-Positive Rate falsifications (see Appendix **Table A 4.15** for the confusion matrix). Taken together, we can conclude that falsifiers within the experimental data indeed produced distinct patterns which were detectable using the different supervised machine learning algorithms. Especially algorithms based on decision trees showed good performance for these data.

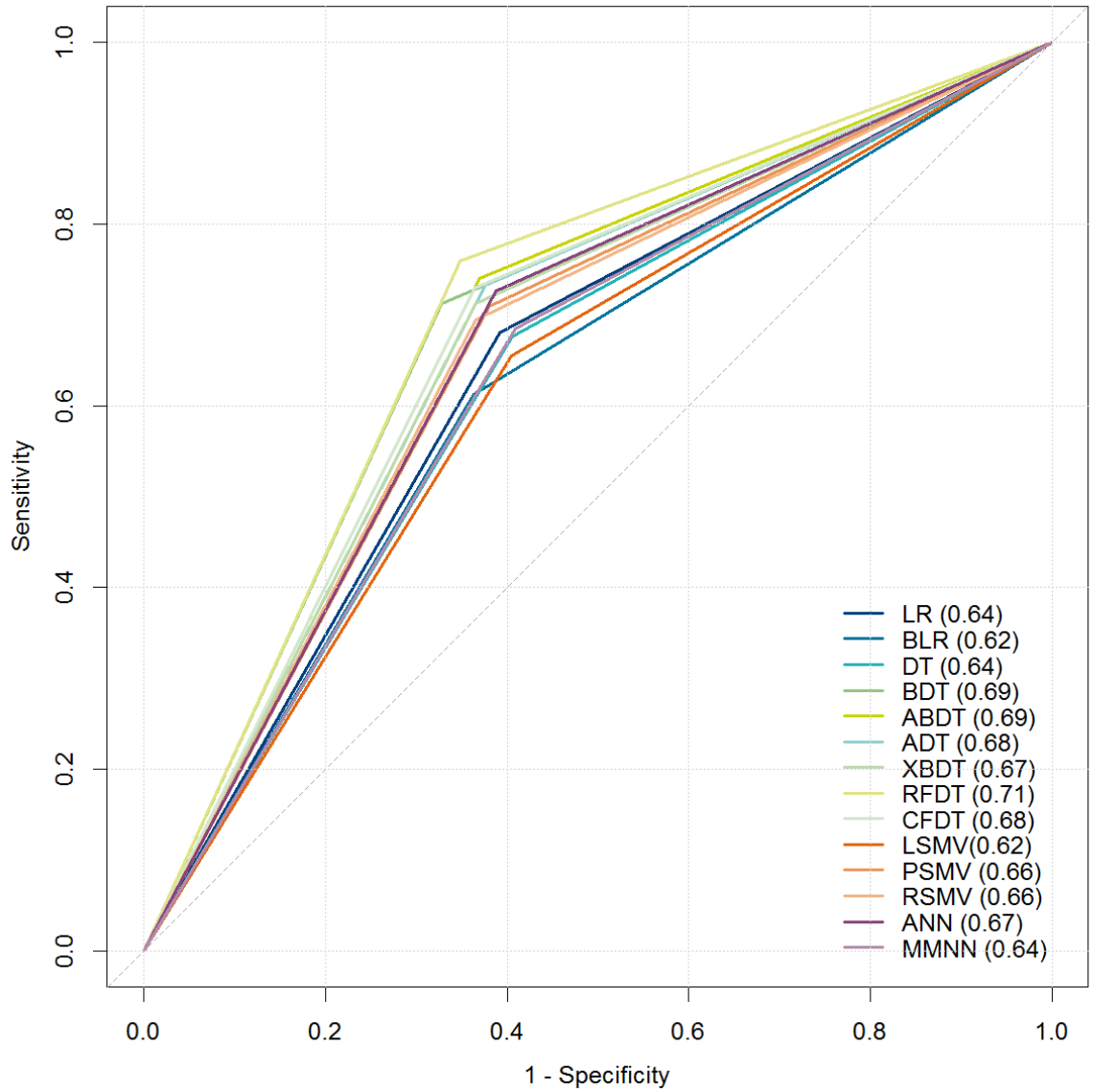

Figure 4.8: ROC curve of selected models, test data (RQ1a).
Source: Experimental data, University of Giessen, 2011.

Table 4.6: Final performance measures of selected models, test data (RQ1a).

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.07	0.85	0.15	0.93	0.72	0.51	0.25
BLR	0.58	0.21	0.79	0.42	0.61	0.62	0.69
DT	0.00	1.00	0.00	1.00	-	0.46	-
BDT	0.33	0.28	0.72	0.67	0.71	0.69	0.71
ABDT	0.20	0.52	0.48	0.80	0.74	0.63	0.58
ADT	0.36	0.38	0.63	0.64	0.66	0.63	0.64
XBDT	0.31	0.44	0.56	0.69	0.67	0.62	0.61
RFDT	0.23	0.36	0.64	0.77	0.76	0.70	0.70
CFDT	0.23	0.48	0.52	0.77	0.72	0.64	0.61
LSMV	0.38	0.37	0.63	0.62	0.66	0.63	0.64
PSMV	0.00	1.00	0.00	1.00	-	0.46	-
RSMV	0.00	1.00	0.00	1.00	-	0.46	-
ANN	0.25	0.41	0.59	0.75	0.73	0.67	0.65
MMNN	0.38	0.38	0.63	0.62	0.66	0.62	0.64
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	*AUC*
LR	0.34	0.37	0.63	0.66	0.68	0.64	0.65
BLR	0.58	0.21	0.79	0.42	0.61	0.62	0.62
DT	0.33	0.39	0.61	0.67	0.68	0.63	0.64
BDT	0.33	0.28	0.72	0.67	0.71	0.69	0.69
ABDT	0.25	0.38	0.62	0.75	0.74	0.68	0.69
ADT	0.26	0.39	0.61	0.74	0.73	0.67	0.68
XBDT	0.30	0.35	0.65	0.70	0.71	0.67	0.67
RFDT	0.23	0.36	0.64	0.77	0.76	0.70	0.71
CFDT	0.27	0.36	0.64	0.73	0.73	0.68	0.68
LSMV	0.39	0.36	0.64	0.61	0.66	0.63	0.63
PSMV	0.30	0.38	0.63	0.70	0.71	0.66	0.66
RSMV	0.34	0.33	0.67	0.66	0.69	0.67	0.66
ANN	0.26	0.41	0.59	0.74	0.73	0.66	0.67
MMNN	0.31	0.41	0.59	0.69	0.68	0.63	0.64

Source: Experimental data, University of Giessen, 2011.

4.6.3 Research Question 1b (RQ1b)

4.6.3.1 Training of Algorithms

To answer RQ1b, i.e., how effective the algorithms are in detecting falsifications in the real-world survey data, we again train and test the different algorithms, this time based on a

random split of the real-world data. Starting again with the results of the F1-Score for the initial real-world training data, we find that the average performance is even higher as for the experimental training data. **Figure 4.9** shows that—similar to the results for RQ1a—BDT and RFDT show the best performance (F1-Score of 1.00), with nearly perfect classification for all cases. As Appendix **Table A 4.11** further shows, BDT has a False-Negative Rate of 1%, i.e., only 1% of all falsifications were overlooked, and RFDT classified all cases correctly. Besides, XBDT showed a high F1-Score of 0.97 with only 5% misclassified falsifications and 0% misclassified real interviews. The lowest F1-Score is reached by LR and LSVM with an F1-Score of 0.53 and 0.52 respectively. Importantly, all algorithms show an extreme low False-Positive Rate. However, this is also caused by the unbalanced class size between real interviews and falsifications. As a result, the False-Negative Rate is very important in the context of this data. As an example, ADT was able to reach an F1-Score of 0.77 with a False-Positive Rate of only 1%. On the other hand, the False-Negative Rate was 32%. For the confusion matrix, see Appendix **Table A 4.4**. Still, all algorithms show good to moderate results in terms of the F1-Score.

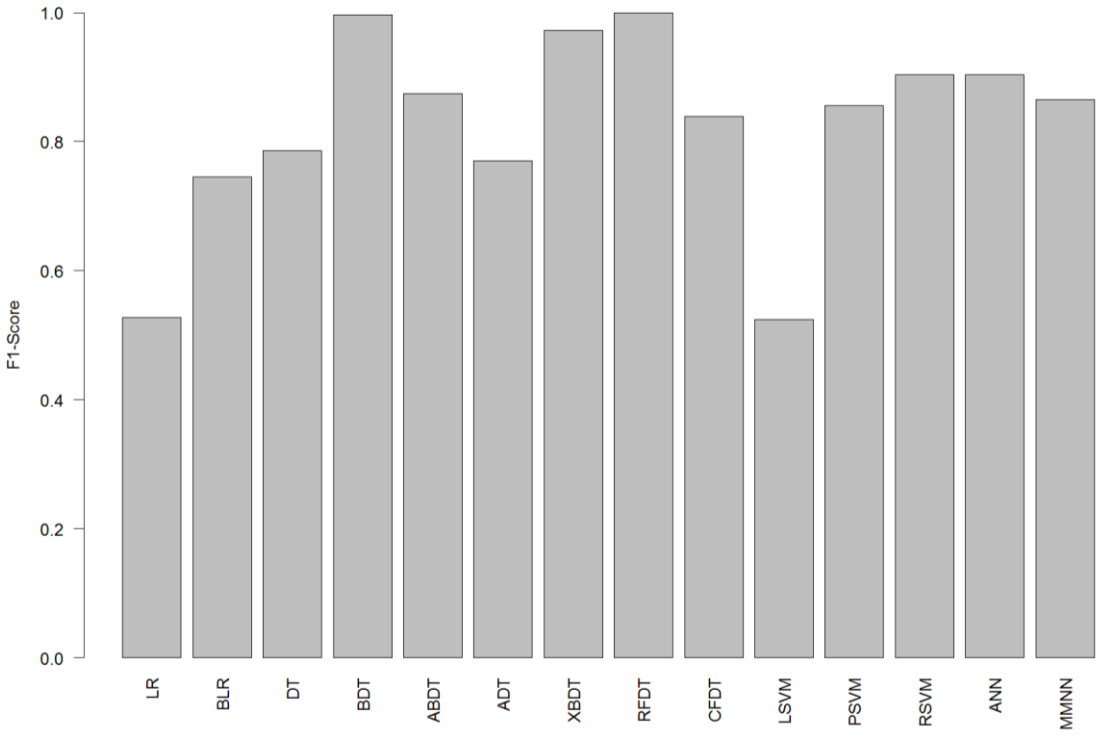

Figure 4.9: F1-Scores of selected models, training data (RQ1b).
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Again, the results of the ROC/AUC metric show similarities to the results of RQ1a (see **Figure 4.10**). Except for LR (AUC of 0.78), DT (AUC of 0.77), and ABDT (AUC of 0.77), all

algorithms show a very high performance. The best performance is reached by RFDT and CFDT (both AUC of 0.95), followed by MMNN (AUC of 0.94), BDT (AUC of 0.93), and BLR as well as ANN (both AUC of 0.92). Looking at the False-Positive and False-Negative Rates at a probability threshold of 0.5 (see Appendix **Table A 4.11**; confusion matrix **Table A 4.5**), however, shows that some of these high performing models have high False-Negative Rates, even though the False-Positive Rates are very low: Some models classified up to 59% of all falsifications wrongly. Taking this into account, BDT, ADT and CBDT show the best classification results, with only 1- 3% overlooked falsifications. Still with some exceptions, nearly all models show a high performance.

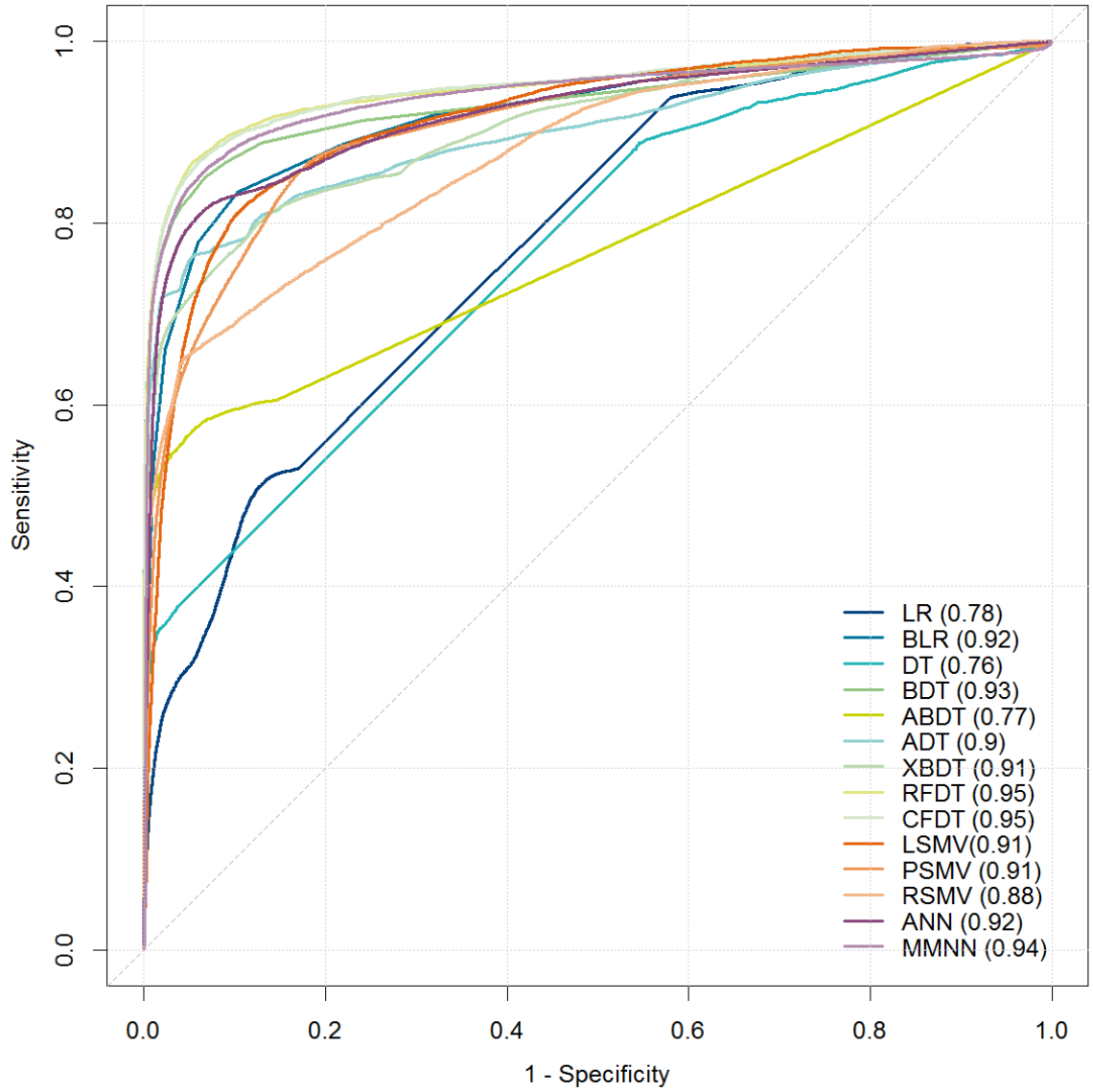

Figure 4.10: ROC curve of selected models, training data (RQ1b).
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

4.6.3.2 Testing of Algorithms

The final evaluation of the results from the training stage, based on the testing data reveals that predictions are more precise than for the experimental data in RQ1a. Even though we see a slight decrease in the performance in terms of the F1-Score compared to the training stage (see **Figure 4.11**) all algorithms reach a high performance. Except for LR and LSVM, which already showed a lower performance in the training stage, the F1-Scores lay between 0.73 and 0.84. The highest F1-Score was achieved by XBDT (F1-Score of 0.84), RFDT (F1-Score of 0.83), BDT (F1-Score of 0.82), ABDT, and MMNN (both F1-Scores of 0.81). This confirms the finding, that especially algorithms based on decision trees constantly show a high performance. As we have found, that due to the unbalanced class distribution False-Positive and False-Negative Rates are also important for the evaluation, **Table 4.7** gives more insights. As for the training stage, False-Positive Rates are all very low. There are, however, some noteworthy differences between the False-Negative Rates (see Appendix **Table A 4.16** for confusion matrix). Clearly, XBDT also shows the best performance in terms of False-Negative Rates, with only 21% overlooked falsifications. Even though the performance in the F1-Score is lower compared to others, BLR also only had a False-Negative Rate of 21% together with a False-Positive Rate of 3%. Most other False-Negative Rates vary around 30%.

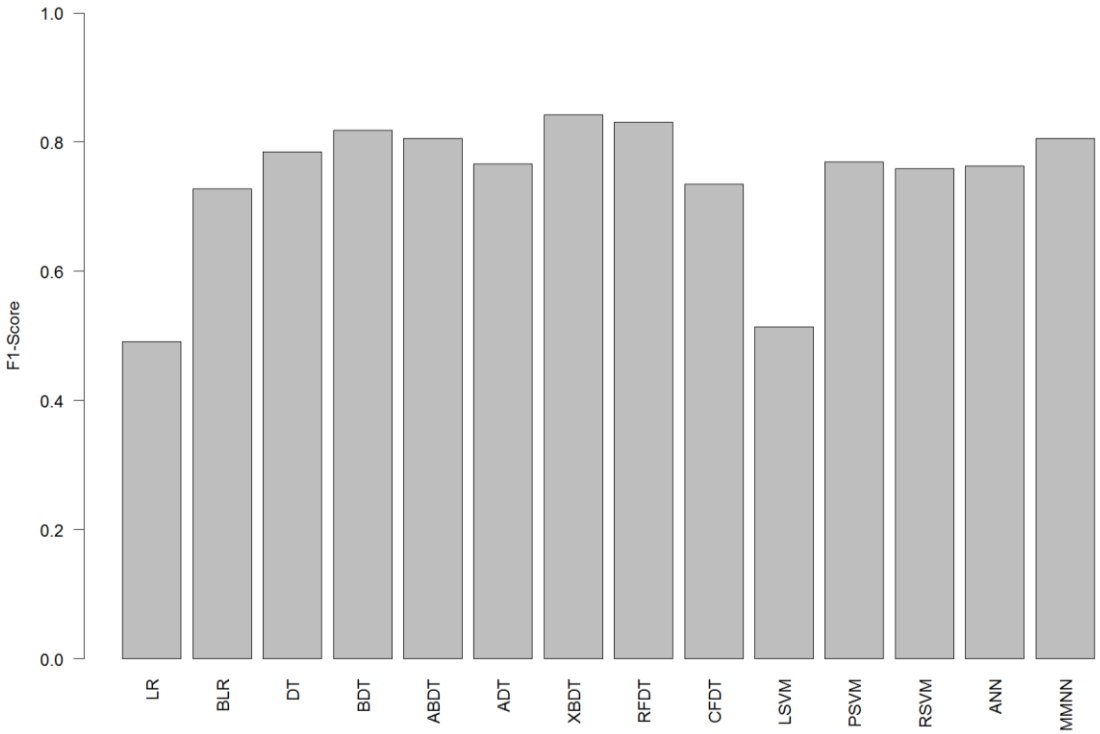

Figure 4.11: F1-Scores of selected models, test data (RQ1b).
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

As before with the results of the F1-Score, also the ROC/AUC metric shows very high performance. All AUCs lie above 0.80 (see **Table 4.7**). **Figure 4.12** shows that especially ABDT (AUC of 0.94), XBDT (AUC of 0.94), and BDT (AUC of 0.93) show high performance outcomes. Focusing solely on the AUC we find, that RFDT (AUC of 0.98) and CFDT (AUC of 0.96) show the best performance. As before, we should also consider the outcomes of the False-Negative Rates. Here, ADT is best performing with only 20% wrongly classified falsifications, followed by BLR with 21% overlooked falsifications, and BDT, ABDT, XBDT, and RSMV with 24% (see Appendix **Table A 4.17** for confusion matrix). In summary, nearly all algorithms were able to classify most cases correctly, with a slightly better performance of most decision tree-based algorithms. For RQ1b we can hence also conclude that in the real-world data falsifiers produced distinct patterns which were detectable using most algorithms.

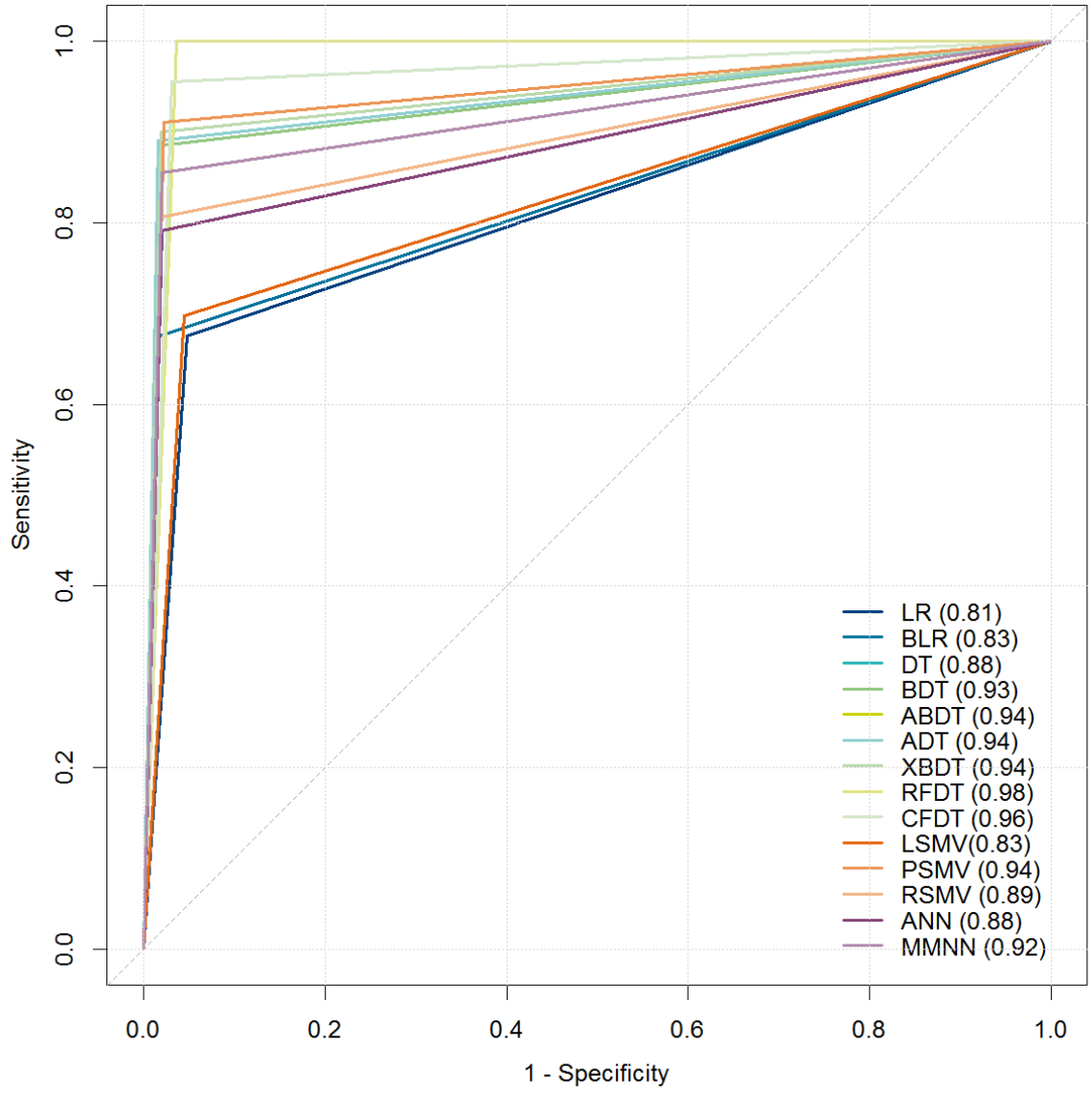

Figure 4.12: ROC curve of selected models, test data (RQ1b).
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table 4.7: Final performance measures of selected models, test data (RQ1b).

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.01	0.62	0.38	0.99	0.69	0.94	0.49
BLR	0.03	0.21	0.79	0.97	0.67	0.96	0.73
DT	0.01	0.28	0.72	0.99	0.86	0.97	0.78
BDT	0.01	0.24	0.76	0.99	0.89	0.98	0.82
ABDT	0.01	0.27	0.73	0.99	0.90	0.97	0.81
ADT	0.01	0.31	0.69	0.99	0.86	0.97	0.77
XBDT	0.01	0.21	0.79	0.99	0.90	0.98	0.84
RFDT	0.01	0.24	0.76	0.99	0.92	0.98	0.83
CFDT	0.00	0.39	0.61	1.00	0.93	0.97	0.74
LSMV	0.01	0.61	0.39	0.99	0.74	0.94	0.51
PSMV	0.01	0.30	0.70	0.99	0.85	0.97	0.77
RSMV	0.02	0.27	0.73	0.98	0.79	0.97	0.76
ANN	0.02	0.25	0.75	0.98	0.78	0.97	0.76
MMNN	0.01	0.24	0.76	0.99	0.86	0.97	0.81
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	AUC
LR	0.01	0.62	0.38	0.99	0.68	0.94	0.81
BLR	0.03	0.21	0.79	0.97	0.67	0.96	0.83
DT	0.02	0.25	0.75	0.98	0.79	0.97	0.89
BDT	0.01	0.24	0.76	0.99	0.89	0.98	0.93
ABDT	0.01	0.24	0.76	0.99	0.90	0.98	0.94
ADT	0.01	0.20	0.80	0.99	0.89	0.98	0.94
XBDT	0.01	0.24	0.76	0.99	0.90	0.98	0.94
RFDT	0.00	0.46	0.54	1.00	1.00	0.97	0.98
CFDT	0.00	0.39	0.61	1.00	0.96	0.97	0.96
LSMV	0.01	0.58	0.42	0.99	0.70	0.94	0.83
PSMV	0.01	0.28	0.72	0.99	0.91	0.97	0.94
RSMV	0.01	0.24	0.76	0.99	0.81	0.97	0.89
ANN	0.02	0.25	0.75	0.98	0.79	0.97	0.89
MMNN	0.01	0.25	0.75	0.99	0.85	0.97	0.92

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

4.6.4 Research Question 2 (RQ2)

4.6.4.1 Training of Algorithms

Addressing the question of the algorithms' effectiveness in detecting interviewer falsification caused by different falsifiers (RQ2), results of the F1-Score models in the initial training stage show that—similar to RQ1a and RQ1b—BDT (F1-Score of 1.00) and RFDT (F1-Score of 1.00) show very high model performance (**Figure 4.13**). Beside these two, ADT, RSVM and MMNN were also able to reach the same F1-Score. All five algorithms were able to classify all real-interviews correctly and misclassified no falsification up to 3 falsifications (out of 355), and had therefore nearly perfect classifications (see Appendix **Table A 4.6**). Besides these high-performance outcomes, LR and DT failed the classification task, as they classified all interviewers as real-interviews. All other algorithms showed a moderate high performance with F1-Scores ranging from 0.67 to 0.79 (see Appendix **Table A 4.12**).

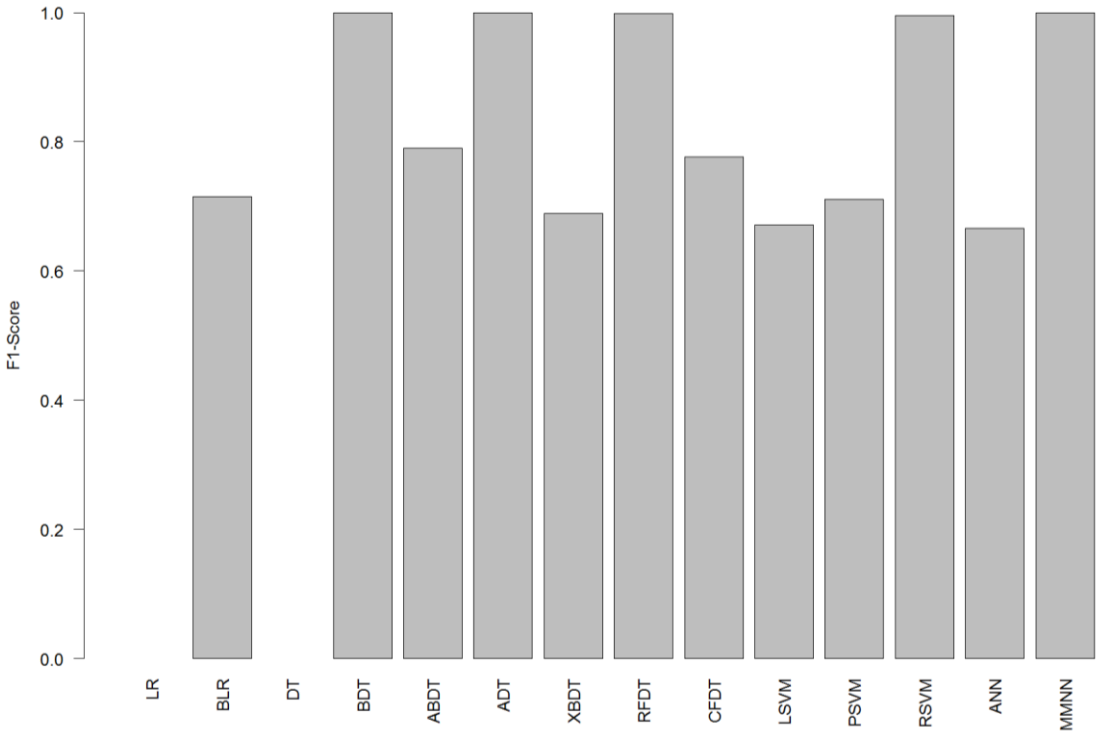

Figure 4.13: F1-Scores of selected models, testing data (RQ2).
Source: Experimental data, University of Giessen, 2011.

Results from the ROC/AUC metric (**Figure 4.14**) are nearly identical with the ones of RQ1a, only resulting in slightly higher AUCs for all algorithms (see Appendix **Table A 4.7**). RFDT (AUC of 0.75) has the highest AUC showing a perfect classification if we use a probability threshold of 0.5. The same holds for BDT with a slightly lower AUC of 0.72. In terms of the AUC, CFDT (AUC of 0.72), XBDT (AUC of 0.71), ADT (AUC of 0.70), and

LSMV (AUC of 0.70) also show high performance, demonstrating again the efficiency of decision tree models. Even without a high AUC, ADBT additionally showed very low False-Positive and False-Negative Rates.

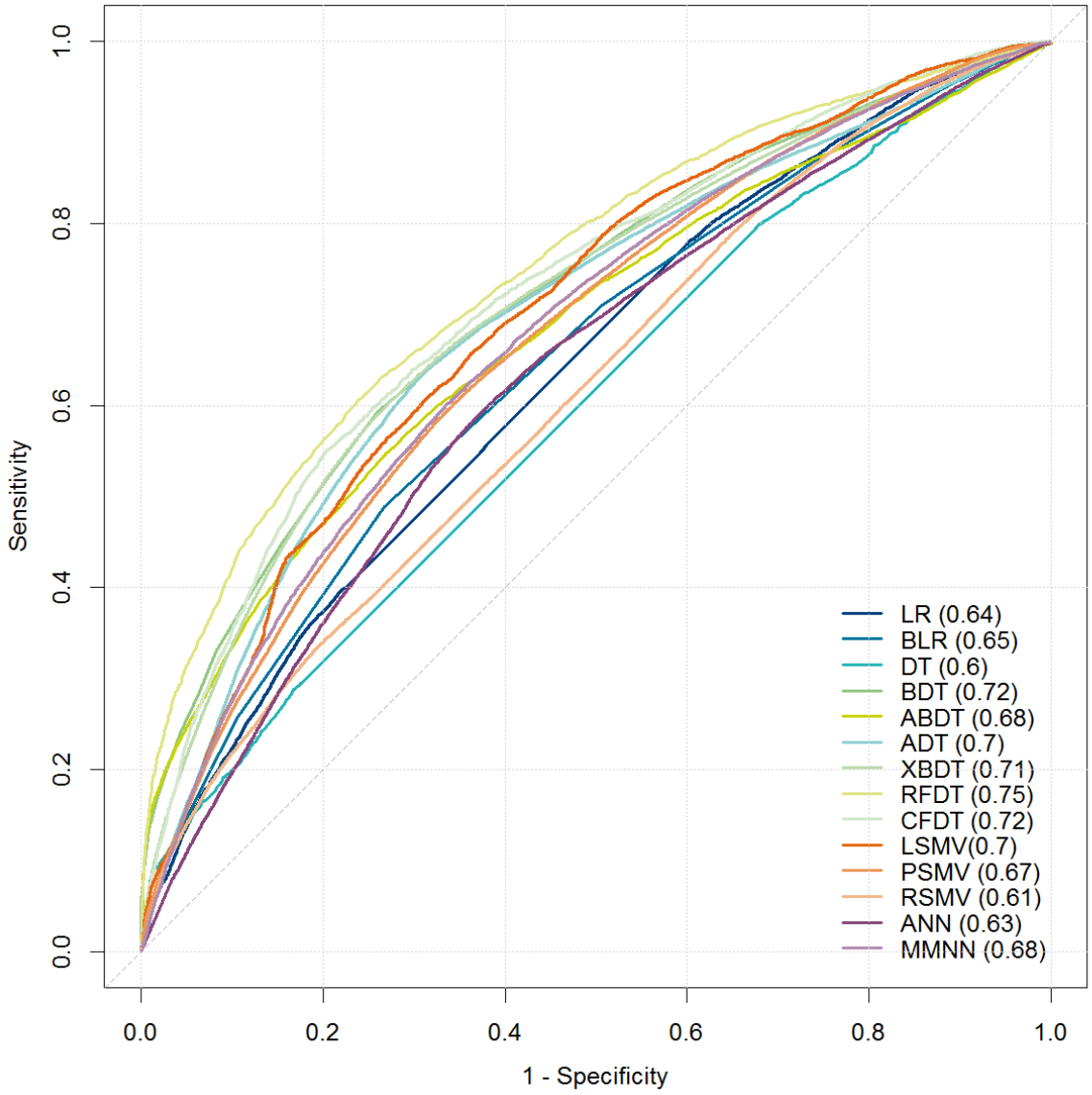

Figure 4.14: ROC curve of selected models, training data (RQ2).
Source: Experimental data, University of Giessen, 2011.

4.6.4.2 Testing of Algorithms

Compared to RQ1a, results of the training stage for RQ2 were slightly better. However, the generally better results are not reflected in the testing stage, which indicates that the split based on interviewers might lead to overfitting. **Figure 4.15** shows that—except for LR and DT that again misclassified all falsifications (see Appendix **Table A 4.18**)—all algorithms resulted in very similar, moderate F1-Scores. The highest F1-Scores is obtained by BLR (F1-Socre of 0.68) and RSMV (F1-Score of 0.64). This is mainly due to the low False-Negative

Rate of 17% and 15% respectively (**Table 4.8**). At the same time, these algorithms showed very high False-Positive Rates (63% and 81%) which would be problematic in a real-world scenario in which we try to identify interviewers for more in-depth controls. Taking this factor into account, algorithms like BDT, RFDT, or CFDT that were able to also reach high Accuracy and a balanced False-Positive and False-Negative rate might still be preferable.

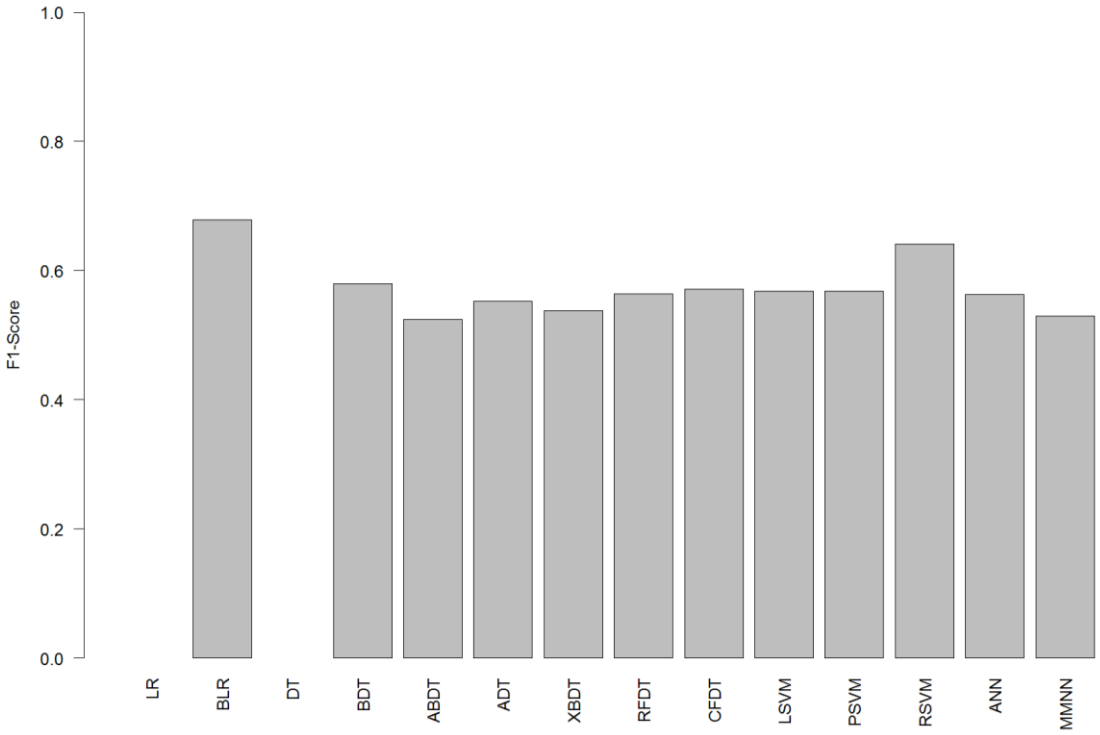

Figure 4.15: F1-Scores of selected models, test data (RQ2).
Source: Experimental data, University of Giessen, 2011.

The results for the ROC/AUC metric again show, that results are very similar for all algorithms (**Figure 4.16**). In terms of AUC, BLR (AUC of 0.61), XDBT (AUC of 0.61), BDT (AUC of 0.60), and ABDT (AUC of 0.60) show the best performance. DT shows the lowest AUC of 0.54. **Table 4.8** shows, that the False-Negative Rate of all algorithms varies between 0.42 and 0.5 while the False-Positive Rate is slightly lower, varying between 0.33 and 0.5 (see Appendix **Table A 4.19** for confusion matrix). In summary, this shows that different falsifiers produce comparable patterns which are detectable by all most algorithms (RQ2), however patterns of the same falsifier are more likely to be recognizable (RQ1). This hints that each falsifier produces own strategies, which overlap in some parts with the strategies of other falsifiers but are also unique in other parts.

Table 4.8: Final performance measures of selected models, test data (RQ2).

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.00	1.00	0.00	1.00	-	0.50	-
BLR	0.63	0.17	0.83	0.37	0.57	0.60	0.68
DT	0.00	1.00	0.00	1.00	-	0.50	-
BDT	0.36	0.45	0.55	0.64	0.61	0.60	0.58
ABDT	0.29	0.54	0.46	0.71	0.61	0.58	0.52
ADT	0.39	0.47	0.53	0.61	0.58	0.57	0.55
XBDT	0.31	0.52	0.48	0.69	0.61	0.58	0.54
RFDT	0.34	0.47	0.53	0.66	0.61	0.59	0.56
CFDT	0.35	0.46	0.54	0.65	0.61	0.59	0.57
LSMV	0.41	0.44	0.56	0.59	0.58	0.57	0.57
PSMV	0.40	0.45	0.55	0.60	0.58	0.58	0.57
RSMV	0.81	0.15	0.85	0.19	0.51	0.52	0.64
ANN	0.39	0.46	0.54	0.61	0.58	0.58	0.56
MMNN	0.39	0.50	0.50	0.61	0.56	0.55	0.53
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	AUC
LR	0.41	0.45	0.55	0.59	0.57	0.57	0.57
BLR	0.35	0.44	0.56	0.65	0.62	0.60	0.61
DT	0.50	0.42	0.58	0.50	0.54	0.54	0.54
BDT	0.36	0.45	0.55	0.64	0.61	0.60	0.60
ABDT	0.34	0.46	0.54	0.66	0.61	0.60	0.60
ADT	0.33	0.48	0.52	0.67	0.61	0.59	0.59
XBDT	0.33	0.46	0.54	0.67	0.62	0.61	0.61
RFDT	0.37	0.46	0.54	0.63	0.60	0.59	0.59
CFDT	0.37	0.45	0.55	0.63	0.60	0.59	0.59
LSMV	0.41	0.44	0.56	0.59	0.58	0.57	0.58
PSMV	0.37	0.45	0.55	0.63	0.60	0.59	0.59
RSMV	0.37	0.46	0.54	0.63	0.59	0.58	0.59
ANN	0.39	0.45	0.55	0.61	0.58	0.58	0.58
MMNN	0.38	0.50	0.50	0.62	0.57	0.56	0.56

Source: Experimental data, University of Giessen, 2011.

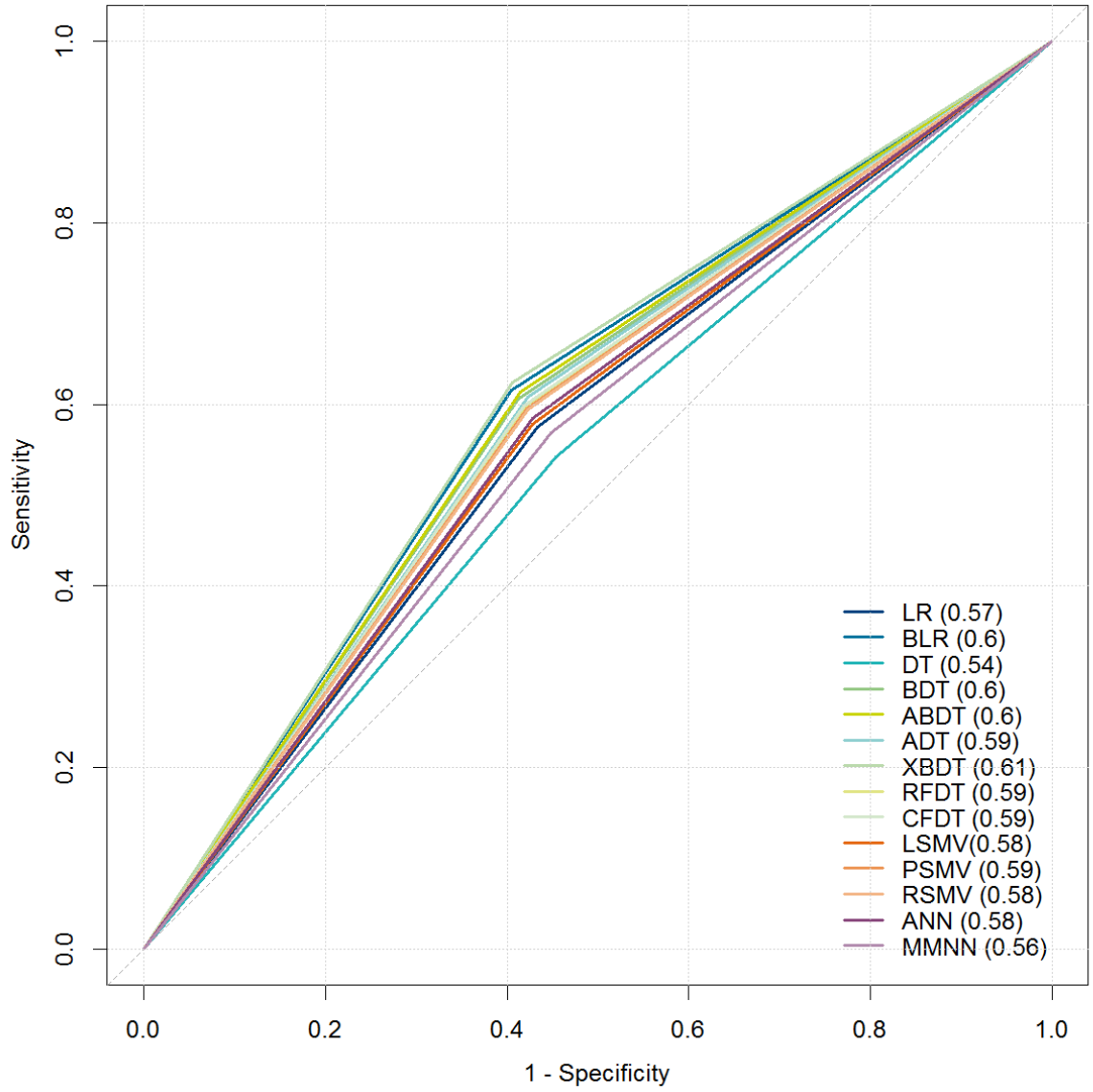

Figure 4.16: ROC curve of selected models, test data (RQ2).
Source: Experimental data, University of Giessen, 2011.

4.6.5 Research Question 3 (RQ3)

4.6.5.1 Training of Algorithms

Using the full experimental dataset in the training stage results in moderate to high F1-Scores for most algorithms. Except for the low performance of LR—due to the misclassification of all falsifications—all algorithms reached an F1-Score of at least 0.58 (see Appendix **Table A 4.13** for more details). The best performing algorithms are again BDT and RFDT with an F1-Score of 1.00 (**Figure 4.17**). In this setting, ADT reached an F1-Score of 1.00 as well. All three algorithms were close to a perfect classification: BDT misclassified one real interview and two falsifications, ADT misclassified one falsification, and RFDT misclassified two falsifications (Appendix **Table A 4.8**). Hence, all three showed a False-Negative and False-Positive Rate of

0%. Very close to this result is ADBT, with an F1-Score of 0.95, with a False-Positive Rate of 1% and a False-Negative Rate of 8%. F1-Scores of the other algorithms range from 0.58 for ANN up to 0.78 for XBDT.

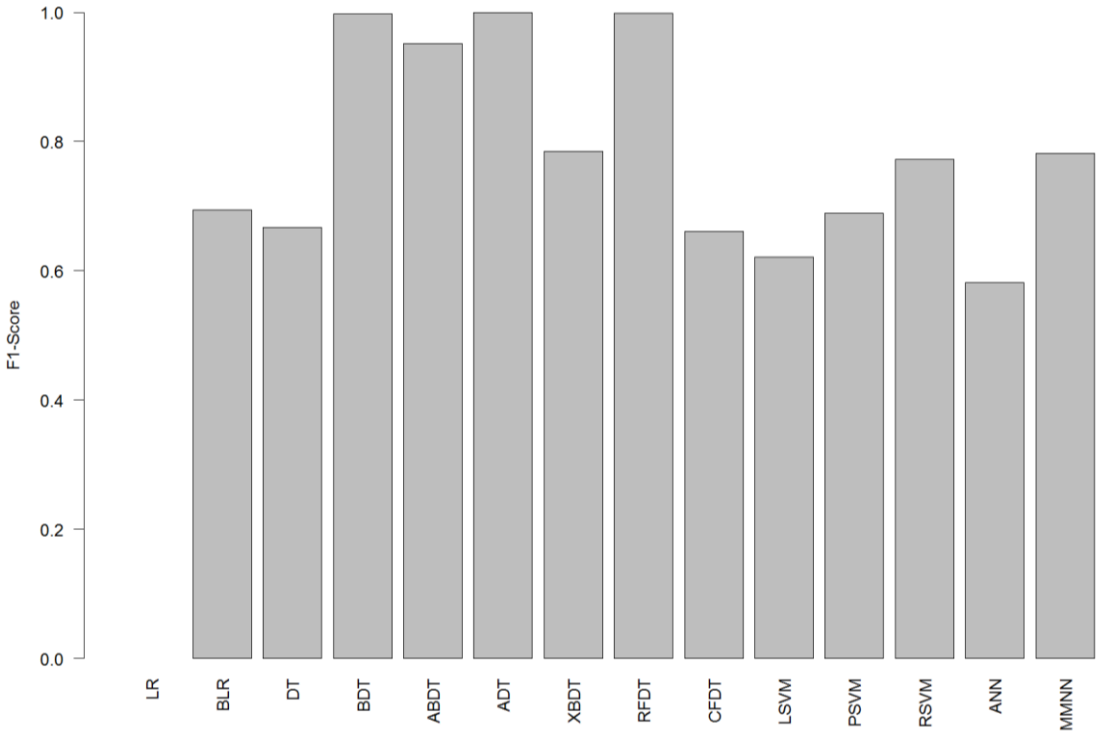

Figure 4.17: F1-Scores of selected models, training data (RQ3).
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Result for the ROC/AUC metric are very consistent with the findings of the F1-Score. RFDT shows the highest AUC of 0.72, followed by BDT with an AUC of 0.79, XBDT and CFDT with an AUC of 0.69 and ADT with an AUC of 0.68 (**Figure 4.18**). Both, BDT and RFDT again have a False-Positive and False-Negative Rate of 0% (Appendix **Table A 4.9**). As in most of the other results, especially the different decision tree-based algorithms showed superior performance in the training stage, compared to the other algorithms. All other groups of algorithms showed False-Positive and False-Negative Rates around 30 to 40%.

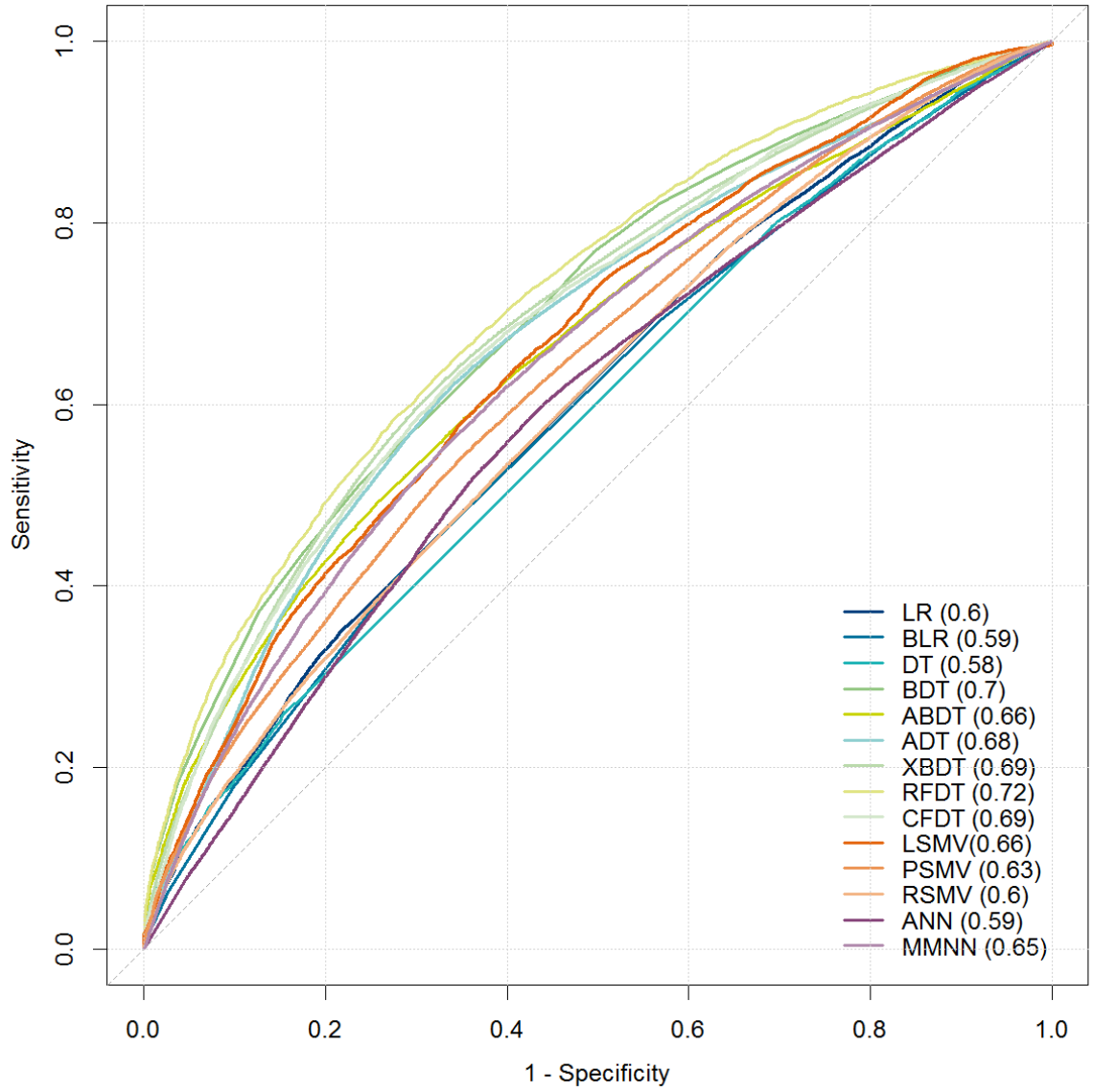

Figure 4.18: ROC curve of selected models, training data (RQ3).
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

4.6.5.2 Testing of Algorithms

In contrast to the other testing results we find that the performance drops drastically. In terms of the F1-Score (**Figure 4.19**), the score only varies between 0.07 and 0.18, which is a very low performance outcome. The highest F1-Scores are observed for MMNN (0.18), BDT (0.17), and PSMV (0.17). Although they show a comparable low False-Negative Rate, and hence identify a relatively high share of correctly classified falsification, simultaneously they show a high False-Positive Rate (**Table 4.9**). This holds for all algorithms. For example, BLR shows a low False-Negative Rate of 21%—hence 277 out of 351 falsifications were correctly classified (see Appendix **Table A 4.20**)—the False-Positive Rate is extremely high with 63%–

–hence 2,823 out of 4,465 real interviews were wrongly classified as falsifications. Such results do not increase the precision of quality controls, as one would still need to analyze roundabout two thirds of the survey.

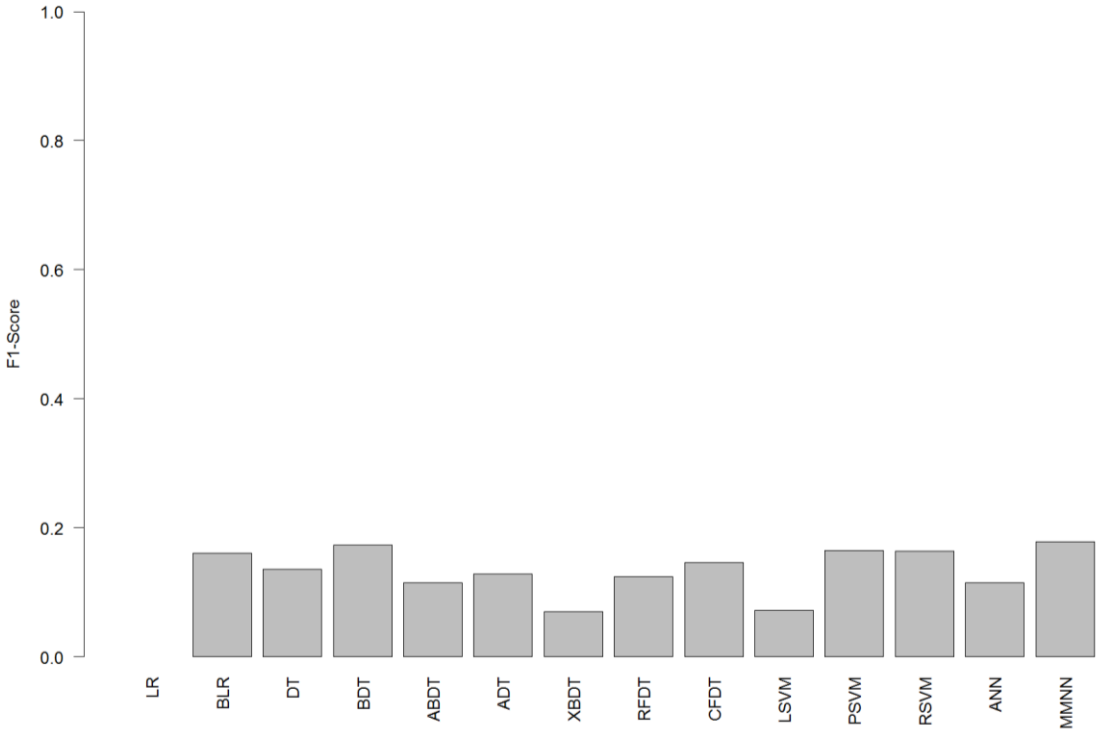

Figure 4.19: F1-Scores of selected models, test data (RQ3).
Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Even worse results are obtained for the ROC/AUC metric (**Figure 4.20**; Appendix **Table A 4.21**). All results lie barley above or even below 0.5 which means that they are not better than a random guess. This is also reflected in the False-Positive and False-Negative Rates (**Table 4.9**). Hence, we have to reject RQ3. We were not able to identify falsifications of other falsifiers in another survey using supervised machine learning. However, we are not able to clearly say if this is due to differences in the falsification behavior and hence the patterns produced, or due to the differences in survey characteristics and/or class balance.

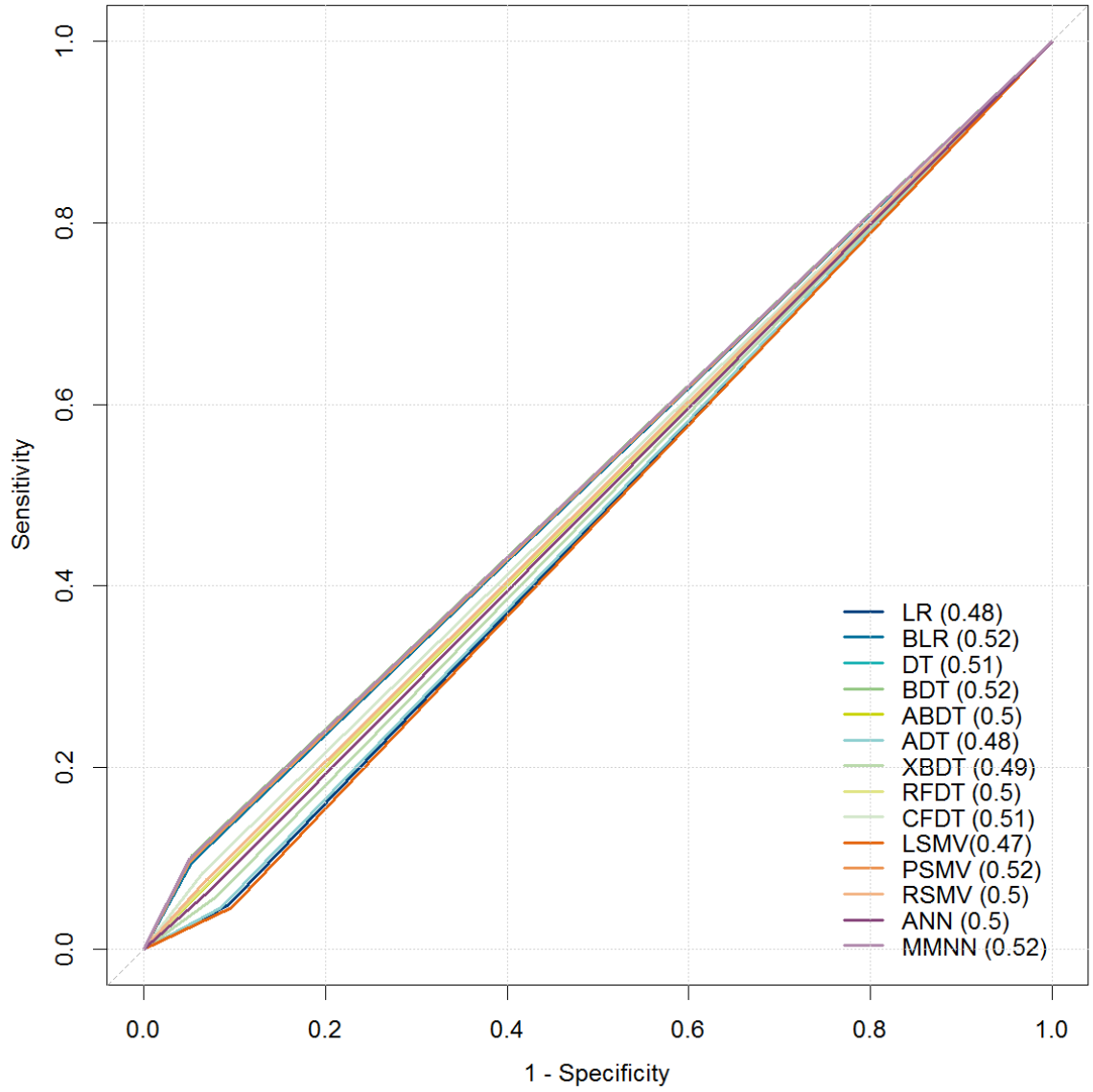

Figure 4.20: ROC curve of selected models, test data (RQ3).

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table 4.9: Final performance measures of selected models, test data (RQ3).

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.00	1.00	0.00	1.00	-	0.93	-
BLR	0.63	0.21	0.79	0.37	0.09	0.40	0.16
DT	1.00	0.00	1.00	0.00	0.07	0.07	0.14
BDT	0.38	0.45	0.55	0.62	0.10	0.62	0.17
ABDT	0.34	0.68	0.32	0.66	0.07	0.64	0.11
ADT	0.41	0.58	0.42	0.59	0.08	0.58	0.13
XBDT	0.35	0.80	0.20	0.65	0.04	0.62	0.07
RFDT	0.34	0.64	0.36	0.66	0.08	0.63	0.12
CFDT	0.37	0.55	0.45	0.63	0.09	0.61	0.15
LSMV	0.48	0.74	0.26	0.52	0.04	0.50	0.07
PSMV	0.48	0.36	0.64	0.52	0.10	0.53	0.17
RSMV	0.44	0.41	0.59	0.56	0.10	0.56	0.16
ANN	0.31	0.70	0.30	0.69	0.07	0.66	0.11
MMNN	0.46	0.33	0.67	0.54	0.10	0.55	0.18
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	AUC
LR	0.48	0.69	0.31	0.52	0.05	0.51	0.48
BLR	0.48	0.37	0.63	0.52	0.09	0.53	0.52
DT	0.40	0.54	0.46	0.60	0.08	0.59	0.51
BDT	0.38	0.45	0.55	0.62	0.10	0.62	0.53
ABDT	0.33	0.66	0.34	0.67	0.07	0.64	0.50
ADT	0.32	0.80	0.20	0.68	0.05	0.64	0.48
XBDT	0.35	0.72	0.28	0.65	0.06	0.62	0.49
RFDT	0.37	0.62	0.38	0.63	0.07	0.61	0.50
CFDT	0.41	0.52	0.48	0.59	0.08	0.58	0.51
LSMV	0.48	0.71	0.29	0.52	0.05	0.50	0.47
PSMV	0.49	0.34	0.66	0.51	0.10	0.52	0.52
RSMV	0.47	0.50	0.50	0.53	0.08	0.52	0.50
ANN	0.44	0.59	0.41	0.56	0.07	0.55	0.50
MMNN	0.46	0.36	0.64	0.54	0.10	0.55	0.52

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

4.7 Discussion

In this study, we provide comprehensive insights into the performance of supervised machine learning algorithms for detecting falsified interviews in survey data. While previous research mainly focused on unsupervised algorithms, such as cluster analysis, we evaluated four different supervised algorithm groups: regression-based models, tree-based models, support vector machines, and neural networks. Thereby, the assessment is based on both experimental and real survey data with identified falsifications, allowing for training and testing of the algorithms in a realistic setting. Furthermore, by simulating three distinct scenarios, we were able to address the question of each algorithms' efficiency when trained on falsifications within the same survey (RQ1), when trained on falsifications induced by disparate falsifiers within the same survey (RQ2), and when trained on falsifications from a different survey (RQ3).

In general, we found positive results for two of our three research questions. Falsifiers produced patterns that could be identified within the same survey, regardless of whether the algorithms were trained on a random split of the data or were specifically trained and tested on data from different falsifiers within the same survey (RQ1 and RQ2). We found that especially the algorithms based on decision trees showed solid outcomes throughout the different research questions: In particular, bagged decision trees (BDT) as well as random forest (RFDT) frequently reached the highest performance. Our finding regarding the suitability of random forest thus corroborate those of Birnbaum (2012) and Birnbaum et al. (2013). In some cases, even very simple techniques like boosted logistic regression obtained solid results. However, inferring from the patterns of falsifications in one survey on interviews at risk in another survey did not prove to be a successful strategy (RQ3). None of the algorithms were able to accurately predict falsifications in the real-world data when trained on the experimental data, resulting in a high number of false-negative and false-positive cases.

4.7.1 Practical Implications of Results

The insights gained in this study also have some significant practical implications. The primary objective of this paper was to evaluate the efficiency of supervised machine learning algorithms, with the aim of determining whether and under which circumstances they can be integrated into quality control routines. In order to support common strategies such as monitoring and re-interviewing, making them more targeted and cost-effective, algorithms need to precisely predict falsifications. In this context, an appropriate algorithm should result in low false-negative rates and a moderate-sized "at risk" group of interviewers, who are selected for further controls. Theoretically, supervised algorithms offer an optimal foundation for this

objective, as they are capable of learning from real-world falsifications. However, our results indicate that this assumption is only confirmed in certain situations, in practice. Training algorithms based on falsifications within the same survey proofed as highly efficient. Therefore, practitioners could utilize falsifications identified in the field to detect similar instances in subsequent phases of the fieldwork. In this context, it is also possible to retrain the algorithms once new falsifications are identified. A comparable situation may arise in the context of panel studies. Practitioners could utilize falsifications identified in one wave of the survey as training data for quality controls of a subsequent wave. Importantly, practitioners should refrain from using data from different surveys, which use a different questionnaire or rely on a different population. Dissimilarities between the dataset make it virtually impossible for the algorithms to make precise predictions regarding the falsification status.

4.7.2 Limitations and Future Work

The empirical setting used in our analysis comes with some limitations, which should be kept in mind when interpreting the findings and for future research. First, the two datasets used address two different populations: Students and refugees and also differ in terms of survey topics, questionnaires, and interview situation. This might lead to dataset shifting for RQ3, i.e., differences between training and test data that reduce the performance of the resulting models. In terms of topic or questions, we might get different results in indicators just based on the questions used to calculate them. As each question might also impact respondent's response behavior differently (Biemer and Lyberg 2003), the question alone could lead to differences. Furthermore, while the interviewers in the experimental setting were students themselves, interviewers in the real dataset were professionals. Therefore, also the same data-based features could be used for both samples, it might be natural to assume that deviations of falsifiers might have a different impact on these features depending on the falsifiers and, possibly, also on the underlying population. Further the interview situation was very different. Experimental data were collected in a controlled environment; hence, each falsification is also documented as such. For the real-world data, on the other hand, some interviews labeled as "real" could unknowingly still be a falsification. Further, behavior of student interviewers and professional interviewers might also be different or driven by the population their working with. Future research should therefore replicate the findings using further real-world or experimental data focused on different populations, e.g., including business surveys. In addition, it might be valuable to include information on the interviewers such as experience, number of interviews conducted in the survey, time needed for the interviews in the supervised learning procedure.

Second, when it comes to transfer learned patterns from one survey to another, (dis)-similarity between the surveys with regard to population and interviewers' behavior (e.g., share of falsifications) might play a relevant role. Thus, it appears more promising to train models based on one wave of a panel survey to identify fraud in the next wave as compared to the transfer between two quite dissimilar surveys in the present analysis. As this option exists only for panel studies, for other surveys, it might be sensible to select a training dataset as similar as possible to the survey under consideration. In our application, a major difference between the two datasets consisted in the different share of falsifiers, which was 50% in the experimental setting, while it was just around 7% in der real dataset. Given that some supervised learning methods exhibit difficulties when trained on unbalanced data, the failure to confirm RQ3 might be a result of such differences. Therefore, future research might replicate the analysis making use of artificial datasets generated from the experimental data by means of bootstrapping (as in Storfinger and Winker 2013) with different shares of falsifiers to analyze the respective impact.

Lastly, we only included a limited number of algorithms. New or other algorithms might lead to better results making the usage of supervised machine learning for this use case more precise and applicable. Besides considering further algorithms, it might also be a promising venue for future research considering ensemble methods, i.e., combinations of the results of our different algorithms to increase the performance.

Appendix

Table A 4.1: Overview of tuning parameters for each algorithm.

Algorithm	Method	Model Parameters	Model Values
LR	'glmnet'	alpha	0, 0.5, 1
		lambda	0.0001, 0.001, 0.01, 0.1, 1, 3, 5, 7, 9
BLR	'LogitBoost'	nIter	5, 10, 15, 20, 25, 30
DT	'rpart'	cp	0.0001, 0.001, 0.01, 0.1, 1, 3, 5, 9
BDT	'treebag'	-	
ABDT	'AdaBag'	mfinal	50, 100, 150, 200, 250
		maxdepth	1, 6, 11
ADT	'ada'	iter	50, 100, 150, 200, 250
		maxdepth	1, 6, 11
		nu	0.01, 0.1, 0.5
		nrounds	100, 250
		max_depth	1, 6, 11
		eta	0.01, 0.1
XBDT	'xgbDART'	gamma	0.1, 1
		subsample	0.5, 1.0
		colsample_bytree	0.2, 0.6, 1.0
		rate_drop	0.01, 0.1, 0.5
		skip_drop	0.5, 1.0
RFDT	'rf'	mtry	1, 3, 5, 7, 9, 11
CFDT	'cforest'	mtry	1, 3, 5, 7, 9, 11
LSMV	'svmLinear'	C	0.1, 0.18, 0.32, 0.56, 1. 1.78, 3.16, 5.62, 10, 17.78, 31.62
PSMV	'svmPoly'	degree	1, 2, 3, 4
		scale	0.1, 1, 10
		C	0.1, 0.18, 0.32, 0.56, 1. 1.78, 3.16, 5.62, 10, 17.78, 31.62
RSMV	'svmRadial'	sigma	0.1, 1, 10
		C	0.1, 0.18, 0.32, 0.56, 1. 1.78, 3.16, 5.62, 10, 17.78, 31.62
ANN	'nnet'	size	1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51
		decay	0.0001, 0.001, 0.01, 0.1
MMNN	'monmlp'	hidden1	5, 10, 15, 20, 25
		n.ensemble	2, 5, 7, 10

Table A 4.2: Confusion matrix according to F1-Score (RQ1a); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	113	34	147
	Interview	445	544	989
	Total	558	578	1136
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	406	301	707
	Interview	152	277	429
	Total	558	578	1136
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	558	578	1136
	Total	558	578	1136
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	558	0	558
	Interview	0	578	578
	Total	558	578	1136
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	312	60	372
	Interview	246	518	764
	Total	558	578	1136
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	558	0	558
	Interview	0	578	578
	Total	558	578	1136
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	417	80	497
	Interview	141	498	639
	Total	558	578	1136
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	557	0	557
	Interview	1	578	579
	Total	558	578	1136
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	293	104	397
	Interview	265	474	739
	Total	558	578	1136

Source: Experimental data, University of Giessen, 2011.

Table A 4.2 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	344	217	561
	Interview	214	361	575
	Total	558	578	1136
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	558	578	1136
	Total	558	578	1136
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	578	578
	Interview	558	0	558
	Total	558	578	1136
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	296	175	471
	Interview	262	403	665
	Total	558	578	1136
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	492	48	540
	Interview	66	530	596
	Total	558	578	1136

Source: Experimental data, University of Giessen, 2011.

Table A 4.3: Confusion matrix according to ROC (RQ1a); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	332	202	534
	Interview	226	376	602
	Total	558	578	1136
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	406	301	707
	Interview	152	277	429
	Total	558	578	1136
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	364	154	518
	Interview	194	424	618
	Total	558	578	1136
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	558	0	558
	Interview	0	578	578
	Total	558	578	1136
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	532	0	532
	Interview	26	578	604
	Total	558	578	1136
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	343	139	482
	Interview	215	439	654
	Total	558	578	1136
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	376	145	521
	Interview	182	433	615
	Total	558	578	1136
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	557	0	557
	Interview	1	578	579
	Total	558	578	1136
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	456	84	540
	Interview	102	494	596
	Total	558	578	1136

Source: Experimental data, University of Giessen, 2011.

Table A 4.3 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	344	215	559
	Interview	214	363	577
	Total	558	578	1136
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	369	167	536
	Interview	189	411	600
	Total	558	578	1136
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	402	148	550
	Interview	156	430	586
	Total	558	578	1136
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	297	177	474
	Interview	261	401	662
	Total	558	578	1136
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	416	98	514
	Interview	142	480	622
	Total	558	578	1136

Source: Experimental data, University of Giessen, 2011.

Table A 4.4: Confusion matrix according to F1-Score (RQ1b); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	116	44	160
	Interview	164	3529	3693
	Total	280	3573	3853
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	225	99	324
	Interview	55	3474	3529
	Total	280	3573	3853
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	191	15	206
	Interview	89	3558	3647
	Total	280	3573	3853
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	278	0	278
	Interview	2	3573	3575
	Total	280	3573	3853
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	222	6	228
	Interview	58	3567	3625
	Total	280	3573	3853
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	190	23	213
	Interview	90	3550	3640
	Total	280	3573	3853
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	266	1	267
	Interview	14	3572	3586
	Total	280	3573	3853
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	280	0	280
	Interview	0	3573	3573
	Total	280	3573	3853
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	206	5	211
	Interview	74	3568	3642
	Total	280	3573	3853

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.4 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	111	32	143
	Interview	169	3541	3710
	Total	280	3573	3853
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	214	6	220
	Interview	66	3567	3633
	Total	280	3573	3853
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	240	11	251
	Interview	40	3562	3602
	Total	280	3573	3853
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	243	15	258
	Interview	37	3558	3595
	Total	280	3573	3853
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	225	15	240
	Interview	55	2558	2613
	Total	280	2573	2853

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.5: Confusion matrix according to ROC (RQ1b); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	116	44	160
	Interview	164	3529	3693
	Total	280	3573	3853
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	225	99	324
	Interview	55	2474	2529
	Total	280	2573	2853
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	226	29	255
	Interview	54	3544	3598
	Total	280	3573	3853
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	278	0	278
	Interview	2	3573	3575
	Total	280	3573	3853
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	251	0	251
	Interview	29	3573	3602
	Total	280	3573	3853
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	273	0	273
	Interview	7	3573	3580
	Total	280	3573	3853
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	272	1	273
	Interview	8	3572	3580
	Total	280	3573	3853
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	258	0	258
	Interview	22	3573	3595
	Total	280	3573	3853
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	197	4	201
	Interview	83	3569	3652
	Total	280	3573	3853

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.5 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	118	45	163
	Interview	162	3528	3690
	Total	280	3573	3853
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	204	12	216
	Interview	76	3561	3637
	Total	280	3573	3853
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	242	7	249
	Interview	38	3566	3604
	Total	280	3573	3853
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	248	7	255
	Interview	32	3566	3598
	Total	280	3573	3853
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	213	22	235
	Interview	67	3551	3618
	Total	280	3573	3853

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.6: Confusion matrix according to F1-Score (RQ2); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	355	357	712
	Total	355	357	712
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	312	205	517
	Interview	43	152	195
	Total	355	357	712
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	355	357	712
	Total	355	357	712
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	355	0	355
	Interview	0	357	357
	Total	355	357	712
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	251	29	280
	Interview	104	328	432
	Total	355	357	712
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	355	0	355
	Interview	0	357	357
	Total	355	357	712
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	225	73	298
	Interview	130	284	414
	Total	355	357	712
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	354	0	354
	Interview	1	357	358
	Total	355	357	712
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	260	55	315
	Interview	95	302	397
	Total	355	357	712

Source: Experimental data, University of Giessen, 2011.

Table A 4.6 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	239	118	357
	Interview	116	239	355
	Total	355	357	712
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	250	98	348
	Interview	105	259	364
	Total	355	357	712
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	352	0	352
	Interview	3	357	360
	Total	355	357	712
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	232	110	342
	Interview	123	247	370
	Total	355	357	712
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	355	0	355
	Interview	0	357	357
	Total	355	357	712

Source: Experimental data, University of Giessen, 2011.

Table A 4.7: Confusion matrix according to ROC (RQ2); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	240	121	361
	Interview	115	236	351
	Total	355	357	712
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	205	93	298
	Interview	150	264	414
	Total	355	357	712
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	231	126	357
	Interview	124	231	355
	Total	355	357	712
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	355	0	355
	Interview	0	357	357
	Total	355	357	712
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	343	0	343
	Interview	12	357	369
	Total	355	357	712
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	254	83	337
	Interview	101	274	375
	Total	355	357	712
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	252	78	330
	Interview	103	279	382
	Total	355	357	712
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	355	0	355
	Interview	0	357	357
	Total	355	357	712
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	287	49	336
	Interview	68	308	376
	Total	355	357	712

Source: Experimental data, University of Giessen, 2011.

Table A 4.7 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	239	118	357
	Interview	116	239	355
	Total	355	357	712
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	255	90	345
	Interview	100	267	367
	Total	355	357	712
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	271	82	353
	Interview	84	275	359
	Total	355	357	712
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	237	113	350
	Interview	118	244	362
	Total	355	357	712
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	341	13	354
	Interview	14	344	358
	Total	355	357	712

Source: Experimental data, University of Giessen, 2011.

Table A 4.8: Confusion matrix according to F1-Score (RQ3); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	710	710	1420
	Total	710	710	1420
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	617	451	1068
	Interview	93	259	352
	Total	710	710	1420
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	710	710	1420
	Interview	0	0	0
	Total	710	710	1420
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	708	1	709
	Interview	2	709	711
	Total	710	710	1420
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	651	7	658
	Interview	59	703	762
	Total	710	710	1420
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	709	0	709
	Interview	1	710	711
	Total	710	710	1420
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	542	129	671
	Interview	168	581	749
	Total	710	710	1420
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	708	0	708
	Interview	2	710	712
	Total	710	710	1420

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.8 (continued)

CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	432	166	598
	Interview	278	544	822
	Total	710	710	1420
LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	440	267	707
	Interview	270	443	713
	Total	710	710	1420
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	485	212	697
	Interview	225	498	723
	Total	710	710	1420
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	540	148	688
	Interview	170	562	732
	Total	710	710	1420
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	346	133	479
	Interview	364	577	941
	Total	710	710	1420
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	549	145	694
	Interview	161	565	726
	Total	710	710	1420

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.9: Confusion matrix according to ROC (RQ3); training data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	443	273	716
	Interview	267	437	704
	Total	710	710	1420
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	526	377	903
	Interview	184	333	517
	Total	710	710	1420
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	455	204	659
	Interview	255	506	761
	Total	710	710	1420
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	708	1	709
	Interview	2	709	711
	Total	710	710	1420
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	660	7	667
	Interview	50	703	753
	Total	710	710	1420
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	540	104	644
	Interview	170	606	776
	Total	710	710	1420
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	634	48	682
	Interview	76	662	738
	Total	710	710	1420
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	710	0	710
	Interview	0	710	710
	Total	710	710	1420

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.9 (continued)

CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	558	115	673
	Interview	152	595	747
	Total	710	710	1420
LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	441	263	704
	Interview	269	447	716
	Total	710	710	1420
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	473	233	706
	Interview	237	477	714
	Total	710	710	1420
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	514	218	732
	Interview	196	492	688
	Total	710	710	1420
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	422	241	663
	Interview	288	469	757
	Total	710	710	1420
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	563	152	715
	Interview	147	558	705
	Total	710	710	1420

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.10: Final performance measures of different algorithms (RQ1a); training data.

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.06	0.80	0.20	0.94	0.77	0.58	0.32
BLR	0.52	0.27	0.73	0.48	0.57	0.60	0.64
DT	0.00	1.00	0.00	1.00	-	0.51	-
BDT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
ADBT	0.10	0.44	0.56	0.90	0.84	0.73	0.67
ADT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
XBDT	0.14	0.25	0.75	0.86	0.84	0.81	0.79
RFDT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
CFDT	0.18	0.47	0.53	0.82	0.74	0.68	0.61
LSMV	0.38	0.38	0.62	0.62	0.61	0.62	0.61
PSMV	0.00	1.00	0.00	1.00	-	0.51	-
RSMV	1.00	1.00	0.00	0.00	0.00	0.00	-
ANN	0.30	0.47	0.53	0.70	0.63	0.62	0.58
MMNN	0.08	0.12	0.88	0.92	0.91	0.90	0.90
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	*AUC*
LR	0.35	0.41	0.59	0.65	0.62	0.62	0.60
BLR	0.52	0.27	0.73	0.48	0.57	0.60	0.60
DT	0.27	0.35	0.65	0.73	0.70	0.69	0.58
BDT	0.00	0.00	1.00	1.00	1.00	1.00	0.68
ADBT	0.00	0.05	0.95	1.00	1.00	0.98	0.64
ADT	0.24	0.39	0.61	0.76	0.71	0.69	0.66
XBDT	0.25	0.33	0.67	0.75	0.72	0.71	0.68
RFDT	0.00	0.00	1.00	1.00	1.00	1.00	0.70
CFDT	0.15	0.18	0.82	0.85	0.84	0.84	0.68
LSMV	0.37	0.38	0.62	0.63	0.62	0.62	0.65
PSMV	0.29	0.34	0.66	0.71	0.69	0.69	0.62
RSMV	0.26	0.28	0.72	0.74	0.73	0.73	0.59
ANN	0.31	0.47	0.53	0.69	0.63	0.61	0.58
MMNN	0.17	0.25	0.75	0.83	0.81	0.79	0.63

Source: Experimental data, University of Giessen, 2011.

Table A 4.11: Final performance measures of different algorithms (RQ1b); training data.

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.01	0.59	0.41	0.99	0.73	0.95	0.53
BLR	0.03	0.20	0.80	0.97	0.69	0.96	0.75
DT	0.00	0.32	0.68	1.00	0.93	0.97	0.79
BDT	0.00	0.01	0.99	1.00	1.00	1.00	1.00
ADBT	0.00	0.21	0.79	1.00	0.97	0.98	0.87
ADT	0.01	0.32	0.68	0.99	0.89	0.97	0.77
XBDT	0.00	0.05	0.95	1.00	1.00	1.00	0.97
RFDT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
CFDT	0.00	0.26	0.74	1.00	0.98	0.98	0.84
LSMV	0.01	0.60	0.40	0.99	0.78	0.95	0.52
PSMV	0.00	0.24	0.76	1.00	0.97	0.98	0.86
RSMV	0.00	0.14	0.86	1.00	0.96	0.99	0.90
ANN	0.00	0.13	0.87	1.00	0.94	0.99	0.90
MMNN	0.01	0.20	0.80	0.99	0.94	0.98	0.87
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	AUC
LR	0.01	0.59	0.41	0.99	0.73	0.95	0.78
BLR	0.04	0.20	0.80	0.96	0.69	0.95	0.92
DT	0.01	0.19	0.81	0.99	0.89	0.98	0.77
BDT	0.00	0.01	0.99	1.00	1.00	1.00	0.93
ADBT	0.00	0.10	0.90	1.00	1.00	0.99	0.77
ADT	0.00	0.03	0.98	1.00	1.00	1.00	0.90
XBDT	0.00	0.03	0.97	1.00	1.00	1.00	0.91
RFDT	0.00	0.08	0.92	1.00	1.00	0.99	0.95
CFDT	0.00	0.30	0.70	1.00	0.98	0.98	0.95
LSMV	0.01	0.58	0.42	0.99	0.72	0.95	0.91
PSMV	0.00	0.27	0.73	1.00	0.94	0.98	0.91
RSMV	0.00	0.14	0.86	1.00	0.97	0.99	0.88
ANN	0.00	0.11	0.89	1.00	0.97	0.99	0.92
MMNN	0.01	0.24	0.76	0.99	0.91	0.98	0.94

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.12: Final performance measures of different algorithms (RQ2); training data.

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.00	1.00	0.00	1.00	-	0.50	-
BLR	0.57	0.12	0.88	0.43	0.60	0.65	0.72
DT	0.00	1.00	0.00	1.00	-	0.50	-
BDT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
ADBT	0.08	0.29	0.71	0.92	0.90	0.81	0.79
ADT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
XBDT	0.20	0.37	0.63	0.80	0.76	0.71	0.69
RFDT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
CFDT	0.15	0.27	0.73	0.85	0.83	0.79	0.78
LSMV	0.33	0.33	0.67	0.67	0.67	0.67	0.67
PSMV	0.27	0.30	0.70	0.73	0.72	0.71	0.71
RSMV	0.00	0.01	0.99	1.00	1.00	1.00	1.00
ANN	0.31	0.35	0.65	0.69	0.68	0.67	0.67
MMNN	0.00	0.00	1.00	1.00	1.00	1.00	1.00
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	*AUC*
LR	0.34	0.32	0.68	0.66	0.66	0.67	0.64
BLR	0.26	0.42	0.58	0.74	0.69	0.66	0.65
DT	0.35	0.35	0.65	0.65	0.65	0.65	0.60
BDT	0.00	0.00	1.00	1.00	1.00	1.00	0.72
ADBT	0.00	0.03	0.97	1.00	1.00	0.98	0.68
ADT	0.23	0.28	0.72	0.77	0.75	0.74	0.70
XBDT	0.22	0.29	0.71	0.78	0.76	0.75	0.71
RFDT	0.00	0.00	1.00	1.00	1.00	1.00	0.75
CFDT	0.14	0.19	0.81	0.86	0.85	0.84	0.72
LSMV	0.33	0.33	0.67	0.67	0.67	0.67	0.70
PSMV	0.25	0.28	0.72	0.75	0.74	0.73	0.67
RSMV	0.23	0.24	0.76	0.77	0.77	0.77	0.61
ANN	0.32	0.33	0.67	0.68	0.68	0.68	0.63
MMNN	0.04	0.04	0.96	0.96	0.96	0.96	0.68

Source: Experimental data, University of Giessen, 2011.

Table A 4.13: Final performance measures of different algorithms (RQ3); training data.

Best model according to F1							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	$F1_{score}$
LR	0.00	1.00	0.00	1.00	-	0.50	-
BLR	0.64	0.13	0.87	0.36	0.58	0.62	0.69
DT	1.00	0.00	1.00	0.00	0.50	0.50	0.67
BDT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
ADBT	0.01	0.08	0.92	0.99	0.99	0.95	0.95
ADT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
XBDT	0.18	0.24	0.76	0.82	0.81	0.79	0.78
RFDT	0.00	0.00	1.00	1.00	1.00	1.00	1.00
CFDT	0.23	0.39	0.61	0.77	0.72	0.69	0.66
LSMV	0.38	0.38	0.62	0.62	0.62	0.62	0.62
PSMV	0.30	0.32	0.68	0.70	0.70	0.69	0.69
RSMV	0.21	0.24	0.76	0.79	0.78	0.78	0.77
ANN	0.19	0.51	0.49	0.81	0.72	0.65	0.58
MMNN	0.20	0.23	0.77	0.80	0.79	0.78	0.78
Best model according to ROC							
	FP_{rate}	FN_{rate}	Q_{sens}	Q_{spec}	Q_{prec}	Q_{acc}	*AUC*
LR	0.38	0.38	0.62	0.62	0.62	0.62	0.60
BLR	0.53	0.26	0.74	0.47	0.58	0.60	0.59
DT	0.29	0.36	0.64	0.71	0.69	0.68	0.59
BDT	0.00	0.00	1.00	1.00	1.00	1.00	0.70
ADBT	0.01	0.07	0.93	0.99	0.99	0.96	0.66
ADT	0.15	0.24	0.76	0.85	0.84	0.81	0.68
XBDT	0.07	0.11	0.89	0.93	0.93	0.91	0.69
RFDT	0.00	0.00	1.00	1.00	1.00	1.00	0.72
CFDT	0.16	0.21	0.79	0.84	0.83	0.81	0.69
LSMV	0.37	0.38	0.62	0.63	0.63	0.63	0.66
PSMV	0.33	0.33	0.67	0.67	0.67	0.67	0.63
RSMV	0.31	0.28	0.72	0.69	0.70	0.71	0.60
ANN	0.34	0.41	0.59	0.66	0.64	0.63	0.59
MMNN	0.21	0.21	0.79	0.79	0.79	0.79	0.65

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.14: Confusion matrix according to F1-Score (RQ1a); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	23	9	32
	Interview	129	123	252
	Total	152	132	284
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	120	76	196
	Interview	32	56	88
	Total	152	132	284
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	152	132	284
	Total	152	132	284
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	109	44	153
	Interview	43	88	131
	Total	152	132	284
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	73	26	99
	Interview	79	106	185
	Total	152	132	284
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	95	48	143
	Interview	57	84	141
	Total	152	132	284
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	85	41	126
	Interview	67	91	158
	Total	152	132	284
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	98	31	129
	Interview	54	101	155
	Total	152	132	284
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	79	30	109
	Interview	73	102	175
	Total	152	132	284

Source: Experimental data, University of Giessen, 2011.

Table A 4.14 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	96	50	146
	Interview	56	82	138
	Total	152	132	284
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	152	132	284
	Total	152	132	284
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	152	132	284
	Total	152	132	284
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	90	33	123
	Interview	62	99	161
	Total	152	132	284
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	95	50	145
	Interview	57	82	139
	Total	152	132	284

Source: Experimental data, University of Giessen, 2011.

Table A 4.15: Confusion matrix according to ROC (RQ1a); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	96	45	141
	Interview	56	87	143
	Total	152	132	284
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	120	76	196
	Interview	32	56	88
	Total	152	132	284
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	92	44	136
	Interview	60	88	148
	Total	152	132	284
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	109	44	153
	Interview	43	88	131
	Total	152	132	284
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	94	33	127
	Interview	58	99	157
	Total	152	132	284
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	93	34	127
	Interview	59	98	157
	Total	152	132	284
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	99	40	139
	Interview	53	92	145
	Total	152	132	284
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	98	31	129
	Interview	54	101	155
	Total	152	132	284
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	97	36	133
	Interview	55	96	151
	Total	152	132	284

Source: Experimental data, University of Giessen, 2011.

Table A 4.15 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	97	51	148
	Interview	55	81	136
	Total	152	132	284
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	95	39	134
	Interview	57	93	150
	Total	152	132	284
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	102	45	147
	Interview	50	87	137
	Total	152	132	284
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	90	34	124
	Interview	62	98	160
	Total	152	132	284
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	89	41	130
	Interview	63	91	154
	Total	152	132	284

Source: Experimental data, University of Giessen, 2011.

Table A 4.16: Confusion matrix according to F1-Score (RQ1b); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	27	12	39
	Interview	44	880	924
	Total	71	892	963
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	56	27	83
	Interview	15	865	880
	Total	71	892	963
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	51	8	59
	Interview	20	884	904
	Total	71	892	963
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	54	7	61
	Interview	17	885	902
	Total	71	892	963
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	52	6	58
	Interview	19	886	905
	Total	71	892	963
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	49	8	57
	Interview	22	884	906
	Total	71	892	963
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	56	6	62
	Interview	15	886	901
	Total	71	892	963
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	54	5	59
	Interview	17	887	904
	Total	71	892	963
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	43	3	46
	Interview	28	889	917
	Total	71	892	963

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.16 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	28	10	38
	Interview	43	882	925
	Total	71	892	963
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	50	9	59
	Interview	21	883	904
	Total	71	892	963
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	52	14	66
	Interview	19	878	897
	Total	71	892	963
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	53	15	68
	Interview	18	877	895
	Total	71	892	963
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	54	9	63
	Interview	17	883	900
	Total	71	892	963

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.17: Confusion matrix according to ROC (RQ1b); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	27	13	40
	Interview	44	879	923
	Total	71	892	963
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	56	27	83
	Interview	15	865	880
	Total	71	892	963
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	53	14	67
	Interview	18	878	896
	Total	71	892	963
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	54	7	61
	Interview	17	885	902
	Total	71	892	963
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	54	6	60
	Interview	17	886	903
	Total	71	892	963
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	57	7	64
	Interview	14	885	899
	Total	71	892	963
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	54	6	60
	Interview	17	886	903
	Total	71	892	963
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	38	0	38
	Interview	33	892	925
	Total	71	892	963
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	43	2	45
	Interview	28	890	918
	Total	71	892	963

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.17 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	30	13	43
	Interview	41	879	920
	Total	71	892	963
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	51	5	56
	Interview	20	887	907
	Total	71	892	963
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	54	13	67
	Interview	17	879	896
	Total	71	892	963
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	53	14	67
	Interview	18	878	896
	Total	71	892	963
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	53	9	62
	Interview	18	883	901
	Total	71	892	963

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33).

Table A 4.18: Confusion matrix according to F1-Score (RQ2); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	355	353	708
	Total	355	353	708
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	296	221	517
	Interview	59	132	191
	Total	355	353	708
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	355	353	708
	Total	355	353	708
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	197	128	325
	Interview	158	225	383
	Total	355	353	708
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	163	103	266
	Interview	192	250	442
	Total	355	353	708
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	188	137	325
	Interview	167	216	383
	Total	355	353	708
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	171	110	281
	Interview	184	243	427
	Total	355	353	708
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	187	121	308
	Interview	168	232	400
	Total	355	353	708
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	191	123	314
	Interview	164	230	394
	Total	355	353	708

Source: Experimental data, University of Giessen, 2011.

Table A 4.18 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	199	146	345
	Interview	156	207	363
	Total	355	353	708
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	197	141	338
	Interview	158	212	370
	Total	355	353	708
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	302	285	587
	Interview	53	68	121
	Total	355	353	708
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	193	138	331
	Interview	162	215	377
	Total	355	353	708
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	178	139	317
	Interview	177	214	391
	Total	355	353	708

Source: Experimental data, University of Giessen, 2011.

Table A 4.19: Confusion matrix according to ROC (RQ2); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	196	145	341
	Interview	159	208	367
	Total	355	353	708
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	200	125	325
	Interview	155	228	383
	Total	355	353	708
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	207	175	382
	Interview	148	178	326
	Total	355	353	708
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	197	128	325
	Interview	158	225	383
	Total	355	353	708
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	190	120	310
	Interview	165	233	398
	Total	355	353	708
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	183	118	301
	Interview	172	235	407
	Total	355	353	708
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	193	116	309
	Interview	162	237	399
	Total	355	353	708
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	193	130	323
	Interview	162	223	385
	Total	355	353	708
CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	195	131	326
	Interview	160	222	382
	Total	355	353	708

Source: Experimental data, University of Giessen, 2011.

Table A 4.19 (continued)

LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	199	145	344
	Interview	156	208	364
	Total	355	353	708
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	194	132	326
	Interview	161	221	382
	Total	355	353	708
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	193	132	325
	Interview	162	221	383
	Total	355	353	708
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	194	138	332
	Interview	161	215	376
	Total	355	353	708
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	178	135	313
	Interview	177	218	395
	Total	355	353	708

Source: Experimental data, University of Giessen, 2011.

Table A 4.20: Confusion matrix according to F1-Score (RQ3); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	0	0	0
	Interview	351	4465	4816
	Total	351	4465	4816
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	277	2823	3100
	Interview	74	1642	1716
	Total	351	4465	4816
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	351	4465	4816
	Interview	0	0	0
	Total	351	4465	4816
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	193	1677	1870
	Interview	158	2788	2946
	Total	351	4465	4816
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	113	1505	1618
	Interview	238	2960	3198
	Total	351	4465	4816
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	149	1814	1963
	Interview	202	2651	2853
	Total	351	4465	4816
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	70	1566	1636
	Interview	281	2899	3180
	Total	351	4465	4816
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	125	1536	1661
	Interview	226	2929	3155
	Total	351	4465	4816

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.20 (continued)

CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	159	1663	1822
	Interview	192	2802	2994
	Total	351	4465	4816
LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	93	2133	2226
	Interview	258	2332	2590
	Total	351	4465	4816
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	225	2143	2368
	Interview	126	2322	2448
	Total	351	4465	4816
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	206	1956	2162
	Interview	145	2509	2654
	Total	351	4465	4816
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	106	1391	1497
	Interview	245	3074	3319
	Total	351	4465	4816
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	235	2038	2273
	Interview	116	2427	2543
	Total	351	4465	4816

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.21: Confusion matrix according to ROC (RQ3); test data.

LR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	110	2139	2249
	Interview	241	2326	2567
	Total	351	4465	4816
BLR		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	221	2123	2344
	Interview	130	2342	2472
	Total	351	4465	4816
DT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	163	1767	1930
	Interview	188	2698	2886
	Total	351	4465	4816
BDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	193	1677	1870
	Interview	158	2788	2946
	Total	351	4465	4816
ABDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	118	1492	1610
	Interview	233	2973	3206
	Total	351	4465	4816
ADT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	70	1450	1520
	Interview	281	3015	3296
	Total	351	4465	4816
XBDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	98	1576	1674
	Interview	253	2889	3142
	Total	351	4465	4816
RFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	134	1670	1804
	Interview	217	2795	3012
	Total	351	4465	4816

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

Table A 4.21 (continued)

CFDT		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	167	1820	1987
	Interview	184	2645	2829
	Total	351	4465	4816
LSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	102	2139	2241
	Interview	249	2326	2575
	Total	351	4465	4816
PSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	230	2176	2406
	Interview	121	2289	2410
	Total	351	4465	4816
RSMV		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	176	2120	2296
	Interview	175	2345	2520
	Total	351	4465	4816
ANN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	143	1954	2097
	Interview	208	2511	2719
	Total	351	4465	4816
MMNN		Correct Classification		Total
		Falsification	Real Interview	
Prediction	Falsification	224	2045	2269
	Interview	127	2420	2547
	Total	351	4465	4816

Source: IAB-BAMF-SOEP Survey of Refugees in Germany (version SOEP.v33) and Experimental data, University of Giessen, 2011.

References

AAPOR. 2023. “Standard Definitions - Final Dispositions of Case Codes and Outcome Rates for Surveys”. American Association for Public Opinion Research. Available at https://aapor.org/wp-content/uploads/2024/03/Standards-Definitions-10th-edition.pdf

Aggarwal, Charu C. 2018. Neural networks and deep learning. Springer.

Alfaro, Esteban, Matias Gamez, and Noelia Garcia. 2013. “Adabag: An R package for classification with boosting and bagging.” Journal of Statistical Software 54:1–35.

Bergmann, Michael, Karin Schuller, and Frederic Malter. 2019. “Preventing Interview Falsifications During Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE).” Longitudinal and Life Course Studies 10(4):513–30.

Biemer, Paul P., and Lars E. Lyberg. 2003. Introduction to survey quality. John Wiley & Sons.

Biemer, Paul P., and S. Lynne Stokes 1989. “The Optimal Design of Quality Control Samples to Detect Interviewer Cheating.” Journal of Official Statistics 5(1):23–39.

Birnbaum, Benjamin, Gaetano Borriello, Abraham D. Flaxman, Brian DeRenzi, and Anna R. Karlin. 2013. “Using behavioral data to identify interviewer fabrication in surveys.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Available at https://bbirnbaum.com/assets/publications/chi13.pdf.

Birnbaum, Benjamin. 2012. “Algorithmic Approaches to Detecting Interviewer Fabrication in Surveys.” Dissertation, University of Washington. Available at http://hdl.handle.net/1773/22011.

Bishop, Christopher M. 2006. Pattern recognition and machine learning. Springer.

Blasius, Jörg, and Victor Thiessen. 2012. Assessing the Quality of Survey Data. SAGE Publications.

Blasius, Jörg, and Victor Thiessen. 2013. “Detecting Poorly Conducted Interviews.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 67–88. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Blasius, Jörg, and Victor Thiessen. 2015. “Should we trust survey data? Assessing response simplification and data fabrication.” Social Science Research 52:479–93.

Blasius, Jörg, and Victor Thiessen. 2021. “Perceived corruption, trust, and interviewer behavior in 26 European Countries.” Sociological Methods & Research 50(2):740–77.

Bredl, Sebastian, Peter Winker, and Kerstin Kötschau. 2012. “A Statistical Approach to Detect Interviewer Falsification of Survey Data.” Survey Methodology Journal 38(1):1–10. Available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012001/article/11680-eng.pdf.

Breiman, Leo. 1996. “Bagging predictors.” Machine learning 24:123–40.

Breiman, Leo. 2001. “Random forests.” Machine learning 45:5–32.

Breiman, Leo, Jerome Friedman, Richard Olshen, and Charles Stone. 1984. Classification and Regression Trees. Chapman & Hall/CRC.

Brücker, Herbert, Nina Rother, and Jürgen Schupp. 2017. “IAB-BAMF-SOEP-Befragung von Geflüchteten 2016: Studiendesign, Feldergebnisse sowie Analysen zu schulischer wie beruflicher Qualifikation, Sprachkenntnissen sowie kognitiven Potenzialen.” In IAB-Forschungsbericht, Institut für Arbeitsmarkt und Berufsforschung. Available at https://www.iab.de/185/section.aspx/Publikation/k170918302.

Buskirk, Trent D., Antje Kirchner, Adam Eck, and Curtis S. Signorino. 2018. “An introduction to machine learning methods for survey researchers.” Survey Practice 11(1).

Chen, Tianqi, and Carlos Guestrin. 2016. “Xgboost: A scalable tree boosting system.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–94. Available at https://dl.acm.org/doi/pdf/10.1145/2939672.2939785.

Cohen, Mollie J., and Zach Warner. 2021. “How to Get Better Survey Data More Efficiently.” Political Analysis 29(2):121–38.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-vector networks.” Machine learning 20:273–97.

de Haas, Samuel, and Peter Winker. 2014. “Identification of partial falsifications in survey data.” Statistical Journal of the IAOS 30(3):271–281.

DeMatteis, Jill M., Linda J. Young, James Dahlhamer, Ronald E. Langley, Joe Murphy, Kristen Olson, and Sharan Sharma. 2020. “Falsification in Surveys: Task Force Final Report.” Washington, DC: American Association for Public Opinion Research. Available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Data_Falsification_Task_Force_Report.pdf.

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser. 2014. “Assessing the Mechanisms of Misreporting to Filter Questions in Surveys.” Public Opinion Quarterly 78(3):721–33.

Fawcett, Tom. 2006. “An introduction to ROC analysis.” Pattern recognition letters 27(8):861–74.

Finn, Arden, and Vimal Ranchhod. 2017. “Genuine fakes: The prevalence and implications of data fabrication in a large South African survey.” The World Bank Economic Review 31(1):129–57.

Freund, Yoav, and Robert E. Schapire. 1997. “A decision-theoretic generalization of on-line learning and an application to boosting.” Journal of computer and system sciences 55(1):119–39.

Friedman, Jerome H. 2001. “Greedy function approximation: a gradient boosting machine.” Annals of statistics 29(5):1189–232.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. “Additive logistic regression: a statistical view of boosting.” The annals of statistics 28(2):337–407.

Groves, Robert M. 2004. “Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects.” Survey Research 35(1):1–5.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. Survey Methodology. Hoboken, NJ: Wiley.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hoerl, Arthur E., and Robert W. Kennard. 1970. “Ridge regression: Biased estimation for nonorthogonal problems.” Technometrics 12(1): 55–67.

Hoffmann, Frank, Torsten Bertram, Ralf Mikut, Markus Reischl, and Oliver Nelles. 2019. “Benchmarking in classification and regression.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(5):e1318.

Hood, Catherine C., and John M. Bushery. 1997. “Getting More Bang from the Reinterview Buck: Identifying ‘At Risk’ Interviewers.” Proceedings of the Survey Research Method Section, American Statistical Association, 820–24. Available at http://www.asasrms.org/proceedings/papers/1997_141.pdf.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2015. “ctree: Conditional inference trees.” The comprehensive R archive network 8:1-34. Available at http://cran.irsn.fr/web/packages/partykit/vignettes/ctree.pdf.

Jacobsen, Jannes. 2018. “Language Barriers During the Fieldwork of the IAB-BAMF-SOEP Survey of Refugees in Germany.” In Surveying the Migrant Population: Consideration of Linguistic and Cultural Issues, edited by Dorothee Behr, 75–84. Köln: GESIS–Leibniz-Institut für Sozialwissenschaften.

Jebreel, Najeeb Moharram, Rami Haffar, Ashneet Khandpur Singh, David Sánchez, Josep Domingo-Ferrer, and Alberto Blanco-Justicia. 2020. “Detecting bad answers in survey data through unsupervised machine learning.” In Privacy in Statistical Databases, UNESCO Chair in Data Privacy, International Conference Proceedings, edited by Josep Domingo-Ferrer and Krishnamurty Muralidhar, 309–20. Springer International Publishing.

Jesske, Birgit. 2013. “Concepts and Practices in Interviewer Qualification and Monitoring.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 91–102. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Kern, Christoph, Thomas Klausch, and Frauke Kreuter. 2019. “Tree-based machine learning methods for survey research.” Survey research methods 13(1):73–93.

Kosyakova, Yuliya, Jan Skopek, and Stephanie Eckman. 2015. “Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach.” International Journal of Public Opinion Research 27(3):417–31.

Kosyakova, Yuliya, Lukas Olbrich, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2019. “Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany.” Institut für Arbeitsmarkt- und Berufsforschung. Available at https://fdz.iab.de/187/section.aspx/Publikation/k190404302.

Kroh, Martin, Simon Kühne, Jannes Jacobsen, Manuel Siegert, and Rainer Siegers. 2017. “Sampling, Nonresponse, and Integrated Weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4)–Revised Version.” SOEP Survey Papers, No. 477. Berlin: DIW–German Institute for Economic Research. Available at http://hdl.handle.net/10419/172792.

Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, R. Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. 2023. “Package ‘caret’.” CRAN Documentation. Available at http://cran.radicaldevelop.com/web/packages/caret/caret.pdf.

Kursa, Miron B., and Witold R. Rudnicki. 2010. “Feature selection with the Boruta package.” Journal of statistical software 36:1–13.

Lang, Bernhard. 2005. “Monotonic multi-layer perceptron networks as universal approximators.” In Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005, edited by Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny, 31–7. Springer.

Lantz, Brett. 2019. Machine learning with R: expert techniques for predictive modeling. Packt publishing ltd.

Li, Jianzhu, J. Michael Brick, Back Tran, and Phyllis Singer. 2011. “Using Statistical Models for Sample Design of a Reinterview Program.” Journal of Official Statistics 27(3):433–50.

Menold, Natalja, Peter Winker, Nina Storfinger, and Christoph J. Kemper. 2013. “A Method for Ex-Post Identification of Falsification in Survey Data.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 25–47. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Murphy, Joe, Joe Eyerman, Colleen McCue, Christy Hottinger, and Joel Kennet. 2005. “Interviewer Falsification detection using data mining.” Proceedings of Statistics Canada Symposium 2005, Methodological Challenges for Future Information Needs. Available at https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X20050019445.

Murphy, Joe, Rodney Baxter, Joe Eyerman, David Cunningham, and Joel Kennet. 2004. “A System for Detecting Interviewer Falsification.” Proceedings of the American Statistical

Association and the American Association for Public Opinion Research. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000517.pdf.

Olbrich, Lukas, Yuliya Kosyakova, Joseph W. Sakshaug, and Silvia Schwanhäuser. 2023. "Detecting Interviewer Fraud Using Multilevel Models." Journal of Survey Statistics and Methodology 12(1):14–35.

Porras, Javier, and Ned English. 2004. "Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys." Proceedings of the Survey Research Method Section, American Statistical Association, 4223–28. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000879.pdf.

Rebala, Gopinath, Ajay Ravi, and Sanjay Churiwala. 2019. An introduction to machine learning. Springer.

Robbins, Michael. 2018. "New frontiers in detecting data fabrication." In Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC), edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop, and Brita Dorer, 771–805 Wiley & Sons, Inc.

Rosmansyah, Yusep, Ibnu Santoso, Ariq Bani Hardi, Atina Putri, and Sarwono Sutikno. 2019. "Detection of Interviewer Falsification in Statistics Indonesia's Mobile Survey." International Journal on Electrical Engineering and Informatics 11(3): 474–84.

Schäfer, Christin, Jörg-Peter Schräpler, Klaus-Robert Müller, and Gert G Wagner. 2004. Automatic identification of faked and fraudulent interviews in surveys by two different methods. DIW Discussion Papers.

Schapire, Robert E., and Yoav Freund. 2012. Boosting: Foundations and Algorithms. The MIT Press.

Schräpler, Jörg-Peter, and Gert G. Wagner. 2005. "Characteristics and Impact of Faked Interviews in Surveys: An Analysis of Genuine Fakes in the Raw Data of SOEP." Allgemeines Statistisches Archiv 89:7–20.

Schwanhäuser, Silvia, Joseph W Sakshaug, Yuliya Kosyakova, and Frauke Kreuter. 2020. "Statistical identification of fraudulent interviews in surveys: improving interviewer controls." In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 91–106. Boca Raton, FL: Taylor & Francis Group.

Schwanhäuser, Silvia, Joseph W. Sakshaug, and Yuliya Kosyakova. 2022. "How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification." Public Opinion Quarterly 86(1):51–81.

Shah, Neha, Diwakar Mohan, Jean Juste Harisson Bashingwa, Osama Ummer, Arpita Chakraborty, and Amnesty E. LeFevre. 2020. "Using machine learning to optimize the

quality of survey data: protocol for a use case in India." JMIR Research Protocols 9(8). Available at https://www.researchprotocols.org/2020/8/e17619/.

Slomczynski, Kazimierz Maciek, Przemek Powalko, and Tadeusz Krauze. 2017. "Non-Unique Records in International Survey Projects: The Need for Extending Data Quality Control." Survey Research Methods 11(1):1–16.

Stitson, M.O., J.A.E. Weston, A. Gammerman, V. Vovk, and V. Vapnik. 1996. "Theory of support vector machines." University of London 117(827):188–91. Available at https://ynucc.yu.ac.kr/~shkwon/lectures/ic/svm/svm_1.pdf.

Stokes, S. Lynne, and Patty Jones. 1989. "Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey." Proceedings of the Survey Research Method Section, American Statistical Association, 696–98. Available at http://www.asasrms.org/Proceedings/papers/1989_127.pdf.

Storfinger, Nina, and Peter Winker. 2013. "Assessing the Performance of Clustering Methods in Falsification Using Bootstrap." In Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 46–65. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Sun, Hanyu, and Ting Yan. 2023. "Applying Machine Learning to the Evaluation of Interviewer Performance." Survey Practice 16(1).

Tharwat, Alaa. 2020. "Classification assessment methods." Applied computing and informatics 17(1):168–92.

Thissen, M Rita. 2014. "Computer audio-recorded interviewing as a tool for survey research." Social Science Computer Review 32(1):90–104.

Thissen, M. Rita, and Susan K. Myers. 2016. "Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality." Statistical Journal of the IAOS 32(3):339–47.

Thissen, M. Rita, Sridevi Sattaluri, Emily McFarlane, and Paul P. Biemer. 2008. "The evolution of audio recording in field surveys." Survey Practice 1(5).

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society Series B: Statistical Methodology 58(1):267–288.

Wagner, James, Kristen Olson, and Minako Edgar. 2017. "The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata." Survey Research Methods 11(3):218–33.

Walzenbach, Sandra. 2021. "Do falsifiers leave traces? Finding recognizable response patterns in interviewer falsifications." Methods, data, analyses: a journal for quantitative methods "and survey methodology (mda) 15(2):125–60.

Weinauer, Marlene. 2019. "Be a Detective for a Day: How to Detect Falsified Interviews with Statistics." Statistical Journal of the IAOS 35(4):569–75.

Winker, Peter. 2016. “Assuring the quality of survey data: Incentives, detection and documentation of deviant behavior.” Statistical Journal of the IAOS 32(3):295–303.

Witten, Ian H., and Eibe Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, Elsevier.

5. Discussion

5.1 Summary

In many academic disciplines, interviewer-administered surveys represent a crucial data source for research. Ensuring the high quality of this data is of vital importance to guarantee the reliability of research results (Olson et al. 2020). A major thread to the data quality, which presents a substantial challenge in survey research, is the phenomenon of interviewer falsification. This dissertation has therefore undertaken a comprehensive examination of this topic. The malpractice of interviewers deviating from their guidelines or instructions—in the worst case fabricating entire interviews, can have severe effects on the data (Schräpler and Wagner 2005; Brüderl et al. 2013; DeMatteis et al. 2020). Given the systematic nature of this fraudulent interviewer behavior, even a small number of fabricated interviews holds the potential to severely bias the results of analyses (Schräpler and Wagner 2005). To counteract this potential thread, the existing literature proposes a range of strategies for preventing and detecting interviewer falsification. In particular, statistical or data-based detection methods enjoy popularity among survey literature (see, e.g., Stokes and Jones 1989; Hood and Bushery 1997; Turner et al. 2002; Murphy et al. 2004; Porras and English 2004; Li et al. 2011; Bredl, Winker, and Kötschau 2012; Menold et al. 2013; Bergmann, Schuller, and Malter. 2019; Schwanhäuser, Sakshaug, and Kosyakova 2022). This dissertation presented a broad evaluation of several statistical detection methods, including falsification indicators, multivariate approaches combining falsification indicators, approaches focusing on the identification of duplicated response patterns, and innovative machine learning algorithms.

Despite the considerable variety of proposed detection methods in the literature, many studies are subject to a number of limitations: They often focus on complete falsifications, neglecting the more subtle partial falsifications that are harder to detect. Many studies lack external validity as they are based on experimental data or simulated scenarios rather than real-world applications. Rather than comparing different detection methods, they showcase one single detection method. In conclusion, the existing literature lacks comprehensive and systematic evaluations of the various detection methods under different circumstances. This dissertation seeks to address this gap by evaluating a broad range of data-based detection tools within the context of real-world survey data. By conducting multiple systematic examinations of different detection methods, which consider different aspects, like partial falsifications,

longitudinal settings, or advancing technologies, this dissertation offers valuable insights for both researchers and practitioners engaged in the field of survey methodology.

The first paper evaluates the effectiveness of different detection methods, including numerous existing as well as newly proposed falsification indicators, cluster analysis using two different clustering algorithms, and a novel meta-indicator approach. The results confirm the effectiveness of multivariate detection methods in detecting interviewer falsification, even after applying robustness checks. Furthermore, the paper finds high effectiveness of different falsification indicators in discriminating between real and falsified interviews. With these results, the paper highlights the added value of combining multiple indicators when aiming to the detection of falsifiers. It also confirms the assumptions about the direction and usefulness of falsification indicators.

The second paper extends the evaluation of data-based detection methods to the context of partial interviewer falsification in panel surveys. The paper gives important implications for data quality controls in panel survey, as it compares a wide range of established methods used on cross-sectional data (like cluster analysis, outlier detection, and principal component analysis) as well as methods focusing on the longitudinal structure of panel data by examining correlations between adjacent waves. The paper further provides insights into the sensitivity of these methods in detecting partial falsifications. The study reveals that many cross-sectional methods, such as cluster analysis and outlier detection, are also effective in revealing partial falsifications. Importantly, longitudinal methods in this case did not provide a clear indication of partial falsifications. This challenges the common assumption that falsifiers can be easily identified by discrepancies between two adjacent waves. The results underscore that even basic statistical detection methods can help identify partial falsifications and, hence, effectively enhance quality controls.

The final paper further broadens the scope of detection methods by exploring the potentials of innovative supervised or rather trained machine learning algorithms. These methods have rarely been used and evaluated in the context of falsification detection. Using different regression models, decision trees, support vector machines, and neural networks, the paper evaluates the individual performance of these algorithms in three distinct scenarios: first, when training the algorithms on falsifications within the same survey, second, when training them on falsifications by different falsifiers within the same survey, and third, when training them on falsifications from a completely different survey. While many algorithms showed promising results in the first two scenarios, their performance significantly dropped when

applied across different surveys. These findings indicate that, dependent on the context, particularly decision-tree-based algorithms, can assist in improving quality controls. This is especially the case, if training data from the same survey is available.

5.2 Practical Implications

The identification of interviewer falsification and the corresponding improvement of data quality controls are very practical topics. For practitioners in the field of survey research, it is important to make informed decisions about the concrete design of their quality controls, selecting the tools best suited to fit their specific needs. This is not always easy if the respective costs and benefits of each tool are unclear to them. To address this issue, this dissertation contributed to a more nuanced understanding of possible statistical tools for detecting interviewer falsification, and offers some practical implications for selecting and implementing effective methods. In summary, this dissertation has three main implications.

The first implication is that easily implemented detection tools often yield most accurate results. This was for example demonstrated by the meta-indicator approach, which is straightforward to interpret and performed similarly well to cluster analysis. This finding was further emphasized in the second paper, where the commonly used cluster analysis and a rapidly applied outlier detection algorithm were effective in identifying partial falsifications, even though these methods are typically applied to identify complete falsifications. Additionally, both papers indicate that a small set of indicators might suffice to detect interviewer falsification. Importantly, this indicator set should ideally combine different indicator types to cover a broad range of potential fraudulent patterns. For instance, one might combine time-related paradata indicators with item scale indicators and content-related indicators such as Benford's Law.

Secondly, the collective findings of all papers indicate that falsifiers employ a wide range of falsification strategies, which sometimes results in very different patterns. Given that the specific fraud behavior may strongly depend on the falsifiers' motivation, it is possible that some assumptions regarding the measurable patterns may not hold. For instance, the second paper demonstrated that both lower and higher levels of item nonresponse can be an indicator of fraudulent behavior. Additionally, the patterns resulting from fraudulent behavior may vary depending on the type of falsification. In the case of complete falsifications, simplified response behavior was frequently observed, whereas partially falsifying interviewers exhibited a variety of fabrication strategies, and showed learning effects across survey waves. Furthermore,

fraudulent behavior may differ based on the interviewers' demographics, as the falsifying students applied different strategies than the real-world falsifiers. This underscores the importance of combining different control methods that address different types of falsification behavior and hence effectively capture the full spectrum of fraudulent behaviors.

Finally, this dissertation highlights the potential value of exploring new methods, including the meta-indicator approach presented in the first paper and the supervised machine learning techniques discussed in the third paper. Nevertheless, it is crucial to thoroughly evaluate these methods and to compare them with existing ones to determine their utility and limitations. The efficiency of these tools may vary significantly depending on the specific context of the survey, underscoring the importance of having appropriate data for testing various methods. Making falsification data publicly available is therefore of crucial importance in order to enable the replication of results and the assessment of the robustness of various methods. In conclusion, this dissertation encourages practitioners to experiment with new methods and document their findings, regardless of their success.

5.3 Limitations and Future Research

It should be noted that this dissertation is not without limitations and is therefore only able to close some existing gaps related to interviewer falsification in surveys. First, the presented results may depend on the respective dataset used. For instance, the findings may be affected by the survey population, as the data used in this dissertation originate from highly specific groups: A sample of refugees in Germany, recipients of social benefits in Germany, as well as students from one German university. Consequently, the results might be biased towards the European context, including culture-specific patterns, regulations, and survey characteristics. Moreover, as two out of the three datasets originate from real-world surveys, there is a risk that some falsifications were not detected during the control process. Hence, some interviews might be labeled as real interviews even though they were indeed fabricated. This could slightly bias the performance evaluation with regard to false-positive cases. Additionally, all results are based on face-to-face interviews, as telephone interviews including fraudulent interviews were not available for analysis. Telephone interviews are typically conducted in centralized facilities with detailed monitoring, which reduces the likelihood of interviewer falsification (Robbins 2018). Even though all evaluated methods could be applied to data from telephone interviews, this dissertation cannot make claims about the performance of the detection methods in this context.

The second limitation of this dissertation is that it solely focuses on the identification of interviewer falsification using statistical detection methods, thereby addressing a very specific aspect of survey data quality. The three papers do not evaluate detection or quality control methods beyond statistical tools. Consequently, this dissertation does not provide insights into non-statistical methods, such as monitoring techniques or re-interviewing, nor does it investigate prevention strategies for interviewer falsification. Both are, however, essential components of a comprehensive system to ensure survey data quality. As previously stated, data-based methods can only indicate fraudulent behavior, but not confirm it. In practice, data-based detection methods need to be combined with non-statistical approaches. Due to its specific focus on interviewer falsification, this dissertation neglects potential fraudulent behavior by other actors (such as supervisors or even researchers) in the survey process or tendencies of misreporting by respondents, even though both negatively impact data quality. As a result, this dissertation does not provide insights or evaluations of tools used to detect these kinds of behaviors. Moreover, this dissertation does not explicitly focus on milder forms of deviant interviewer behavior, such as deviations in the probing process or reading deviations. Although such deviations might be included in partial falsification behaviors, there is limited evidence regarding their efficiency in detecting such interviewer behavior.

In light of these limitations, this dissertation paves the way for future research in the field of interviewer falsification and fraud in surveys. First, as this dissertation only investigated the detection of interviewer fraud in face-to-face surveys, there is less evidence regarding telephone interviews. Furthermore, the existing literature has largely overlooked the potential of mixed-mode designs in the context of falsification detection. Although mode effects may influence data from two distinct modes, data from a mode less prone to interviewer effects and falsification could serve as a baseline for identifying anomalous responses in face-to-face data. Particularly in the context of content-related indicators, this could open up new avenues for quality control. This could also help reduce the number of false-positives in statistical detection methods, as thresholds for suspicious values could be defined based on results from other modes. To determine the efficiency of this approach, further research is required.

Secondly, all methods were applied ex-post, that is, after the fieldwork of the survey had been completed. Therefore, it is unclear how well these methods perform during the field period. Many detection methods aggregate results at the interviewer-level, but early in their fieldwork, interviewers may have conducted only a small number of interviews. This could lead to respondents' answer patterns being misinterpreted as overall interviewer effects, potentially

resulting in unwarranted suspicion against some interviewers. More research is needed to establish appropriate thresholds and to evaluate the within-field performance of different detection methods. Applying these methods during fieldwork would enhance quality controls, as early intervention could prevent further fraudulent activity and allow for re-interviewing respondents to avoid data loss. Deleting data at the end of fieldwork might cause selection issues, especially if falsifiers primarily fabricate interviews of unwilling or challenging respondents. Therefore, further research on real-time tools for falsification detection is necessary.

Third, as previously described, the different contributions document a wide range of falsification behaviors and strategies. Although numerous assumptions can be made about the directions and motivations of fraudulent interviewer behaviors, very little is known about the true motivations of falsifiers, the corresponding effects on data quality, and how deliberate some of their deviations really are. More research is needed, to also improve feedback loops to interviewers. Some studies address this issue, but further research is necessary (e.g., Edwards, Sun, and Hubbard 2020). Over the time, interviewers have become a scarce resource. Consequently, it is crucial to optimize interviewers' working environment, address their concerns, and to raise the interviewers’ awareness regarding the importance of data quality.

Finally, there is a trend towards increasing use of online surveys (Evans and Mathur 2005). Consequently, concerns about the data quality also shift towards this mode. Other threats, such as inattentive or fraudulent respondents and survey bots, represent another form of falsification in survey research. Given the long tradition of detecting interviewer falsification, the question arises whether some of these techniques can be adapted to identify fraudulent respondents or survey bots. Although an increasing number of studies are addressing this issue, there is a need to develop and validate these tools to ensure data quality in online surveys.

References

Bergmann, Michael, Karin Schuller, and Frederic Malter. 2019. “Preventing Interview Falsifications During Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE).” Longitudinal and Life Course Studies 10(4):513–30.

Bredl, Sebastian, Peter Winker, and Kerstin Kötschau. 2012. “A Statistical Approach to Detect Interviewer Falsification of Survey Data.” Survey Methodology Journal 38(1):1–10. Available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012001/article/11680-eng.pdf.

Brüderl, Josef, Bernadette Huyer-May, and Claudia Schmiedeberg. 2013. “Interviewer behavior and the quality of social network data.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 147–60. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

DeMatteis, Jill M., Linda J. Young, James Dahlhamer, Ronald E. Langley, Joe Murphy, Kristen Olson, and Sharan Sharma. 2020. “Falsification in Surveys: Task Force Final Report.” Washington, DC: American Association for Public Opinion Research. Available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Data_Falsification_Task_Force_Report.pdf.

Edwards, Brad, Hanyu Sun, and Ryan Hubbard. 2020. “Behavior Change Techniques for Reducing Interviewer Contributions to Total Survey Error.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 77–89. Boca Raton, FL: Taylor & Francis Group.

Evans, Joel R., and Anil Mathur. 2005. “The value of online surveys.” Internet Research 15(2):195-219.

Hood, Catherine C., and John M. Bushery. 1997. “Getting More Bang from the Reinterview Buck: Identifying ‘At Risk’ Interviewers.” Proceedings of the Survey Research Method Section, American Statistical Association, 820–24. Available at http://www.asasrms.org/proceedings/papers/1997_141.pdf.

Li, Jianzhu, J. Michael Brick, Back Tran, and Phyllis Singer. 2011. “Using Statistical Models for Sample Design of a Reinterview Program.” Journal of Official Statistics 27(3):433–50.

Menold, Natalja, Peter Winker, Nina Storfinger, and Christoph J. Kemper. 2013. “A Method for Ex-Post Identification of Falsification in Survey Data.” In Interviewers’ Deviations in Surveys: Impact, Reasons, Detection and Prevention, edited by Peter Winker, Natalja Menold, and Rolf Porst, 25–47. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Murphy, Joe, Rodney Baxter, Joe Eyerman, David Cunningham, and Joel Kennet. 2004. “A System for Detecting Interviewer Falsification.” Proceedings of the American Statistical Association and the American Association for Public Opinion Research. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000517.pdf.

Olson, Kristen, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West. 2020. “The Past, Present, and Future of Research on Interviewer Effects.” In Interviewer Effects from a Total Survey Error Perspective, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 3–16. Boca Raton, FL: Taylor & Francis Group.

Porras, Javier, and Ned English. 2004. “Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys.” Proceedings of the Survey Research Method Section, American Statistical Association, 4223–28. Available at http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000879.pdf.

Robbins, Michael. 2018. “New frontiers in detecting data fabrication.” In Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC), edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop, and Brita Dorer, 771–805 Wiley & Sons, Inc.

Schräpler, Jörg-Peter, and Gert G. Wagner. 2005. “Characteristics and Impact of Faked Interviews in Surveys: An Analysis of Genuine Fakes in the Raw Data of SOEP.” Allgemeines Statistisches Archiv 89:7–20.

Schwanhäuser, Silvia, Joseph W. Sakshaug, and Yuliya Kosyakova. 2022. “How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification.” Public Opinion Quarterly 86(1):51–81.

Stokes, S. Lynne, and Patty Jones. 1989. “Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey.” Proceedings of the Survey Research Method Section, American Statistical Association, 696–98. Available at http://www.asasrms.org/Proceedings/papers/1989_127.pdf.

Turner, Charles F., James N. Gribble, Alia A. Al-Tayyib, and James R. Chromy. 2002. “Falsification in Epidemiologic Surveys: Detection and Remediation.” Technical Papers on Health and Behavior Measurement, No. 53. Washington, DC: Research Triangle Institute.